國立交通大學

生物資訊及系統生物研究所

博 士 論 文

同源蛋白質交互作用與複合體剖析蛋白質交互作用體行為

Homologous protein-protein interactions and protein complexes reveal interactome behavior

研 究 生：羅宇書

指導教授：楊進木　教授

中 華 民 國 一百零一 年 八 月

同源蛋白質交互作用與複合體剖析蛋白質交互作用體行為

Homologous protein-protein interactions and protein complexes reveal interactome behavior

研 究 生：羅宇書　　Student：Yu-Shu Lo

指導教授：楊進木　　Advisor：Jinn-Moon Yang

國 立 交 通 大 學

生物資訊及系統生物研究所

博 士 論 文

A Thesis Submitted to Institute of Bioinformatics and Systems Biology
National Chiao Tung University in partial Fulfillment of the Requirements
for the Degree of Ph.D. in
Bioinformatics and Systems Biology

August 2012

Hsinchu, Taiwan, Republic of China

中華民國一百零一年八月

Homologous protein-protein interactions and protein complexes reveal interactome behavior

Student : Yu Shu Lo Adviser : Dr. Jinn-Moon Yang
Institute of Bioinformatics and Systems Biology
National Chiao Tung University

Abstract

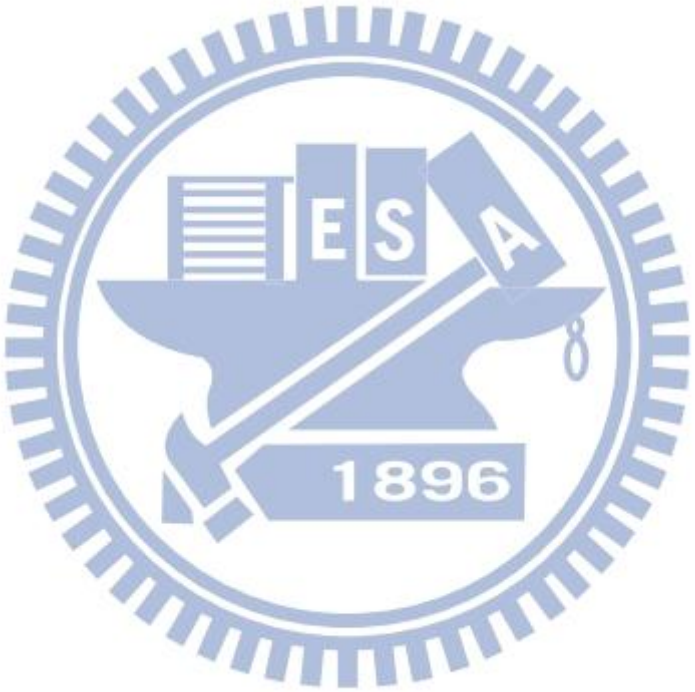
Protein-protein interaction (PPI) networks provide key insights into complex biological systems, from how different processes communicate to the function of individual residues on a single protein. Therefore, several large network databases (e.g. IntAct, DIP, and BioGRID) record hundreds of thousands of physical and genetic interactions from a wide variety of organisms have been purposed. However, these PPI databases are dominated by few species and usually could not provide the binding mechanisms. Therefore, constructing the structure resolved PPI networks across multiple organisms should provide a great value for investigating the behavior of PPI network.

To address the issues, we proposed the concepts of protein interaction family (i.e. protein-protein interaction family and protein complex family) to construct a structure resolved PPI networks and study the behaviors of a specific PPI network. The protein interaction family is a group of protein interactions (PPI or protein complex) which share the consensus interacting domain, binding environment, and have similar biological processes. According to the concept "3D-domain interolog mapping" with a scoring system, we are able to explore all homologous protein-protein interaction pairs (protein-protein interaction family) between two homolog families, derived from a known 3D-structure dimmer (template), across multiple species. Then, we also identify the homologous protein complexes with the binding models (e.g. hydrogen bonds and conserved amino acids in the interfaces), functional modules, and the conserved interacting domains and Gene Ontology annotations in multiple organisms.

Based on the PPIs derived from "3D-domain interolog mapping" and "protein complex family", we are able to construct structure resolved PPI networks in multiple organisms (e.g. *Homo sapiens*, *Mus musculus*, and *Danio rerio*). In each network, the PPIs with residue-based binding models have a highly agreement in Gene Ontology similarities. Furthermore, the architecture (i.e. scale-free network properties) of these networks is consistent with some cellular networks of previous studies. In addition, the consensus proteins and PPIs derived form on our method are highly related to the essential genes and disease related proteins recorded in OMIM. We also indicate that the disease related mutations are more enrichment on the interacting residues, especially on the hydrogen bond residues. In addition, for a given PPI

network, we also provided a new characterization (named *MS-matrix*) to describe the modularity and relative importance of proteins. We believe that structure resolved PPI networks derived from the PPI family would provide the insight for understanding the mechanism of biological processes within a given PPI network.

.



同源蛋白質交互作用與複合體剖析蛋白質交互作用體行為

研究生：羅宇書　　指導教授：楊進木博士

國立交通大學 生物資訊及系統生物研究所 博士班

中文摘要

透過蛋白質交互作用網路（protein interaction network）可以對於複雜的生物系統有更進一步的了解，例如探討不同生化途經之間的協同作用、蛋白質上特定殘基對功能的影響。因此，大量的交互作用資料庫（如：IntAct、DIP 和 BioGRID 等）被建立來探討蛋白質交互作用網路。然而，這些資料庫中的蛋白質交互作用資料往往集中在少數的物種，而且也缺乏對於交互作用介面機制的解釋。

針對此議題，我們提出了蛋白質交互作用家族的觀念（包含蛋白質－蛋白質交互作用家族(protein-protein interaction family)和蛋白質複合體家族(protein complex family)）以協助多個物種中建立具有結構解析的蛋白質交互作用網路，並探討單一物種的蛋白質交互作用網路行為。蛋白質交互作用家族為一群有擁有保留性交互作用區塊、結合環境和相似生物途徑的蛋白質交互作用所組成。而透過“3D-domain interolog mapping”與一個新的能量函式，我們將透過已知結構的模板（template），來探索多個物種間所有的同源蛋白質相互作用。此外，我們也在多個物種中找尋同源蛋白複合體，並描述了結合模型（例如在交互作用介面上的氫鍵和保留性氨基酸配對）、功能模塊、保留性相互作用區塊（interacting domain）和 Gene Ontology。

透過由“3D-domain interolog mapping”與蛋白質複合體家族，我們在人類、老鼠與斑馬魚中建立了具有結構解析的蛋白質交互作用網路。在每一個網路中，這些具有結構解析的蛋白質交互作用與 Gene Ontology 相似度有很高的一致性。此外，這些網路也都具有之前對於生化網路研究中所指出的拓譜特性(scale-free network)。而透過蛋白質交互作用家族我們也可指出網路中在跨物種間具有高度保留的蛋白質與交互作用，而這些蛋白質往往是生存必須基因(essential gene)或是跟疾病相關。更進一步的，這些跟疾病相關的基因突變往往位於蛋白質交互作用介面，並擔任重要的交互作用（例如氫鍵）。此外，對於單一蛋白質交互作用網路，我們提出了一個新的概念“*MS-matrix*”來描述網路上重要的蛋白質以及模組化特性。基於上述這些研究，我們認為透過交互作用家族所構建的結構解析交互作用網路對於了解生化途徑的機制是很有幫助的。

誌謝

在順利撰寫完這份博士論文並取得博士學位的過程中，我獲得非常多的貴人相助。首先，我必須感謝我的指導教授楊進木老師，老師是一位非常有研究熱誠的研究者與教育家，在我遭遇到瓶頸或是對事情產生消極態度的時候，總是可以透過跟老師的開導，而讓我重新燃起希望或是突破瓶頸。除了在研究專業上的教導與訓練之外，在生活與人生態度上，更是教導了我許多，指出了我所欠缺的積極思考、積極爭取的態度與其他那些該糾正的缺點。更感謝老師在我博士班研讀期間所提供的學習環境、研究資源以及出國參與研討會的機會，不僅使我在求學時能夠更為順利亦增廣了我的見聞。

接著我要感謝我的口試委員，包含我的指導教授楊進木教授、口試召集人熊昭教授、江安世教授、黃鎮剛教授、莊永仁教授、陳豐奇教授。感謝每位教授在百忙之中抽空擔任我的口試委員並且評鑑我的論文，更感謝各位教授在口試期間對我的研究所提出的寶貴建議，有了各位教授的建議與指導才使得這份研究論文能夠更臻完美。

我也要感謝我的實驗室其他一起努力的夥伴們，特別感謝系統生物組的峻宇、怡馨、星翰、尚文、采凌、俊辰、怡瑋。我們系統生物組都是靠全組成員一起努力才能有不錯的研究成果。特別感謝峻宇，他是一位非常優秀的學弟，在研究上跟我可以互相討論、互相激勵，也可以容忍我這個組長沒有耐心的脾氣，讓很多研究可以順利進行。也感謝怡馨，在研究上她是個非常有耐心跟細心的研究者，可以注意到我沒有想到的一些地方，彼此相輔相成，也在行政事務上給了我很大的幫助，讓我能輕鬆很多。並感謝星翰跟采凌兩位學弟妹，他們認真的態度讓我們的研究可以順利的進行。並感謝實驗室其他同仁，特別是凱程、章維、志達與一原在討論時往往給予一些很有助益的建議與看法，而且一原也在程式方面給予了我很大的協助。因為實驗室的同仁讓辛苦的博士生活更顯得豐富與多采多姿。

最後我想特別感謝我父母跟怡馨，我的父母能夠體諒我在博士研讀期間的苦悶，對我十分的體貼與關心，常常原諒我的無理取鬧並傾聽我很多日常瑣碎的抱怨與訴苦，支持我讓我能夠面對一切困難。怡馨是一位非常非常好的伴侶，能在我不開心的時候安慰我、鼓勵我。開心的時候則一起分享喜悅。研究上也可以互相扶持，忍耐與糾正我的很多缺點。有她陪伴在身邊讓很多很困難的事情，都變成兩個人可以克服的小事。希望我們兩個可以一直這樣走下去。

最後，僅將此論文獻給我這些敬愛的人以及幫助過我的人。

Contents

List of Figures

List of Tables

Chapter 1. Introduction

1-1. Background

Protein-protein interaction (PPI) networks provide key insights into complex biological systems, from how different processes communicate to the function of individual residues on a single protein. For instance, the systematic identification of protein-protein interactions[1-3] or protein complexes[4-7] has been a widely used strategy for understanding the physical architecture of the cell. Therefore, several large network databases such as IntAct[8], DIP[9], and BioGRID[10] record hundreds of thousands of physical and genetic interactions from a wide variety of organisms have been purposed.

A wealth of investigations have been undertaken to deepen our understanding of hereditary diseases. As a result of that, databases such as the Online Mendelian Inheritance in Man (OMIM)[11] and UniProt[12] together contain almost 30,000 experimentally verified mutations. Nevertheless, the exact mechanisms by which mutations alter a protein's function are in many cases poorly understood. Therefore, researchers have recently begun to use PPI networks to explore the genotype-to-phenotype relationships[13-16], on the basis that many proteins function by interacting with other proteins. However, this idea has only been applied in Human based on the requirement of high-quality PPI with the binding mechanism.

In addition, the concept "homologs" is useful for identifying consensus proteins across multiple organisms and could provide the key residues related to the functions within a given protein. Previous studies have been compared PPI network across multiple organisms to identify the essential pathways and the mechanisms of evolution[17-19]. For example, Peterson, G. J. *et al.* have shown that interaction change through binding site evolution is faster than through gene gain or loss[19] based on the comparison between 23 fungal PPI networks.

However, these studies only focused on a small sub-network or on few organisms which have an enrichment PPI data (e.g. *Homo sapiens* and *Saccharomyces cerevisiae*).

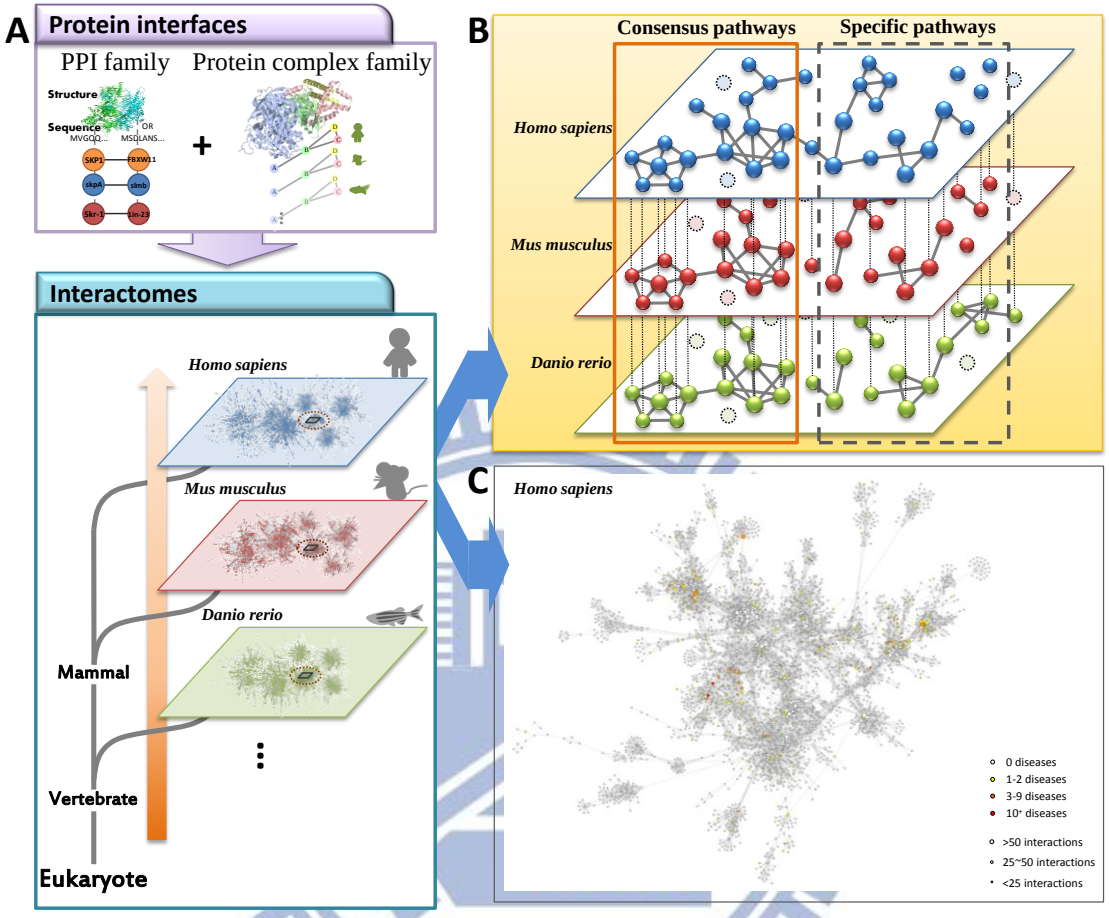


Figure 1-1. The overview of constructing the structure resolved PPI networks and studying the interactome behavior

(A) Using protein-protein interaction family and protein complex family to construct the structure resolved PPI networks in multiple oragnsims. (B) The "interactome behavior" through the consensus component. (C) The structure resolved PPI networks would provide the insight for understanding the mechanism of biological processes.

To address these issues, the structure resolved interaction family (i.e. protein-protein interaction family and protein complex family) are the basic elements and the core idea of our research to construct structure resolved PPI networks and study the behaviors of a specific PPI network. The PPI family is a group of molecular interactions which share the consensus interacting domain, binding environment, and have similar biological processes. The concepts

of PPI families not only help us to construct the highly reliable PPI network in a specific organisms (e.g. *Homo sapiens*, *Mus musculus*, and *Danio rerio*) but also provide the consensus and the diversity behavior of interactome through comparing with multiple species (Fig. 1-1). The methods of inferring interface families and interactomes are briefly summarized as follows.

In protein-protein interaction family, the concept of PPI families is similar to that of protein sequence family[20,21] and protein structure family[22]. Here, the members of a PPI family are conserved on specific functions and in interacting domain(s). Using these conservations of homologous PPIs, it can be used to annotate the protein functions and provide high quality PPIs.

Protein complexes are fundamental units of macromolecular organization and their composition is also known to vary according to cellular requirements[7]. According to these homologous complexes across multiple species, protein complex family provides the binding models (e.g. hydrogen bonds and conserved amino acids in the interfaces), functional modules, and the conserved interacting domains and Gene Ontology annotations of the members.

Based on the members (protein-protein or protein complexes) of protein-protein interface family[23] and protein complex family[24] that are consensus of functional annotation across multiple species, we are able to identify the conserved components in the PPI networks across multiple species and indicate the changes of the conserved components at the interspecific level. Therefore, we would use the strategies to reveal "interactome behavior".

1-2. Current state of constructing protein-protein interaction networks

Many high throughput experimental and computational approaches, such as high-throughput yeast two-hybrid screening[25,26] and co-affinity purification[27], have been

proposed to construct the PPI network within an organism. These large-scale methods are often unable to respond how a protein interacts with another one and describe the relationship between the mutation of proteins and disease syndrome. Previous studies have combined protein structure information with protein interaction data to investigate how mutations affect protein interactions in disease[14-16]. For instance, Wang, X. J. *et al.* generated a structurally resolved human protein interaction network to systematically examine relationship genes, mutations and associated disorders[16].

Table 1-1. The list of the members of proteins and protein-protein interactions in 11 common used organisms

NCBI Taxonomy ID	Organisms	No. Proteins in Integr8 database	No. PPIs in five annotated database
9606	*Homo sapiens*	56,006	67,596
10090	*Mus musculus*	36,379	7,535
3702	*Arabidopsis thaliana*	35,825	6,985
6239	*Caenorhabditis elegans*	23,154	10,095
7227	*Drosophila melanogaster*	15,155	37,674
7955	*Danio rerio*	21,601	221
10116	*Rattus norvegicus*	13,807	2,199
9913	*Bos taurus*	12,235	281
9031	*Gallus gallus*	6,279	70
36329	*Plasmodium falciparum*	5,353	2,956
4932	*Saccharomyces cerevisiae*	5,727	237,193
	Total	231,521	372,805

However, the experimental PPI data is necessary for these methods. The experimental PPI databases (e.g. IntAct[8], DIP[9], MIPS[28], BioGRID[10], and MINT[29]) are dominated by few species, especially *Saccharomyces cerevisiae*. Table 1-1 presents the number of PPIs and proteins in organisms that are commonly used in molecular researches. For example, there are 56,006 proteins (24.19% of 11 common organisms) and 67,596 PPIs (18.1% of 11 common organisms) of *Homo sapiens* are recorded in Integr8 database [30] (which are collected the complete sequencing genomes) and the five public interaction databases, respectively. On the contrary, the *Saccharomyces cerevisiae* only has 5,727 proteins (2.4%), but it has the dominant experimental PPI recorded in the databases (i.e. 237,193; 63.6% of 11 common organisms). This statistical data indicate that current interaction databases are overestimated and have many false-positive recorded PPIs in some organisms (e.g. *Saccharomyces cerevisiae*). Moreover,

these databases are underestimated and incomplete in most organisms (e.g. *Homo sapiens* and *Mus musculus*). Both of the overestimated and underestimated protein interaction data could influence the low reliable construction of protein interactome in a specific organism.

Protein Data Bank (PDB)[31] stores three-dimensional (3D) structure complexes, from which physical interacting domains can be identified to study DDIs and PPIs using comparative modeling[32,33]. As the number of protein structures increases rapidly, some domain-domain interaction databases, such as 3did[34], and iPfam[35], have recently been derived from PDB. Additionally, some methods have utilized template-based methods (i.e. comparative modeling[32] and fold recognition[33]), which search a 3D-complex library to identify homologous templates of a pair of query protein sequences, in order to predict the protein-protein interactions by accessing interface preference, and score query pair protein sequences according to how they fit the known template structures. However, these methods[32,33] are time-consuming to search all possible protein-protein pairs in a large genome-scale database. For example, the possible protein-protein pairs on the UniProt[12] database (4,826,134 sequences) are about 2.33×10^{13}. In addition, these methods are unable to form homologous PPIs to explore the protein-protein evolution for a specific structure template.

In this thesis, we presented the "3D-domain interologs mapping" and "protein complex family" to construct the structure resolved PPI networks across multiple organisms. "3d-domain interolos mapping" is a concept for efficiently enlarging protein interactions annotated through the homologous PPIs with residue-based binding models. We verified the structure resolved PPI networks on Gene Ontology annotations[36] and the architecture of topology (i.e. scale-free network properties). In addition, we also provide the consensus proteins across three networks based on "3D-domain interologs mapping". These consensus proteins are highly related to the essential genes and disease related proteins. We believe that structure resolved PPI networks would provide the insight for understanding the mechanism of

biological processes within a given PPI network.

1-3. Thesis overview

The thesis is organized as follows. In Chapter 2, for efficiently enlarging protein interactions annotated with residue-based binding models, we proposed a new concept "3D-domain interolog mapping" with a scoring system to explore all homologous protein-protein interaction pairs between the two homolog families, derived from a known 3D-structure dimmer (template), across multiple species. Each family consists of homologous proteins which have interacting domains of the template for studying domain interface evolution of two interacting homolog families. The 3D-interologs database records the evolution of protein-protein interactions database across multiple species. Based on "3D-domain interolog mapping" and a template-based scoring function, we infer 173,294 homologous protein-protein interactions by using 1,895 three-dimensional (3D) structure heterodimers to search the UniProt database (4,826,134 protein sequences). The 3D-interologs database comprises 15,124 species and 283,980 protein-protein interactions, including 173,294 interactions (61%) and 110,686 interactions (39%) summarized from the IntAct database. For a protein-protein interaction, the 3D-interologs database shows functional annotations (e.g. Gene Ontology), interacting domains and binding models (e.g. hydrogen-bond interactions and conserved residues). Additionally, this database provides couple-conserved residues and the interacting evolution by exploring the interologs across multiple species. Experimental results reveal that the proposed scoring function obtains good agreement for the binding affinity of 275 mutated residues from the ASEdb. The precision and recall of our method are 0.52 and 0.34, respectively, by using 563 non-redundant heterodimers to search on the Integr8 database[30] (549 complete genomes). Experimental results demonstrate that the proposed method can infer reliable physical protein-protein interactions and be useful for studying the

protein-protein interaction evolution across multiple species. In addition, the top-ranked strategy and template interface score are able to significantly improve the accuracies of identifying protein-protein interactions in a complete genome.

In Chapter 3, we presented the PCFamily server to identify template-based homologous protein complexes (called protein complex family) and infer functional modules of the query proteins. This server first finds homologous structure complexes of the query using BLASTP to search the structural template database (11,263 complexes). PCFamily then searches the homologous complexes of the templates (query) from a complete genomic database (Integr8 with 6,352,363 protein sequences in 2,274 species). According to these homologous complexes across multiple species, this sever infers binding models (e.g. hydrogen bonds and conserved amino acids in the interfaces), functional modules, and the conserved interacting domains and Gene Ontology annotations of the protein complex family. Experimental results demonstrate that the PCFamily server can be useful for binding model visualizations and annotating the query proteins. We believe that the server is able to provide valuable insights for determining functional modules of biological networks across multiple species.

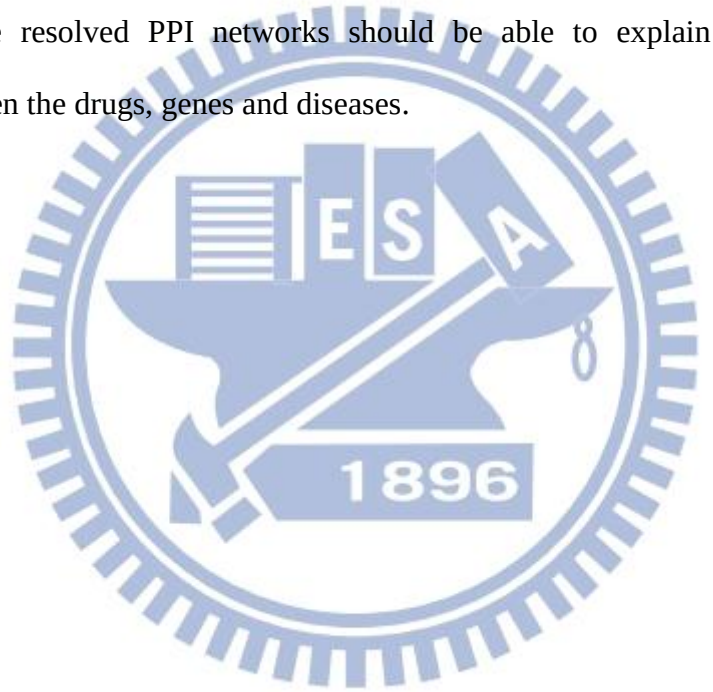
In chapter 4, we provide the structure resolved PPI networks across multiple species, including *H. sapiens*, *M. musculus*, and *D. rerio*. According to structure-based homologous PPIs in multiple species, the PPIs with atomic residue-based binding models in the derived structure resolved network achieved highly agreement with Gene Ontology (BP, CC, and MF terms) similarities. Furthermore, the architecture of these networks is a scale-free network which is consistent with most of the cellular networks. In addition, our derived networks can be used to observe the consensus proteins and modules (a fundamental unit forming with highly connected proteins) which are high conserved appearing in multiple organisms. These consensus proteins are often the essential genes and related to diseases recorded in OMIM. Experimental results also indicate that the mutations of interacting residues on the PPIs often

related to diseases are often on. Our results demonstrate that the structure resolved PPI networks can provide valuable insights for understanding the mechanisms of biological processes.

In chapter 5, we provide a method to characterize a given PPI network. Although, many graphic features have been purposed to measure the role of proteins and identify local modularity structures of high connectivity in a PPI network, the pseudoinverse of the Laplacian matrix plays a key role, has a nice interpretation in terms of random walk on a network, and defines the kernels on a given network. Therefore, we proposed the modularity structure matrix (*MS-matrix*), which is the pseudoinverse of the Laplacian matrix for a given network, to evaluate the modularity structure properties of a PPI network. According to our knowledge, the *MS-matrix* is the first property to identify both global important proteins and local density regions within a network. For a given PPI network of *S. cerevisiae*, our results demonstrate that the important proteins identified by the *MS-matrix* are related to the essential biological processes (i.e. essential genes) and highly consistence with the topology features (i.e. degree, closeness centrality, and betweenness centrality). Then, the relationship between proteins derived from the *MS-matrix* could reflect the similarity of Gene Ontology and could be useful for the module identification. Furthermore, biological characterization (e.g. Gene Onotology) of the modules derived from the *MS-matrix* is similar to the modules collected from the experiment database (e.g. MIPS). Our results demonstrate that the *MS-matrix* would provide the insight for investigating a PPI network through important proteins and local modularity structures.

In the final chapter, we summarized the results of this thesis, and then discuss the future works. To further investigate the behavior of PPI network within a given cell, gene expression data would provide an aspect of in-depth understanding of the dynamic organization of the PPI network and its role in the regulation of cellular processes. For example, the Connectivity Map

(also known as cmap) provided by Lamb, J. *et al.* is a collection of genome-wide transcriptional expression data from cultured human cells treated with bioactive small molecules and simple pattern-matching algorithms that together enable the discovery of functional connections between drugs, genes and diseases through the transitory feature of common gene-expression changes [37]. Therefore, we will combine the gene expression data into the PPI network. We will try to illustrate the behavior of PPI networks under different cell types and different conditions. For example, because the Connectivity Map could provide the up-regulated and down-regulated proteins of given drugs and diseases, combining these data with our structure resolved PPI networks should be able to explain the mechanism of relationship between the drugs, genes and diseases.



Chapter 2. 3D-interologs: An evolution database of physical protein-protein interactions across multiple genomes

Interactions between proteins are critical to most biological processes. To identify and characterize protein-protein interactions (PPIs) and their networks, many high-throughput experimental approaches, such as yeast two-hybrid screening, mass spectroscopy, and tandem affinity purification, and computational methods (phylogenetic profiles[38], known 3D complexes[39], and interologs[40]) have been proposed[41]. Some PPI databases, such as IntAct[8], BioGRID[10], DIP[9], MIPS[28], and MINT[29], have accumulated PPIs submitted by biologists, and those from mining literature, high-throughput experiments, and other data sources. As these interaction databases continue growing in size, they become increasingly useful for analysis of newly identified interactions.

The discovery of sequence homologs to a known protein often provides clues for understanding the function of a newly sequenced gene. As an increasing number of reliable PPIs become available, identifying homologous PPIs should be useful to understand a newly determined PPI. Recently, several PPI databases (e.g., IntAct and BioGRID) allow users to input one or a pair of proteins or gene names to acquire the PPIs associated with the query protein(s). Few computational methods[42,43] applied homologous interactions to assess the reliability of PPIs.

To address this issue, we proposed the concept called "homologous protein-protein interaction"[23]. We define a homologous PPI as follows: (1) homologs of A and B are proteins with significant sequence similarity BLASTP *E*-values $\leq 10^{-10}$ [40,44]; (2) significant joint sequence similarity ($J_E \leq 10^{-40}$) between two pairs, *i.e.*, (A, A_1') and (B, B_1'), of the query

protein pair (A and B) and their respective homologs (A_1' and B_1') recorded in annotated PPI databases. In addition, we constructed the PPISearch server for searching homologous PPIs across multiple species and annotating the query protein pair. According to our knowledge, PPISearch is the first public server that identifies homologous PPIs from annotated PPI databases and infers transferability of interacting domains and functions between homologous PPIs and the query. Our results demonstrate that this server achieves high agreements on interacting domain-domain pairs and function pairs between query protein pairs and their respective homologous PPIs.

Furthermore, a known 3D structure of interacting proteins provides interacting domains and atomic details for thousands of direct physical interactions. It is usually possible to build the binding model of a protein-protein interaction by comparative modeling if a known complex structure comprising homologs of these two sequences is available[32,45]. Therefore, we developed a new scoring function[39], which includes the contact residue interacting score (e.g. the steric, hydrogen bonds, and electrostatic interactions) and the template consensus score (e.g. couple-conserved residue and the template similarity scores), to evaluate how well the interfaces between the query and interacting candidates.

For efficiently enlarging protein interactions annotated with residue-based binding models, we proposed a new concept "3D-domain interolog mapping" with a scoring system[39] to explore all possible homologous protein-protein interaction pairs between the two homolog families, derived from a known 3D-structure dimmer (template), across multiple species. Each family consists of homologous proteins which have interacting domains of the template for studying domain interface evolution of two interacting homolog families.

The 3D-interologs database records the evolution of protein-protein interactions database across multiple species. Based on “3D-domain interolog mapping” and a new scoring function, we infer 173,294 homologous protein-protein interactions by using 1,895 three-dimensional

(3D) structure heterodimers to search the UniProt database (4,826,134 protein sequences). The 3D-interologs database comprises 15,124 species and 283,980 protein-protein interactions, including 173,294 interactions (61%) and 110,686 interactions (39%) summarized from the IntAct database. For a protein-protein interaction, the 3D-interologs database shows functional annotations (e.g. Gene Ontology), interacting domains and binding models (e.g. hydrogen-bond interactions and conserved residues). Additionally, this database provides couple-conserved residues and the interacting evolution by exploring the interologs across multiple species. Experimental results reveal that the proposed scoring function obtains good agreement for the binding affinity of 275 mutated residues from the ASEdb. The precision and recall of our method are 0.52 and 0.34, respectively, by using 563 non-redundant heterodimers to search on the Integr8 database (549 complete genomes).

Experimental results demonstrate that the proposed method can infer reliable physical protein-protein interactions and be useful for studying the protein-protein interaction evolution across multiple species. In addition, the top-ranked strategy and template interface score are able to significantly improve the accuracies of identifying protein-protein interactions in a complete genome. The 3D-interologs database is available at http://3D-interologs.life.nctu.edu.tw.

2-1. Introduction

A major challenge of post genomic biology is to understand the networks of interacting genes, proteins and small molecules that produce biological functions. The large number of protein interactions [8,9,28], generated by large-scale experimental methods [26,46,47], computational methods [32,38,39,44,48-50], and integrated approaches [51,52], provides opportunities and challenges in annotating protein functions, protein-protein interactions (PPI) and domain-domain interactions

(DDI), and in modeling the cellular signaling and regulatory networks. An approach based on evolutionary cross-species comparisons, such as PathBLAST [53,54] and interologs (i.e. interactions are conserved across species [40,44]), is a valuable framework for addressing these issues. However, these methods often cannot respond how a protein interacts with another one across multiple species.

Protein Data Bank (PDB) [31] stores three-dimensional (3D) structure complexes, from which physical interacting domains can be identified to study DDIs and PPIs using comparative modeling [32,33]. Some DDI databases, such as 3did [34], and iPfam [35], have recently been derived from PDB. Additionally, some methods have utilized template-based methods (i.e. comparative modeling [32] and fold recognition [33]), which search a 3D-complex library to identify homologous templates of a pair of query protein sequences, in order to predict the protein-protein interactions by accessing interface preference, and score query pair protein sequences according to how they fit the known template structures. However, these methods [32,33] are time-consuming to search all possible protein-protein pairs in a large genome-scale database (Fig. 2-1A). For example, the possible protein-protein pairs on the UniProt database (4,826,134 sequences) are about 2.33×10^{13} [12]. In addition, these methods are unable to form homologous PPIs to explore the protein-protein evolution for a specific structure template.

To address these issues, we proposed a new concept "3D-domain interolog mapping" (Fig. 2-1B): for a known 3D-structure complex (template T with chains A and B), domain *a* (in chain A) interacts with domain *b* (in chain B) in one species. Homolog families A' and B' of A and B are proteins, which are significant sequence similarity BLASTP *E*-values $\leq10^{-10}$ and contain domains *a* and *b*, respectively. All possible protein pairs between these two homolog families are considered as protein-protein interaction candidates using the template T. Based on this concept, protein sequence databases can be searched to predict protein-protein interactions across multiple species efficiently. When the genome was deciphered completely for a species,

we considered the rank of protein-protein interaction candidates in each species into our previous scoring system [39] to reduce a large number of false positives. The 3D-interologs database which can indicate interacting domains and contact residues in order to visualize molecular details of a protein-protein interaction. Additionally, this database can provide couple-conserved residues and evolutionary clues of a query sequence and its partners by examining the interologs across multiple species.

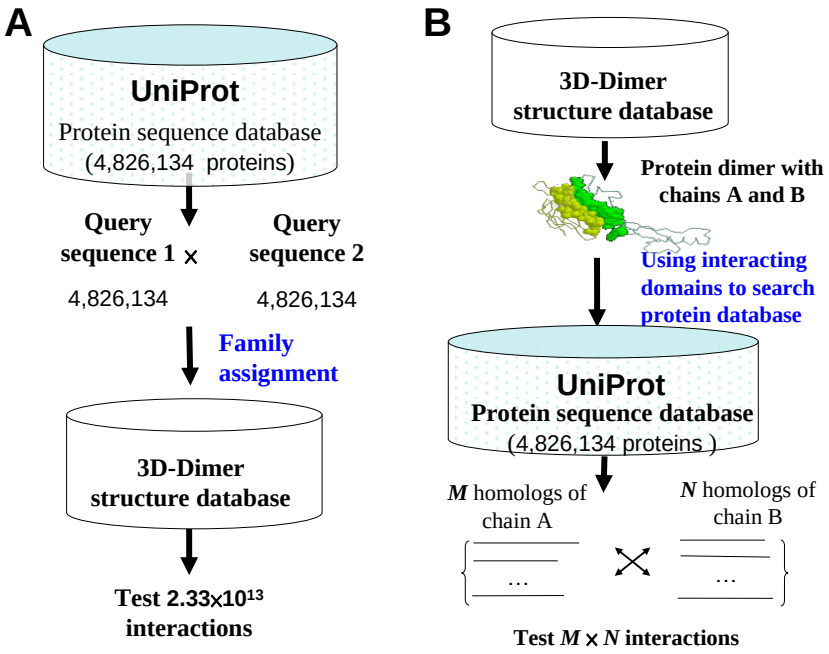


Figure 2-1. Two frameworks of template-based methods for protein-protein interactions (PPI).
(A) For each query protein sequence pair, the method searches 3D-dimer template library to identify homologous templates for exploring the query protein pair, such as MULTIPROSPECTOR [33]. (B) For each structure in 3D-dimer template library, the method searches protein sequence database to identify homologous PPIs of the query structure, such as 3D-interologs.

2-2. Methods and Materials

Figure 2-2 illustrates the overview of the 3D-interologs database. The 3D-interologs allows users to input the UniProt accession number (UniProt AC [12]) or the sequence with FASTA format of the query protein (Fig. 2-2A). When the input is a sequence, 3D-interologs uses BLAST to identify the hit interacting proteins. We identified protein-protein interactions

in 3D-interologs database through structure complexes and a new scoring function using the following steps (Fig. 2-2B). First, a 3D-dimer template library comprising 1,895 heterodimers (3,790 sequences, called NR1895) was selected from the PDB released in Feb 24, 2006. Duplicate complexes, defined by sequence identity of above 98%, were removed from the library. Dimers containing chains shorter than 30 residues were also excluded [33,55]. Interacting domains and contact residues of two chains were identified for each complex in the 3D-dimer library. Contact residues, in which any heavy atoms should be within a threshold distance of 4.5 Å to any heavy atoms of another chain, were regarded as the core parts of the 3D-interacting domains in a complex. Each domain was required to have at least 5 contact residues and more than 25 interacting contacted-residue pairs to ensure that the interface between two domains was reasonably extensive. After the interacting domains were determined, its SCOP domains [22] were identified, and its template profiles were constructed by PSI-BLAST. PSI-BLAST was adopted to search the domain sequences against the UniRef90 database [12], in which the sequence identity < 90% of each other and the number of iteration was set to 3.

After 3D-dimer template library and template profiles were built, we inferred candidates of interacting proteins by 3D-domain interolog mapping. To identify the interacting-protein candidates against protein sequences in the UniProt version 11.3 (containing 4,826,134 protein sequences), the chain profile was used as the initial position-specific score matrix (PSSM) of PSI-BLAST in each template consisting of two chains (e.g. C_A and C_B, Fig. 2-2C). The number of iterations was set to 1. Therefore, this search procedure can be considered as a profile-to-sequence alignment. A pairing-protein sequence (e.g. S1 and S2) was considered as a protein-protein interaction candidate if the sequence identity exceeded 30% and the aligned contact residue ratio (*CR*) was greater than 0.5 for both alignments (i.e. S1 aligning to C_A and S2 aligning to C_B). For each interacting candidate, the scoring function was applied to calculate the interacting score and the *Z*-value, which indicates the statistical significance of the

interacting score. An interacting candidate was regarded as a protein-protein interaction if its *Z*-value was above 3.0 and it ranked in the Top 25 in one species. The candidate rank was considered in one species to reduce the ill-effect of the out-paralogs that arose from a duplication event before the speciation [56]. These inferred interacting protein pairs were collected in the database.

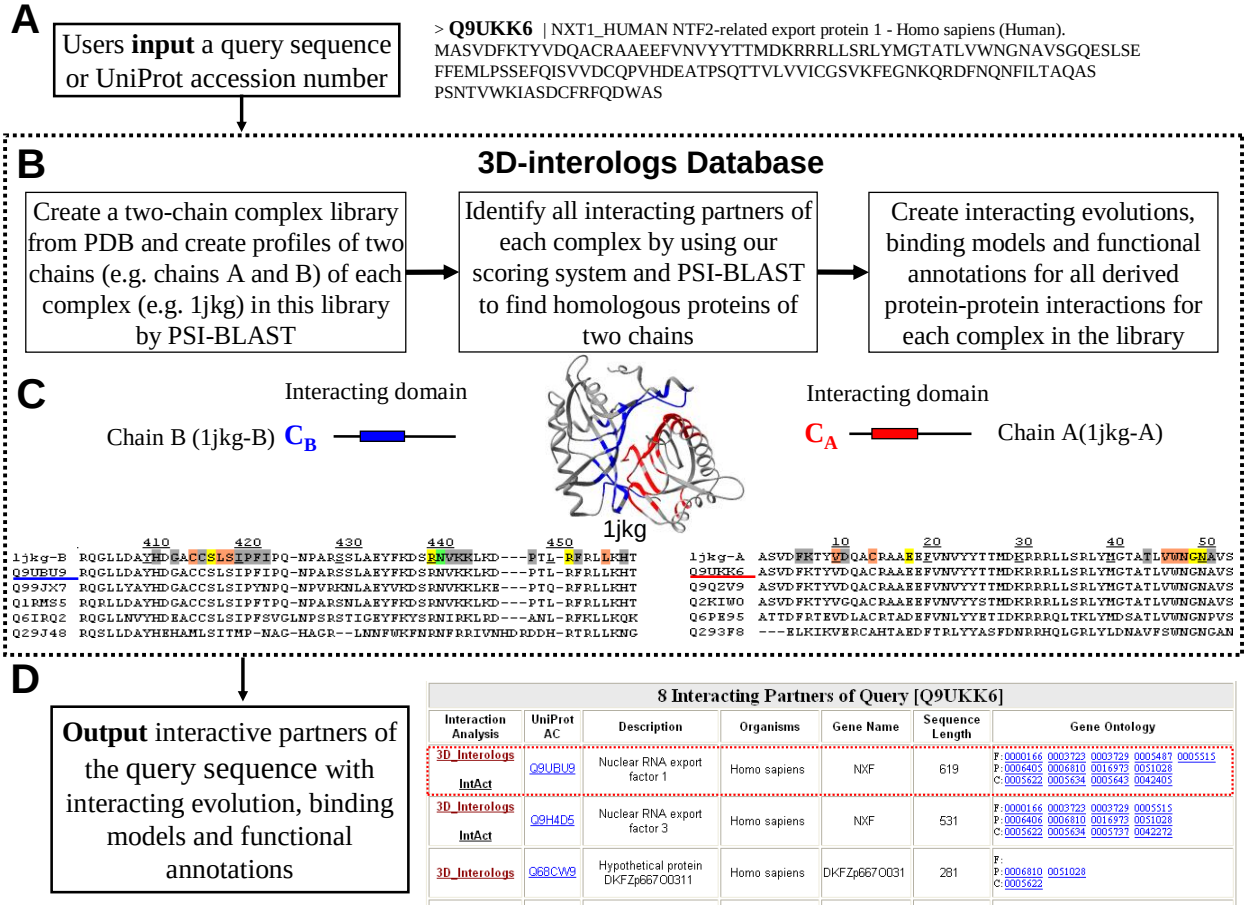


Figure 2-2. Overview of the 3D-interologs database for protein-protein interacting evolution, protein functions annotations and binding models across multiple species.

Finally, for the hit interacting partner derived from 3D-domain interolog mapping, this database provides functional annotations (e.g. UniProt AC, organism, descriptions, and Gene Ontology (GO) annotations [36], Fig. 2-2D), and the visualization of the binding models and interaction evolutions (Fig. 2-2C) between the query protein and its partners. We then constructed two multiple sequence alignments of the query protein and its interacting partner

(Fig. 2-2C) across multiple species. Here, the interacting-protein pair with the highest Z-score in a species was chosen as interologs for constructing multiple sequence alignments using a star alignment. The chains (e.g. Chains A and B, Fig. 2-2C) of the hit structure template were considered as the centers, and all selected interacting-protein pairs across species were aligned to respective chains of the template by PSI-BLAST. The 3D-interologs database annotates the important contact residues in the interface according to the following formats: hydrogen-bond residues (green); conserved residues (orange), conserved residues with hydrogen bonds (yellow) and other (gray).

Data Sets

Two data sets were used to assess 3D-domain interolog mapping and the scoring functions. To determine the contribution of a residue to the binding affinity, the alanine-scanning mutagenesis is frequently used as an experimental probe. We selected 275 mutated (called BA-275) residues from the ASEdb [57] with 16 heterodimers whose 3D structures were known. Those mutated residues are contact residues and positioned at protein–protein interfaces. ASEdb gives the corresponding delta G value representing the change in free energy of binding upon mutation to alanine for each experimentally mutated residue. Residues that contribute a large amount of binding energy are often labeled as hot spots.

In addition, we selected a non-redundant set (NR-563), comprising 563 dimer protein structures from the set NR1895 to evaluate the performance of our scoring functions for predicting PPIs in *S. cerevisiae* and in 549 species collected in Integr8 database (2,102,196 proteins [30]).

2-3. Scoring Function and Matrices

We have recently proposed a scoring function to determine the reliability of a

protein-protein interaction [39]. This study enhances this scoring by dividing the template consensus score into the template similar score and the couple-conserved residue score. Based on this scoring function, the 3D-interologs database can provide the interacting evolution across multiple species and the statistical significance (*Z*-value), the binding models and functional annotations between the query protein and its interacting partners. The scoring function is defined as

$$E_{tot} = E_{vdw} + E_{SF} + E_{sim} + wE_{cons} \quad (1)$$

where E_{vdw} and E_{SF} are the interacting van der Waals energy and the special interacting bond energy (i.e. hydrogen-bond energy, electrostatic energy and disulfide-bond energy), respectively; and E_{sim} is the template interface similar score; and the E_{cons} is couple-conserved residue score. The optimal *w* value was yielded by testing various values ranging from 0.1 to 5.0; *w* is set to 3 for the best performance and efficiency on predicting binding affinity (BA-275) and predicting PPIs in *S. cerevisiae* and in 549 species (Integr8) using the data set NR-563. The E_{vdw} and E_{SF} are given as

$$E_{vdw} = \sum_{i,j}^{CP} (Vss_{ij} + Vsb_{ij} + Vsb_{ji})$$

$$E_{SF} = \sum_{i,j}^{CP} (Tss_{ij} + Tsb_{ij} + Tsb_{ji})$$

where *CP* denotes the number of the aligned-contact residues of proteins *A* and *B* aligned to a hit template; Vss_{ij} and Vsb_{ij} (Vsb_{ji}) are the sidechain-sidechain and sidechain-backbone van der Waals energies between residues *i* (in protein *A*) and *j* (in protein *B*), respectively. Tss_{ij} and Tsb_{ij} (Tsb_{ji}) are the sidechain-sidechain and sidechain-backbone special interacting energies between *i* and *j*, respectively, if the pair residues *i* and *j* form the special bonds (i.e. hydrogen bond, salt bridge, or disulfide bond) in the template structure. The van der Waals energies (Vss_{ij}, Vsb_{ij}, and Vsb_{ji}) and special interacting energies (Tss_{ij}, Tsb_{ij}, and Tsb_{ji}) were calculated from the four knowledge-based scoring matrices (Fig. 2-3), namely sidechain-sidechain (Fig. 2-3A) and sidechain-backbone van der Waals scoring matrices (Fig. 2-3B); and sidechain-sidechain (Fig.

2-3C) and sidechain-backbone special-bond scoring matrices (Fig. 2-3D).

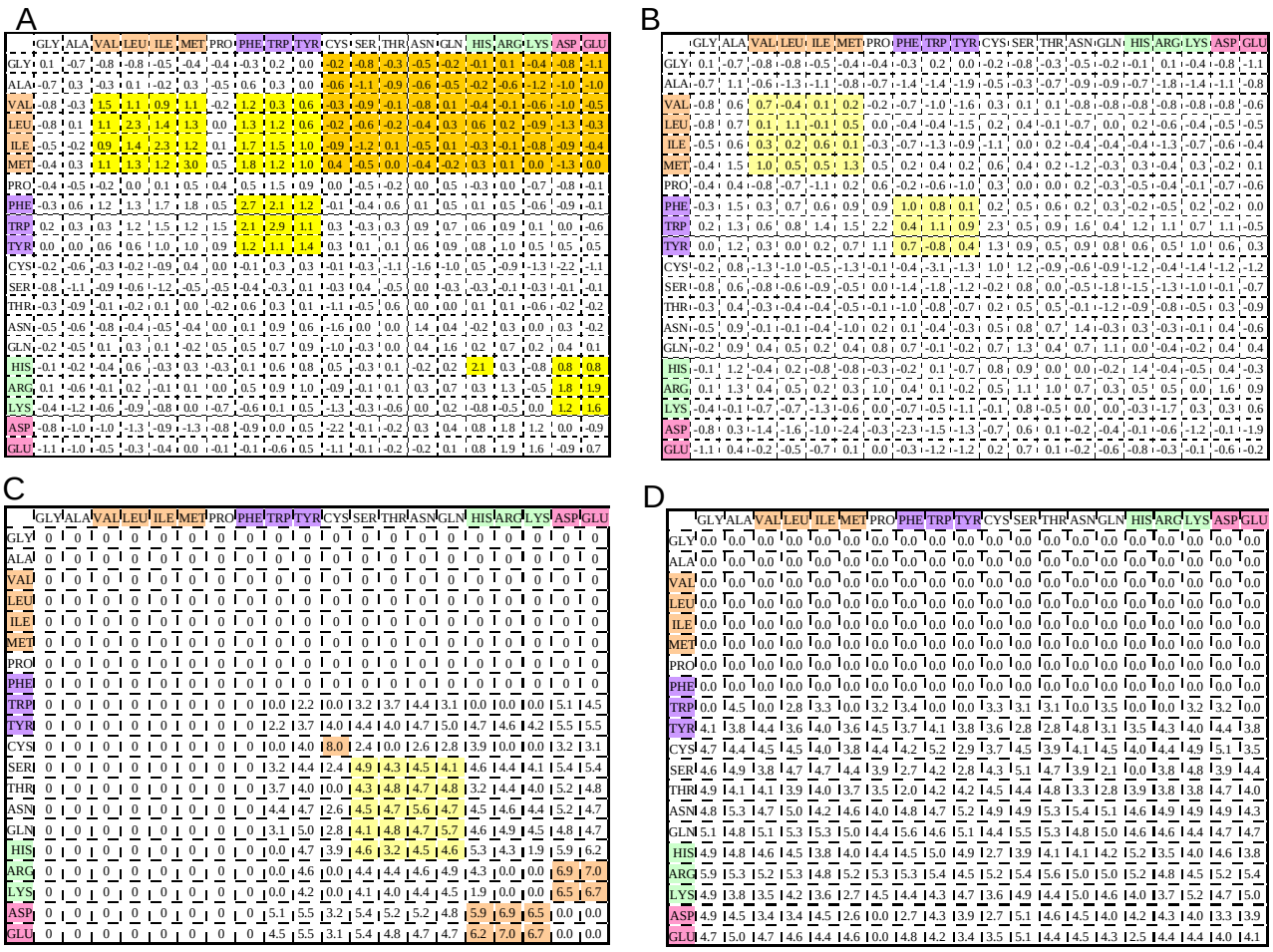


Figure 2-3. Knowledge-based protein-protein interacting scoring matrices: (A) sidechain-sidechain van-der Waals scoring matrix; (B) sidechain-backbone van-der Waals scoring matrix; (C) sidechain-sidechain special-bond scoring matrix; (D) sidechain-backbone special-bond matrix scoring.

The sidechain-sidechain scoring matrices are symmetric and sidechain-backbone scoring matrices are non-symmetric. For sidechain-sidechain van-der Waals scoring matrix, the scores are high (yellow blocks) if large-aliphatic residues (i.e. Val, Leu, Ile, and Met) interact to large-aliphatic residues or aromatic residues (i.e. Phe, Tyr, and Trp) interact to aromatic residue. In contrast, the scores are low (orange blocks) when nonpolar residues interact to polar residues. For sidechain-sidechain special-bond scoring matrix, the scores are high when an interacting resides (i.e. Cys to Cys) form a disulfide bond or basic residues (i.e. Arg, Lys, and His) interact to acidic residues (Asp and Glu). The scoring values are zero if nonpolar residues interact to other residues.

These four knowledge-based matrices, which were derived using a general mathematical structure [58] from a nonredundant set of 621 3D-dimer complexes proposed by Glaser *et al.* [59], are the key components of the 3D-interologs database for predicting protein-protein interactions. This dataset is composed of 217 heterodimers and 404 homodimers and the

sequence identity is less than 30% to each other. The entry (S_{ij}), which is the interacting score for a contact residue *i*, *j* pair ($1 \le i, j \le 20$), of a scoring matrix is defined as $S_{ij} = \ln \frac{q_{ij}}{e_{ij}}$, where q_{ij} and e_{ij} are the observed probability and the expected probability, respectively, of the occurrence of each *i*, *j* pair. For sidechain-sidechain van-der Waals scoring matrix, the scores are high (yellow blocks) if large-aliphatic residues (i.e. Val, Leu, Ile, and Met) interact to large-aliphatic residues or aromatic residues (i.e. Phe, Tyr, and Trp) interact to aromatic residue. In contrast, the scores are low (orange blocks) when nonpolar residues interact to polar residues. The top two highest scores are 3.0 (Met. interacting to Met) and 2.9 (Trp interacting to Trp).

The value of E_{sim} was calculated from the BLOSUM62 matrix [58] based on two alignments between two chains (A and B) of the template and their homologous proteins (A' and B'), respectively. The E_{sim} is defined as

$$E_{sim} = \sum_{i,j}^{CP} \sqrt{\frac{K_{ii'} \times K_{jj'}}{K_{ii} \times K_{jj}}} \qquad (2)$$

where *CP* is the number of contact residue pairs in the template; *i* and *j* are the contact residue in chains A and B, respectively. $K_{ii'}$ is the score of aligning residue *i* (in chain A) to *i′* (in protein A') and $K_{jj'}$ is the score of aligning residue *j* (in chain B) to *j′* (in protein B') according to BLOSUM62 matrix. K_{ii} and K_{jj} are the diagonal scores of BLOSUM62 matrix for residues *i* and *j*, respectively. The couple-conserved residue score (E_{cons}) was determined from two profiles of the template and is given by

$$E_{cons} = \sum_{i,j}^{CP} (\max(0, (M_{ip} - K_{ii}) + (M_{jp'} - K_{jj})) \qquad (3)$$

where *CP* is the number of contact residue pairs; M_{ip} is the score in the PSSM for residue type *i* at position *p* in Protein *A*; $M_{jp'}$ is the score in the PSSM for residue type *j* at position *p′* in Protein *B*, and K_{ii} and K_{jj} are the diagonal scores of BLOSUM62 matrix for residue types *i* and *j*, respectively.

To evaluate statistical significance (Z-value) of the interacting score of a protein-protein interaction candidate, we randomly generated 10,000 interfaces by mutating 60% contact residues for each heterodimer in 3D-dimer template library. The selected residue was substituted with another amino acid residue according to the probability derived from these 621 complexes [59]. The mean and standard deviation for each 3D-dimer were determined from these 10,000 random interfaces which are assuming to form a normal distribution. Based on the mean and standard deviation, the Z-value of a protein-protein candidate predicted by this template can be calculated.

2-4. Inputs and Outputs

The 3D-interologs database server is easy-to-use. Users input the UniProt AC or the FASTA format of the query protein (Fig. 2-2A). The server generally returns a list of interacting partners with functional annotations (e.g. the gene name, the protein description and GO annotations) (Fig. 2-2D) and provides the visualization of the binding model and contact residues between the query protein and its partner by aligning them to respective template sequences and structures. Additionally, the 3D-interologs system indicates the interacting evolution analysis by using multiple sequence alignments of the interologs across multiple species (Fig. 2-2C). The significant contact residues in the interface are indicated. If Java is installed in the user's browser, then the output shows the structures, and users can dynamically view the binding model, interacting domains and important residues in the browser.

2-5. Example Analysis

Figure 2-4 show the search results using the human protein NXT1 (UniProt AC Q9UKK6) as the query sequence. The NXT1, which is a nucleocytoplasmic transport factor and shuttles

between the nucleus and cytoplasm, accumulates at the nuclear pore complexes[60]. For this query, 3D-interologs database yielded 8 hit interacting partners (Fig. 2-4A), comprising 5 partners derived from 3D-interologs database and 5 partners from the IntACT database. Thus, two partners were present in both databases. Among these 8 hits, 3 partners (i.e. Uniprot AC Q68CW9, Q5H9I1 and Q9GZY0) were not recorded in IntAct database, but they very likely interact with NXT1. The Q68CW9, which is part of the protein NXF1 (UniProt AC Q9UBU9), consists of the UBA-like domain and the NTF-like domain, which is responsible for association with the protein NXT1 [61]. The sequence of the protein Q5H9I1 is the same as that of the protein Q9H4D5 (i.e. nuclear RNA export factor 3), which binds to NXT1 [62]. The protein Q9GZY0 (nuclear RNA export factor 2) binds protein NXT1 to export mRNA cargoes from nucleus into cytosol [63].

The protein NXT1 interacts with the protein NXF1 to form a compact heterodimers (PDB code 1jkg [63])and an interacting β surface, which is lined with hydrophobic and hydrophilic residues (Fig. 2-4B). Twenty hydrogen bonds or electrostatic interactions are formed in this compact interface. The salt bridge formed by NXT1 Arg134 and NXF1 Asp482 is especially important in the interface [57]. The interacting evolution analysis built by 10 interologs reveals that two residues (Arg134 and Asp482) are conserved in all species (Fig. 2-4C). Additionally, some interacting residues forming the hydrogen bonds are also couple-conserved, for example NXT1 Asp76 and NXF1 Arg440; NXT1 Gln78 and NXF1 Ser417; NXT1 Pro79 and NXF1 Asn531 [57]. The evolution of interaction is valuable to reflect both couple-conserved and critical residues in the binding site.

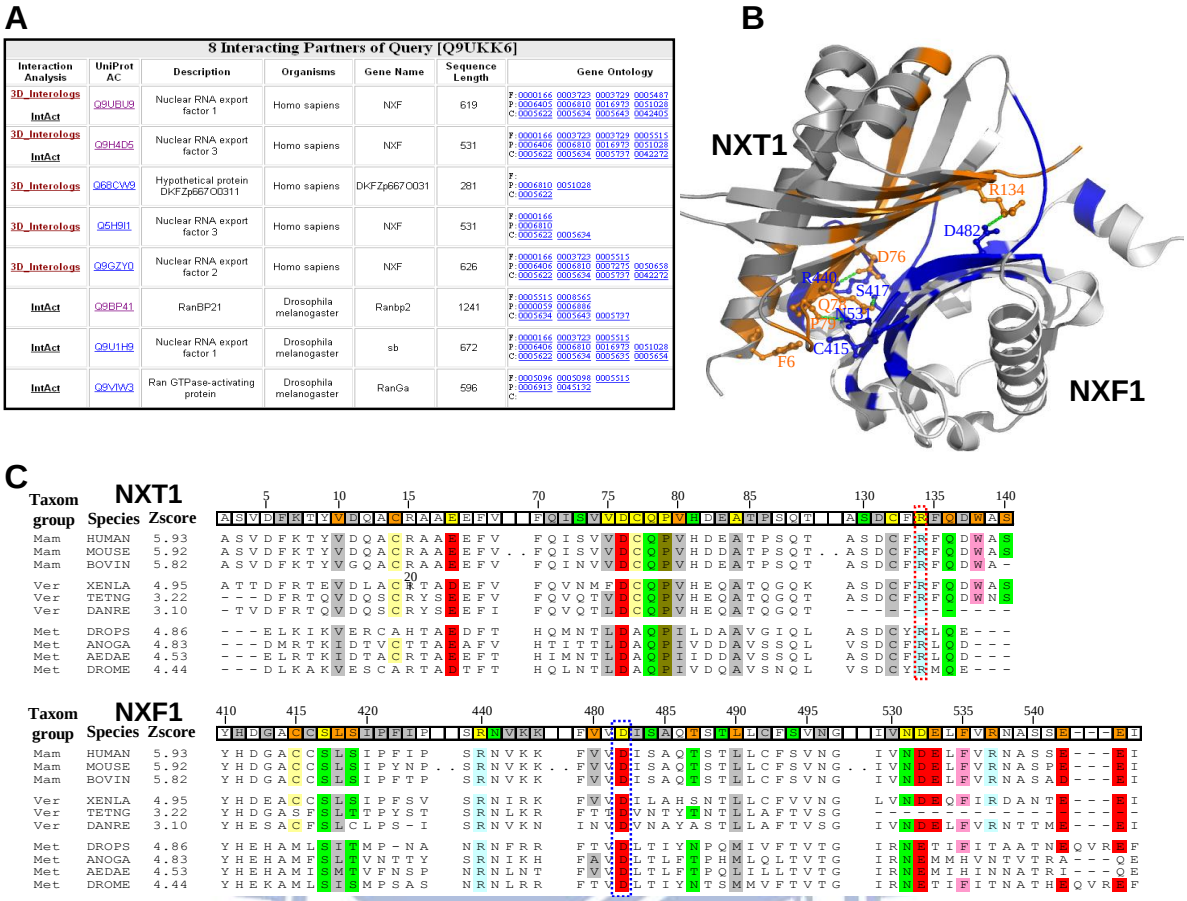


Figure 2-4. The 3D-interologs database search results of using human NXT1 as query.

(A) Eight interacting partners of NXT1 are found in the 3D-Interologs. For each interacting partner, this server provides UniProt accession number, protein description, organism and Gene Ontology annotation. (B) Detailed interactions between the query and its interacting partner (UniProt accession number Q9UBU9) are indicated via the structure template which consists of NXT1 (PDB entry 1jkg-A) and NXF1 (PDB entry 1jkg-B). The contact residues of NXT1 (query side) and NXF1 (partner side) are colored by red and blue, respectively. The contact residues forming hydrogen bonds (green and dash) are given the atom details. (C) The interacting evolution analysis by using multiple sequence alignments of hit interacting partners of the query across multiple species. The 3D-interologs yields 10 interologs of the query template structure. The contacted residues are marked in template structure based on their interacting characteristics, including hydrogen-bond residues (green); conserved residues (orange); both (yellow), and others (gray). The couple-conserved contact positions are colored in the multiple alignments according to the physical-chemical property of amino acid residues. Twenty amino acid types are classified into 7 groups, namely polar positive (His,Arg, and Lys, blue); polar negative (Asp and Glu, red); polar neutral (Ser, Thr, Asn and Gln, green); cystein (yellow); non-polar aliphatic (Ala, Val, Leu, Ile and Met, gray); non-polar aromatic (Phe, Tyr and Trp, pink); and others: (Gly and Pro, brown).

Conversely, some positions, which are not conserved in all species but conserved in an individual taxonomic group, are important for observing the co-evolution across multiple

species. The interacting residue pair (NXT1 Phe6 and NXF1 Cys415) in mammalia and vertebrata is different from that in metazoan (NXF1 Cys415→Met and NXT1 Phe→Leu variant). The van-der Waals potential (1.3 in the sidechain-sidechain van-der Waals scoring matrix, Fig. 2-3A) between Leu and Met is much larger than the potential (−0.1) between Cys and Phe. This co-evolution favors the formation of the hydrophobic interaction in metazoan.

2-6. Results

Database

The 3D-interologs database currently contains 15,124 species and 283,980 protein-protein interactions, including 173,294 interactions (61%) derived from our method based on 3D-domain interolog mapping and 110,686 interactions (39%) summarized from the IntAct database[8]. For the hit interacting partner derived from 3D-domain interolog mapping, this database provides functional annotations (e.g. UniProt AC, organism, descriptions, and Gene Ontology (GO) annotations [36]), and the visualization of the binding models and interaction evolutions between the query protein and its partners. On the other hand, the 3D-interologs database presents only the functional annotations of the hit protein-protein interaction if this interaction was summarized from the IntAct database.

Table 2-1. Statistics of 3D-interologs database on 19 species commonly used in research projects

Species	3D-domain interologs	IntAct
Mus musculus	8,876	2,634
Homo sapiens	8,639	18,716
Danio rerio	4,564	0
Xenopus laevis	4,057	58
Rattus norvegicus	3,685	958
Bos taurus	3,549	174
Drosophila melanogaster	2,644	25,036
Arabidopsis thaliana	2,418	2,111
Caenorhabditis elegans	1,433	4,684
Saccharomyces cerevisiae	443	36,821
Escherichia coli	426	14,007
Schizosaccharomyces pombe	371	341
Dictyostelium discoideum	284	84
Zea mays	219	0

Oryza sativa	193	69
Takifugu rubripes	191	0
Chlamydomonas reinhardtii	122	14
Plasmodium falciparum	68	2,707
Pneumocystis carinii	23	0
other species	131,089	2,272
Total	173,294	110,686

Among 15,124 species in the 3D-interologs database, Table 2-1 shows 19 species commonly used in molecular research projects, such as *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Escherichia coli*. To analyze couple-conserved residues and interface evolutions for providing evolutionary clues, the 15,124 species were divided into 10 taxonomic groups [64], namely mammalia, vertebrata, metazoa, invertebrata, fungi, plant, bacteria, archaea, viruses, and others.

Binding Affinity Prediction

The enhanced scoring functions were first evaluated on 275 mutated residues selected from the ASEdb database[57] to reveal the Pearson correlations between ddG values and predicted energies. The 3D-interologs method applied four scoring functions (Fig. 2-5), including 3D-interologs (red), 3D-partner (blue), E_{sim} (only template similarity, green) and one matrix (black) proposed by Lu, *et al.* [33]. Among these four scoring functions, the 3D-interologs is the best (0.92) and one matrix is the worst (0.55, i.e. Lu, *et al.*). The correlations are 0.91 and 0.88 for 3D-partner and 0.88 (only template similarity), respectively.

The binding free energy is often not evenly distributed across interfaces but involves a small subset of "hot spots" contributed extraordinarily high energy [65]. For instance, the human blood-coagulation complex (PDB code 1dan) has 52 residues whose energy contribution was probed by alanine scanning mutagenesis [66,67]. Among these 52 residues, residues Lys-20 and

Asp-58, which are highly conserved in many species, provide the binding free energy upper 2 kcal/mol; on the other hand, the average energy contribution of the other 50 residues is 0.37 kcal/mol. This result implies that the couple-conserved residue score (E_{cons}) is beneficial to model the binding energy of residues positioned in the interfaces. Although the hotspots of protein-protein binding are often for maintaining their function, the antibodies keep the diversity to recognize a wide variation of antigens. The correlation is 0.143 when the E_{cons} was used to model the binding energy of antigen-antibody complexes. Fortunately, integrating E_{cons}, E_{sim} and E_{SF} is able to improve the correlation to 0.606 for antigen-antibody complexes.

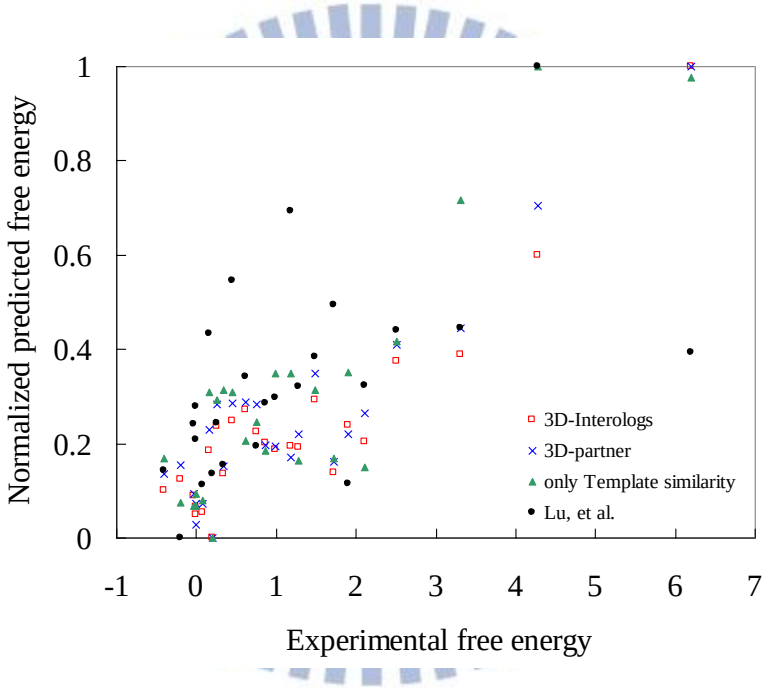


Figure 2-5. Evaluation of the 3D-interologs in binding affinities.

The Pearson correlations between experimental free energies (ddG) and the predicted values of the 3D-interologs using four scoring functions, including 3D-interologs (red), 3D-partner (blue), E_{sim} (only template similarity, green) and one matrix (black) proposed by Lu, *et al.*, on 275 mutated residues selected from Alanine Scanning Energetics database.

Interactions Prediction in *S. cerevisiae*

Additionally, a non-redundant set (NR-563), comprising 563 dimer complexes from the 3D-dimer library, was adopted to evaluate the performance of this enhanced scoring function

for interacting partner predictions in *S. cerevisiae*. This set comprised 5,882 protein-protein interactions, which were recorded as the core subset in the DIP database as the positive cases, and 2,708,746 non-interacting protein pairs, defined by Jansen *et al.* [48] as the negative cases. Figure 2-6A shows the ROC curves of our method and other three scoring functions for predicting PPIs in *S. cerevisiae*. Among these four scoring functions, the 3D-interologs and the template similar score (E_{sim}) were the best and achieved the similar accuracy. Conversely, one matrix (i.e. Lu, *et al.* [33]) was the worst. The average precisions, which was calculated as $(\sum_{i=1}^{A} i/T_h^i)/A$, where T_h^i denotes the number of compounds in a hit list including i correct hits, were 0.84 (3D-interologs), 0.82 (3D-partner), and 0.67 for one matrix (proposed by Lu *et al.*). These results demonstrated that the proposed new scoring function can achieve good agreement for the binding affinity in PPIs and provide statistical significance (Z-value) for predicting PPIs.

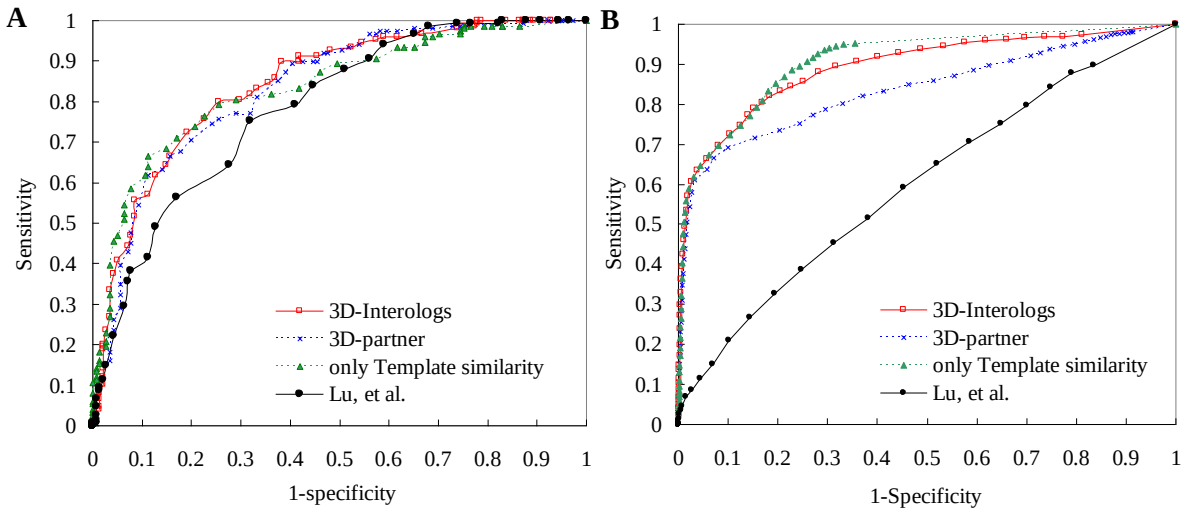


Figure 2-6. The ROC curves of the 3D-interologs for protein-protein interactions.
The 3D-interologs search results on (A) *S. cerevisiae* and (B) 549 species (Intger8) using the data set NR-563 (563 dimer-complex structures) by applying four scoring functions, including 3D-interologs (red), 3D-partner (blue), only template similarity (E_{sim}, green) and one matrix (black) proposed by Lu, *et al.*

Interactions Prediction on Multiple Species

To evaluate the performance of the 3D-domain interolog mapping on multiple species,

563 non-redundant dimer complexes (NR-563) were used as queries to search on the Integr8 database (Release 65) which comprises 2,102,196 proteins in 549 species (Fig. 2-6B and Fig. 2-7). The Integr8 is an integrated database for organisms with completely deciphered genomes, which are mainly obtained from the non-redundant sets of UniProt entries. Experimentally determined protein-protein interactions dataset were collected from IntAct [8] as the gold standard positive set (110,686 interactions). The gold standard negative set was generated according to the assumption that two proteins acting in the same biological process are more likely to interact than two proteins involved in different processes [68]. This study applied the relative specificity similarity (RSS), proposed by Wu *et al.* [69], to measure the biological process similarity and the location similarity of two proteins based on the GO terms of the biological process (BP) and the cellular component (CC), which describes locations at levels of subcellular structures and macromolecular complexes, respectively. Among 110,686 interactions recorded in the IntAct database, 51,049 interactions can be used to calculate the BP and the CC RSS scores. The BP and CC RSS scores of 15.85% and 2.65% interactions, respectively, are less than 0.4. Here, we considered an interacting protein pair as a negative PPI if its CC RSS score is less than 0.4.

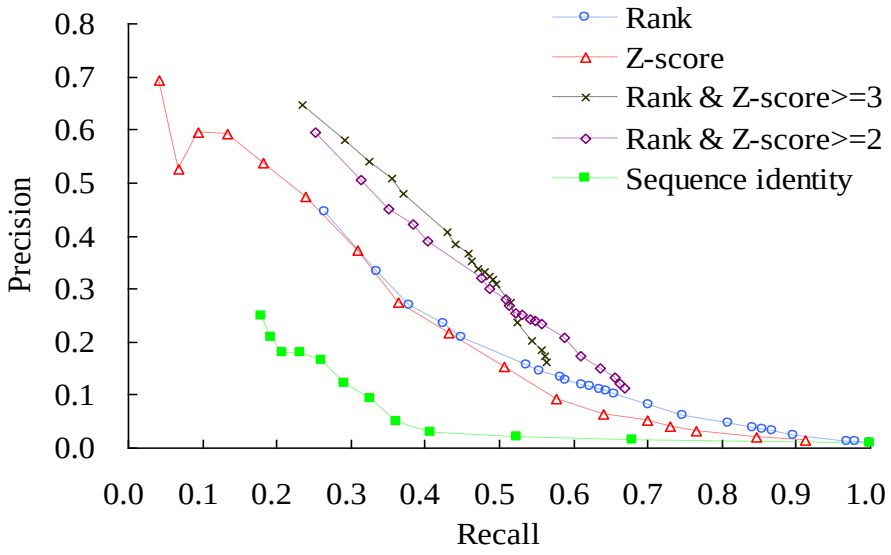


Figure 2-7. Precisions and recalls of 3D-interologs the on Integr8.

The 3D-interologs searches Integr8 database (2,102,196 proteins in 549 species) using the data set NR-563 (563 dimer-complex structures). The 3D-interologs server uses five scoring schemes, including rank in a species (blue), Z-score (red), rank and Z-score >=3 (black), rank and Z-score >=2 (purple), and sequence identity (green).

The structures in the NR-563 as queries to search the Integr8 database yielded 1,063 protein-protein interactions recorded in the IntAct database and 131,831 protein pairs, whose CC RSS scores were less than 0.4 as the negative cases. Based on ROC curves (Fig. 2-6B) for predicting PPIs in 549 species, 3D-interologs and the template similar score (E_{sim}) outperform the 3D-partner server and one-matrix (i.e. Lu, *et al.*) method. In addition, the precision and recall were adopted to access the predicted quality of the 3D-interologs using these four scoring schemes (Fig. 2-7). The precision was defined as $A_h/(A_h+F_h)$, where A_h and F_h denote the numbers of hit positive cases and hit negative cases, respectively. The recall was defined as A_h/A, where A is the total number of positives (here A=1,063). Furthermore, the accuracy of our scoring function (red) is significantly better than that of the sequence identity (green).

The 3D-domain interolog mapping may yield many PPI candidates (e.g. > 200) for one species from a structure template because a eukaryote genome frequently contains multiple paralogous genes. Here, we proposed a top-rank strategy to limit the number of PPIs inferred from a structural template in the same species. For example, we discarded the PPI candidates whose ranks ≥ 25 for a species if the rank threshold is set to 25. Figure 2-7 shows that the performance of the top-rank scores (blue, with different rank thresholds) is similar to that of using Z-score scoring method (red). When we combined the top-rank strategy and the Z-score scoring methods, the precisions (purple and black) are significantly improved. The precision was 0.52 and the recall was 0.34 when Z-score > 3.0 and the rank ≤ 25 in one species.

Adopting the top-rank strategy in one species as the scoring function is useful for distinguishing between positives and negatives when the 3D-domain interolog mapping yielded many protein-protein interactions for one species from a structure template. However,

the rank cannot reflect the binding affinity of a PPI candidate, conversely, the Z-score cannot be adopted to identify the orthologs and in-paralogs arising from a duplication event following the speciation [56]. These results reveal that Z-scores and ranks scoring methods are complementary.

Table 2-2 shows an example for illustrating processes and robustness of combining the top-ranked strategy and Z-score methods. Using human calcineurin heterodimer (PDB code 1aui) structure as query, the 3D-domain interolog mapping yielded 1096 PPI candidates in 38 species if the Z score is set to 2. These 1096 candidates possess the interacting domains (i.e. Metallophos and efand domains) of the query template. Among these PPI candidates, 10 PPIs were recorded in IntACT and 9 candidates were considered as negative PPIs because their CC RSS scores are less than 0.4. The ranks of these 9 negative PPIs are more than 15; conversely, these 10 positive PPIS are top 10 in each species. These observations showed that the top-ranked strategy is useful to dramatically reduce the false positive rate when the 3D-domain interolog mapping for predicting PPIs across multiple complete genomes.

Table 2-2. 3D-interologs search results using human calcineurin heterodimer as the query

Interactor1	Interactor2	Species	Z score	Rank	P / N[a]	RSS of BP[b]	RSS of CC[c]	Interacting domain1	Interacting domain2
P48456	P48451	Fruit fly	8.98	1	P	0.89	0.85	Metallophos	efand
P23287	P25296	Yeast	8.25	1	P	0.88	1.00	Metallophos	efand
P14747	P25296	Yeast	7.95	2	P	0.88	1.00	Metallophos	efand
Q12705	Q9UU93	Yeast	7.94	1	P	-[d]	0.78	Metallophos	efand
P48456	P47948	Fruit fly	4.42	**16**	**N**	0.41	0.30	Metallophos	efand
P48456	P47949	Fruit fly	4.38	**17**	**N**	0.41	0.30	Metallophos	efand
P48456	P49258	Fruit fly	3.99	**23**	P	0.41	0.56	Metallophos	efand
P48456	Q9VQH2	Fruit fly	3.94	**25**	**N**	0.49	0.33	Metallophos	efand
Q8IAM8	P62203	Plasmodium falciparum	3.79	2	P	-	-	Metallophos	efand
P48456	P48593	Fruit fly	3.72	**31**	P	0.35	0.56	Metallophos	efand
P48456	A1ZAE1	Fruit fly	3.59	**34**	**N**	0.00	0.30	Metallophos	efand
Q27889	P48593	Fruit fly	3.42	**40**	P	-	-	Metallophos	efand
P23287	P06787	Yeast	3.36	5	P	0.61	0.88	Metallophos	efand
P48456	Q9VMT2	Fruit fly	3.03	**50**	**N**	0.41	0.30	Metallophos	efand
P48456	Q7K860	Fruit fly	2.99	**53**	**N**	0.41	0.30	Metallophos	efand
P14747	P06787	Yeast	2.86	6	P	0.61	0.88	Metallophos	efand
P48454	Q9NP86	Human	2.33	**90**	**N**	-	0.00	Metallophos	efand
Q08209	Q9NP86	Human	2.31	**91**	**N**	0.41	0.00	Metallophos	efand

P16298	Q9NP86	Human	2.31	**91**	**N**	-	0.00	Metallophos	efand

3D-interologs infers 10 positive and 9 negative protein-protein interactions by human calcineurin heterodimer (PDB code 1aui), including calmodulin-dependent calcineurin A subunit alpha isoform (chain A with interacting domain Metallophos) and calcineurin subunit B type 1 (chain B with interacting domain efhand), searching on Integr8 database.

[a] PPI is a positive (P, recorded in IntACT database) or negative case (N, RSS of cellular component is less than 0.4).

[b,c] The relative specificity similarity (RSS) scores, proposed by Wu *et al*. [69], of Gene Ontology biological process (BP) and cellular component (CC), respectively.

[d] The protein pair is without Gene Ontology annotations in BP or CC.

2-7. Conclusions

This work demonstrates that the 3D-interologs database is robust and feasible for the interacting evolution of PPIs and DDIs across multiple species. This database can provide couple-conserved residues, interacting models and interface evolution through 3D-domain interolog mapping and template-based scoring functions. The scoring function achieves good agreement for the binding affinity in protein-protein interactions. We believe that the 3D-domain interolog mapping should be useful in protein-protein interacting evolution and is able to infer reliable physical protein-protein interactions across multiple genomes.

Chapter 3. PCFamily: a web server for searching homologous protein complexes

The proteins in a cell often assemble into complexes to carry out their functions and play an essential role of biological processes. The PCFamily server identifies template-based homologous protein complexes (called protein complex family) and infers functional modules of the query proteins. This server first finds homologous structure complexes of the query using BLASTP to search the structural template database (11,263 complexes). PCFamily then searches the homologous complexes of the templates (query) from a complete genomic database (Integr8 with 6,352,363 protein sequences in 2,274 species). According to these homologous complexes across multiple species, this sever infers binding models (e.g. hydrogen bonds and conserved amino acids in the interfaces), functional modules, and the conserved interacting domains and Gene Ontology annotations of the protein complex family. Experimental results demonstrate that the PCFamily server can be useful for binding model visualizations and annotating the query proteins. We believe that the server is able to provide valuable insights for determining functional modules of biological networks across multiple species. The PCFamily sever is available at http://pcfamily.life.nctu.edu.tw.

3-1. Introduction

Protein complexes are fundamental units of macromolecular organization and their composition is also known to vary according to cellular requirements [7]. To identify and characterize the protein complexes, genome-scale interaction discovery approaches, such as two-hybrid system or affinity purification [70,71], have been proposed. However, these methods are often unable to respond how a protein interacts with others. Based on increasing

protein-protein interactions (PPI) [8,9,28,29] and structure complexes [31], previous studies have suggested that the total number of protein-protein interaction types are limited (~10,000 types) [72] and the quaternary structures (QS) can be clustered into 3,151 QS families [73].

A known three-dimensional (3D) structure complex provides physical protein interaction topology, interacting domains, and atomic detailed binding models of interactions. Recently, some studies utilized template-based methods (i.e. comparative modeling [32] and fold recognition [33]), which search a 3D-complex library to model a large set of yeast complexes [45,74]. These methods are time-consuming to search all possible homologous PPIs or complexes, which are useful to explore interface evolutions of a specific 3D structure complex, from a large complete genomic database (e.g. Integr8) with many species [30].

To address these issues, we numerously enhanced and modified both PPI family search (sequence-based PPI search method [23]) and 3D-domain interologs with template-based scoring function (3D-template PPI prediction method [39]). According to our knowledge, PCFamily is the first public server that identifies homologous complexes (≥ two proteins) and module evolution of the query. For a set of query protein sequences, this server provides the template-based homologous complexes (called protein complex family (PCF)) in multiple species, graphic visualization of conserved interacting residues and binding models (interfaces), conserved Gene Ontology (GO) annotations [36] and interacting domains. Our results demonstrate that this server achieves high agreements on interacting domains and GO annotations between query proteins and their respective homologous complexes.

3-2. Method and Implementation

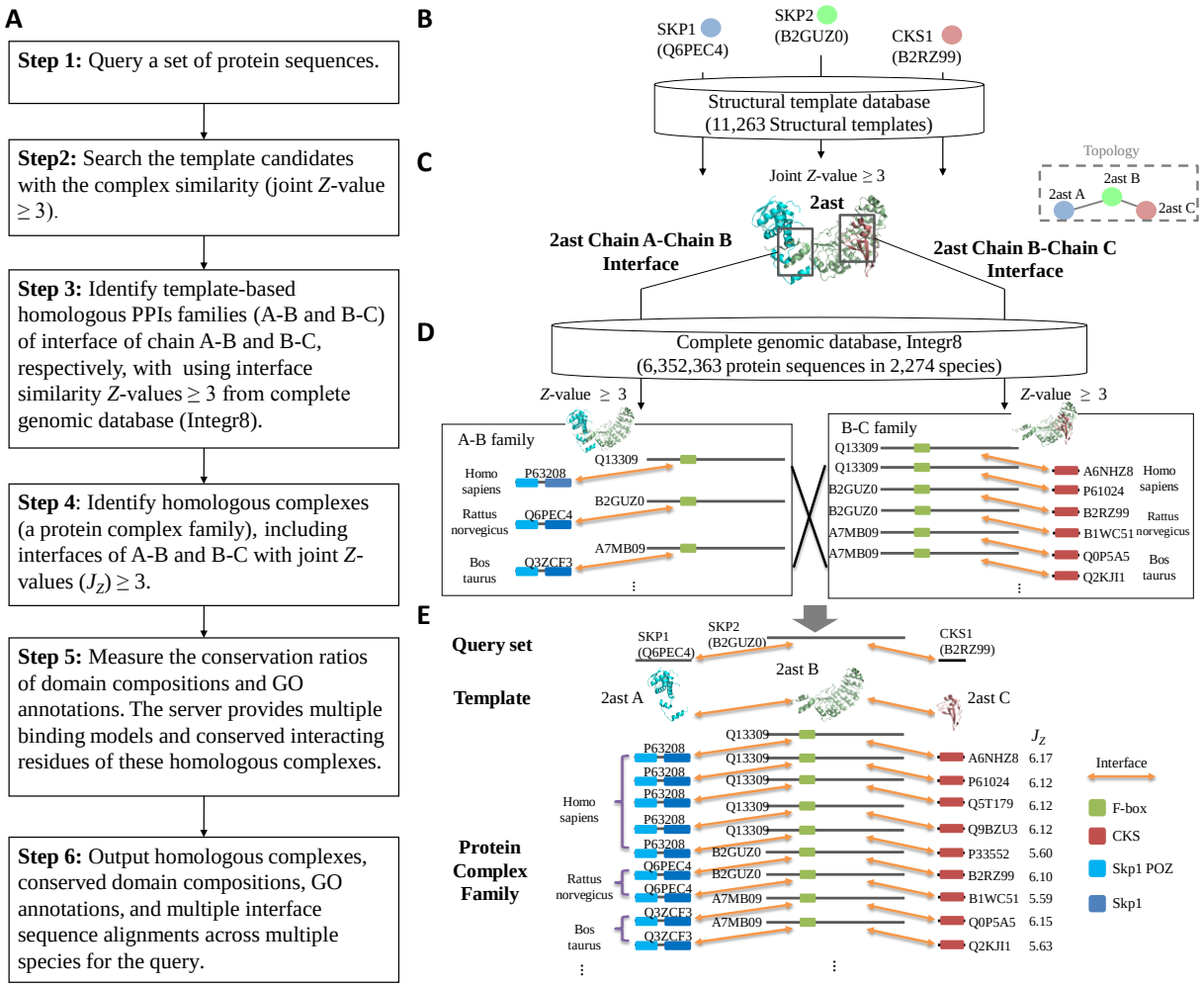


Figure 3-1. Overview of the PCfamily server for homologous complexes search using proteins Skp1, Skp2, and Cks1 of *Rattus norvegicus* as the query.

(A) The main procedure. (B) Identify the template candidate (PDB code 2ast) of the query using BLASTP and template-based scoring function to scan the structural template database. (C) The topology of the template. (D) The homologous PPI families of interfaces A-B and B-C of the template searching on Integr8 database. (E) Template-based homologous complexes of the query.

Figure 3-1 shows the details of the PCFamily server to search the template-based homologous complexes (PCF) of a set of query protein sequences by following steps (Fig. 3-1A). First, the server uses BLASTP to search template candidates from structural template database (11,263 structure complexes selected from Protein Data Bank (PDB)). Then we utilize template-based scoring function [39] to statistically evaluate the complex similarity (joint Z-value ≥ 3.0) between query proteins and candidates (Figs. 3-1B and 3-1C). After a template was

selected, the server searches the PPI family of each interface of template with Z-value ≥ 3.0 from a complete genomic database (Integr8 version 103, containing 6,352,363 protein sequences in 2,274 species) [30] (Figs. 3-2A and 3-1D). These PPI families are combined into homologous complexes with the significant complex similarity (joint Z-value ≥ 3.0) according to the interfaces of the 3D-complex template (Fig. 3-1E). For this PCF including the query, we measured the conservation ratio (CR) of the domain composition (DC) and CRs of biological processes (BP), cellular components (CC), and molecular functions (MF) using Gene Ontology annotations. Finally, this server provides homologous complexes; graphic visualization of complex topology; detailed residues interactions and interface alignments across multiple species (Fig. 3-2); conservations with GO annotations and DCs.

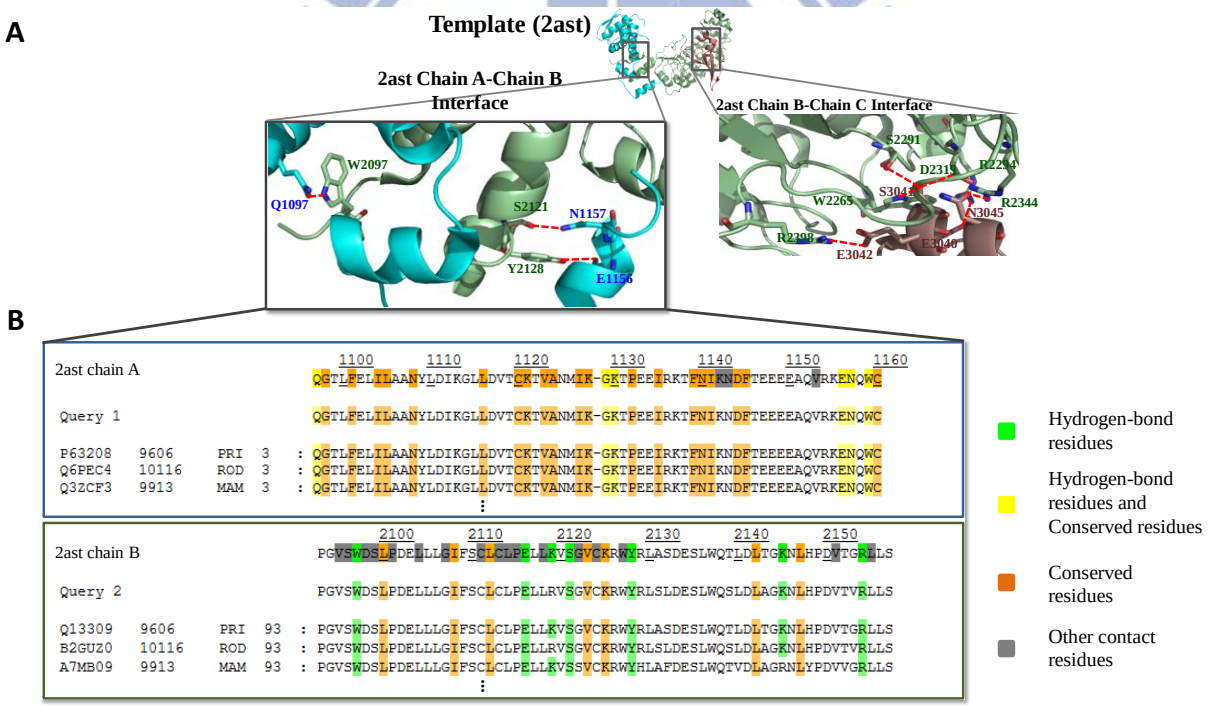


Figure 3-2. Binding models and multiple sequence alignments of PPI family in Skp1-Skp2-Cks1 complex (PDB code 2ast).

(A) The atomic binding model with hydrogen bonds (red dash lines) for each interface of the template. (B) Multiple sequence alignments of PPI family of the interface A (Skp1)-B (Skp2), respectively.

Homologous complex

The concept of homologous complex (≥ two proteins) is extended from homologous PPIs [23] and 3D-domain interologs with template-based scoring function [39]. Here, we used a 3D-trimer template *T* (proteins A, B, and C) with two interfaces A-B and B-C as a simple case to define the homologous complex of *T* as follows: (1) A', B' and C' are the homologous proteins of A, B, and C, respectively, with the significant sequence similarity (BLASTP *E*-values $\leq 10^{-10}$) [40,44]; (2) A'-B' and B'-C' are the template-based homologous PPIs of A-B and B-C, respectively, with the significant interface similarity (Z-value ≥ 3.0) [39]; (3) significant complex similarity (joint *Z*-value ≥ 3.0) between complexes A'-B'-C' and A-B-C. The joint *Z*-value of the complex similarity is defined as

$$J_z = \prod_{i=1}^{n} Z_i \quad (1)$$

where *n* is the number of interfaces of a template (*T*); Z_i is the *Z*-value (interface similarity) of the template-based homologous PPI *i* (e.g. A'-B') based on the template interface (e.g. A-B). Here, $J_Z \geq 3.0$ is considered as significant similarity according to the statistical analysis of 941 3D-structure complexes with 2,138,123 homologous complexes.

Template-based scoring function

We have recently proposed a template-based scoring function to determine the reliability of the PPI derived from a 3D-dimer structure [39]. For a predicted template-based PPI, this scoring function assigned a score, including residue-residue interacting scores, which consist of the steric (E_{vdw}) and hydrogen-bond (E_{SF}) energies, and sequence consensus scores which the couple-conserved residue score (E_{cons}) and contact-residue similarity score (E_{sim}). Finally, we calculated the Z-value of the score for this PPI using the mean and standard deviation of 10,000 random interfaces by mutating 60% interface residues.

Annotations of homologous complexes

A 3D-complex template and its homologous complexes can be considered as a PCF. The concept of the PCF is analogous to the notions of protein sequence family [20], protein structure family [22] and PPI family [23]. We believe that PCFs can be applied widely in biological investigations. We assume that the members of a PCF are conserved on GO annotations, interacting domain(s) and binding model(s). Using these conservations of a PCF, the PCFamily server can annotate the GO terms (BP, CC, and MF) and DCs of query proteins. To statistically evaluate the agreement of GO terms and DCs between the template and its PCF (with N homologous complexes), we define the agreement ratio (AR) using the conservation ratio ($CR=N_a/N$), where N_a is the number of homologous complexes with the same GO term (or DC) in a PCF. The AR is given as

$$AR = \sum_{i \in Q} (A_i(CR \geq c) / T_i(CR \geq c)) \quad (2)$$

where Q is a set of query templates; T_i ($CR \geq c$) is the total number of the GO terms (or DCs) of template i when $CR \geq c$; A_i ($CR \geq c$) is the number of the agreement GO terms (or DCs) of template i when $CR \geq c$.

3-3. Input, Output and Options

PCFamily is an easy-to-use web server (Fig. 3-3). Users input a single or a set of protein sequence(s) in FASTA format or a 3D-complexes protein structure (PDB code) (Fig. 3-3A). Typically, the PCFamily server yields structural template candidates within 25 seconds when querying three sequences and the numbers of amino acids are ≤ 450 (Fig. 3-3B). For the query, this server shows the template candidate and its PCF; detailed atomic interactions of the interfaces and binding models by using Jmol [75]; protein interaction topology (Fig. 3-3C); multiple sequence alignments (MSA) with hydrogen-bond residues and conserved residues (Fig. 3-3D); and CRs of DCs and GO terms (BP, CC and MF) (Fig. 3-3E).

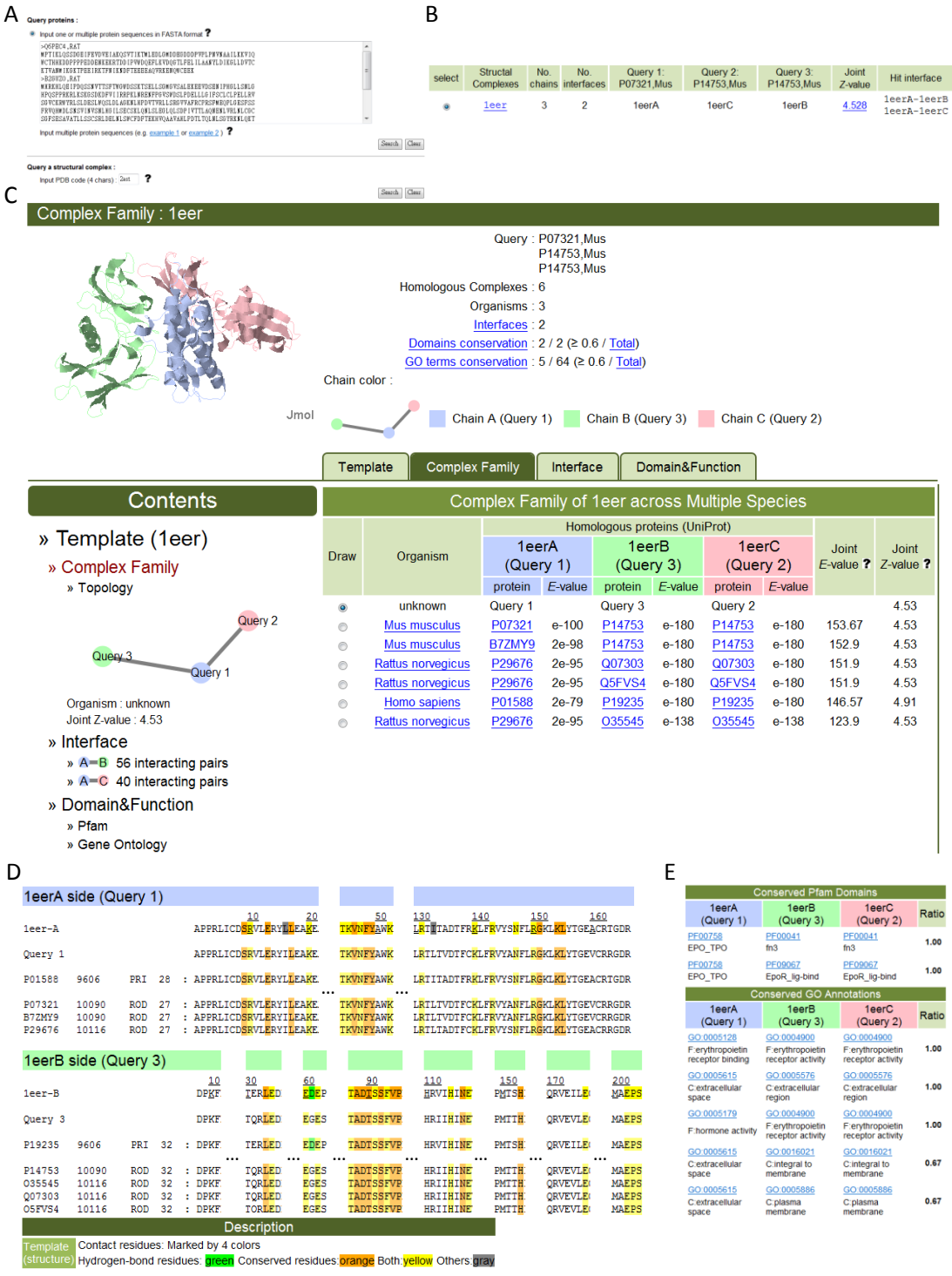


Figure 3-3. The PCfamily server search results using proteins Epor, Epo, and Epor of *Mus musculus* as the query. (A) The user interface for inputting the query protein sequences or PDB code. (B) The template candidate of the query. (C) The numbers of conserved domains and GO term conservations, interfaces, protein interaction topology, homologous complexes of the query (selected template). (D) Multiple sequence alignments and interacting residue conservations of homologous PPIs of the interface A (Epo)-B (Epor), respectively. (E) Conserved domain and GO term compositions of the protein complex family.

3-4. Example Analysis

The complex of Skp1, Skp2, and Cks1

Figure 3-1 shows search results using S-phase kinase-associated protein 1 (Skp1, UniProt accession number: Q6PEC4), S-phase kinase-associated protein 2 (Skp2, B2GUZ0), and RGD1561797 protein (Cks1, B2RZ99) of *Rattus norvegicus* as the query. Skp1 and Skp2 are subunits of the SCF^{Skp2} ubiquitin ligase complex that regulates proteolysis of the $p27^{Kip1}$ protein in cell cycle progression [76,77]. Recognition and ubiquitination of $p27^{Kip1}$ requires the accessory protein Cks1 by the SCF^{Skp2} ubiquitin-ligase complex [76]. According to KEGG pathway database [78], Skp1-Skp2 and Skp2-Cks1 in *Rattus norvegicus* are recorded in the ubiquitin mediated proteolysis pathway and the small cell lung cancer pathway, respectively. For this query, the PCFamily server found the template candidate (PDB code 2ast [76]) (Fig. 3-1C) and 43 homologous complexes (called SCF complex family), from nine species (e.g. *Homo sapiens*, *Rattus norvegicus*, and *Bos taurus* (Fig. 3-1E)). Among these 43 homologous complexes, one complex (*Homo sapiens*) is recorded in the IntAct database [8] and three homologous complexes, including the query in *Rattus norvegicus*, Q9WTX5 (Skp1)-Q9Z0Z3 (Skp2)-P61025 (Cks1b) in *Mus musculus*, and Q3ZCF3 (SKP1)-A7MB09 (SKP2)-Q0P5A5 (CKS1B) in *Bos taurus*, are recorded in KEGG pathway. In addition, 6 members are Skp1-Skp2-Cks1b (or Cks2) complexes which are highly relative to the query and the template. All members of this PCF have the same DC PF01466 (Skp1)-PF00646 (F-box)-PF01111 (CKS) and a high consensus DC PF03931 (Skp1_POZ)-PF00646-PF01111 (*CR*=0.95). The query proteins consist of these two DCs (Fig. 3-1E).

The PCFamily server provides the binding model and MSAs of each interface (Figs. 3-2 and Fig. 3-4) based on the template. Interface A-B (Fig 3-2A) contains 3 main hydrogen bonds, including Gln1097-Trp2097, Glu1156-Tyr2128, and Asn1157-Ser2121. These six residues are conserved in mammals (Fig 3-2B). Additionally, PCFamily identifies six sidechain-sidechain

hydrogen bonds forming the network to stabilize the interface B-C [76] (Fig. 3-4). All interacting residues forming the hydrogen bonds are often highly conserved and useful for observing the interface evolution across multiple species.

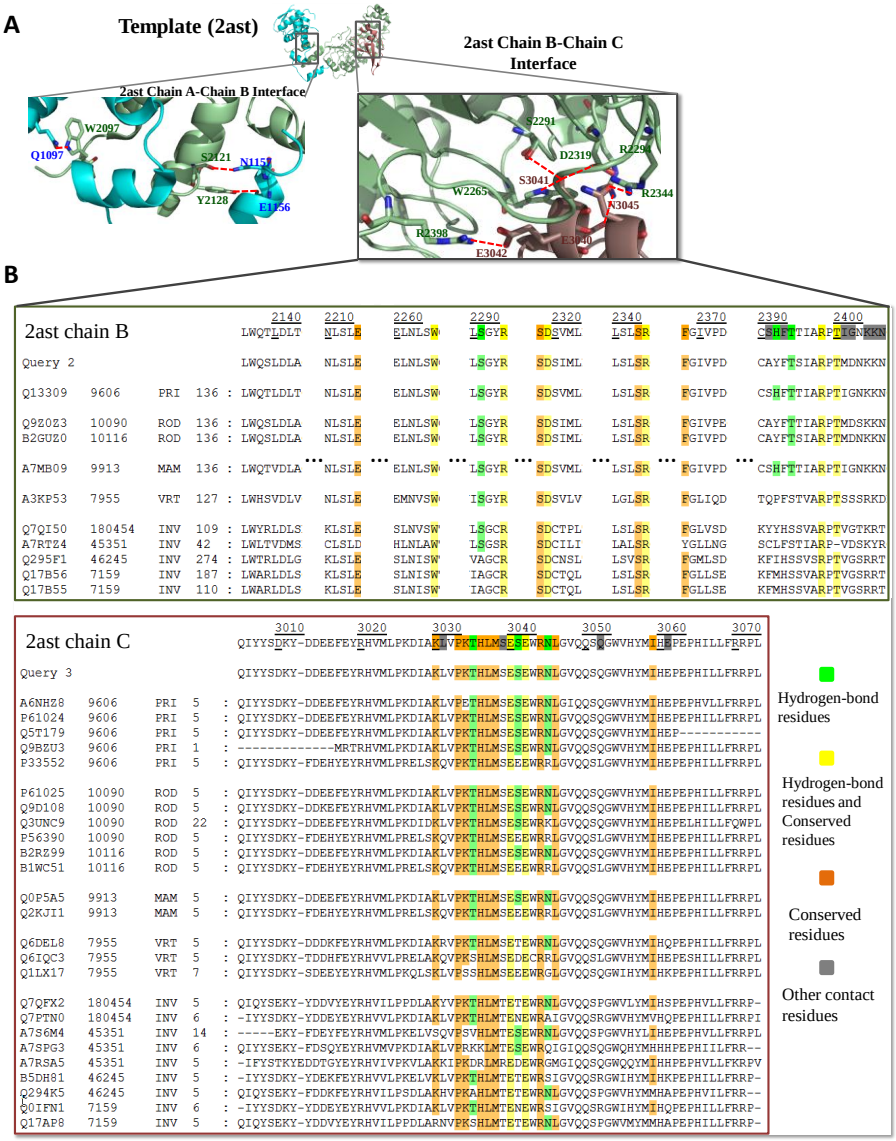


Figure 3-4. Binding models and multiple sequence alignments of PPI family in Skp1-Skp2-Cks1 complex (PDB code 2ast).

(A) The atomic binding model with hydrogen bonds (red dash lines) for each interface of the template. (B) Multiple sequence alignments of the interface B-C (Skp2-Cks1). This interface includes 11 and 26 homologous proteins of the chains B (Skp2) and C (Cks1), respectively.

Epor-Epo-Epor complex

Erythropoietin (Epo) stimulates the proliferation and differentiation of the cells (e.g.

erythroid precursor cells) [79,80]. Epo binds and orientates two cell-surface erythropoietin receptors (Epor) to activate cells and trigger an intracellular phosphorylation cascade [81]. Using *Mus musculus* Epor (P14753), Epo (P07321), and Epor (P14753) as the query proteins (Fig. 3-3A), the PCFamily server found the template candidate (PDB code 1eer) (Fig. 3-3B) and its 6 homologous Epor-Epo-Epor complexes in three species (Fig. 3-3C). Among these 6 complexes, three complexes, P19235-P01588-P19235 (*Homo sapiens*), P14753-P07321-P14753 (*Mus musculus*) and Q5FVS4-P29676-Q5FVS4 (*Rattus norvegicus*) are recorded in KEGG. Two complexes are formed by Epo (P29676) binding to Epors Q07303 [79] and O35545 [82], respectively. PCFamily indicates the MSAs with hydrogen-bond and conserved residues in the interfaces A-B (Fig. 3-3D) and A-C (Fig. 3-5) of Epor-Epo-Epor PCF.

Figure 3-5. Multiple sequence alignments of the (Epo-Epor) A-C interface of template cytokine/receptor complex (PDB code 1eer).

This interface includes five and six homologous proteins of the chains A (erythropoietin) and C (erythropoietin receptors), respectively.

This PCF includes 65 GO term compositions. Among these GO term compositions, the *CR* ratios of two MF compositions and three CC compositions exceed 0.6 (Fig. 3-3E). The query has these five GO term compositions, such as GO:0004900 (erythropoietin receptor activity)-GO:0005128 (erythropoietin receptor binding)-GO:0004900. Additionally, the query and these homologous complexes consistently contain two conserved DCs (*CR*=1), including PF00041-PF00758-PF00041 and PF09067-PF00758-PF09067. PF00758-PF00041 and PF00758-PF09067 are recorded in *i*Pfam [35]. These results reveal that the PCFamily server can identify homologous complexes for the interface evolution and annotations of the query.

3-5. Results

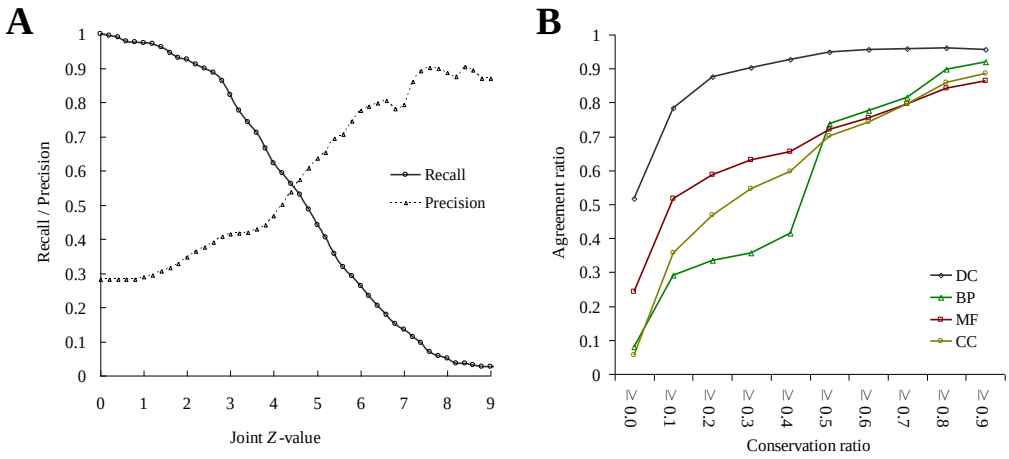


Figure 3-6. Evaluations of the PCFamily server on 941 protein complex families.
(A) The distributions of recall (solid) and precision (dot) with different joint *Z*-value thresholds. (B) The relationships between agreement ratios and the conservation ratios of domain compositions (DC), biological processes (BP), molecular functions (MF), and cellular components (CC).

To evaluate the accuracy of the PCFamily server for discovery of homologous complexes and the annotations of query proteins, we selected a non-redundant query structural template set. This set comprising 941 protein complexes (2,979 sequences and 2,042 interfaces, called NR941) was selected from the PDB released in Feb 24, 2006. For searching homologous

complexes, NR941 was used to assess PCFamily performance and to determine the threshold of joint *Z*-value J_z (Equation (1)) on the Integr8 database (Fig. 3-6A). In addition, the NR941 set was applied to calculate *CR*s of DCs (and GO terms) for each PCF and infer the relations between *CR*s and *AR*s (Equation 2) of DCs and GO terms (Fig. 3-6B).

We defined the gold standard positive and negative sets to measure the performance of the PCFamily server. Here, we used a trimer structural template *T* (proteins A, B, and C) with two interfaces A-B and B-C as a simple case to describe a positive complex (A'-B'-C') of *T* as follows: (1) A', B' and C' are homologs of A, B, and C, respectively, with the significant sequence similarity (BLASTP *E*-values $\leq 10^{-10}$) [40,44]; (2) A'-B' and B'-C' are PPIs recorded in annotated PPI databases (e.g. IntAct) and have the same interacting domains of A-B and B-C, respectively. Based on the rules, the gold standard positive set includes 770 complexes derived from the Integr8 for the set NR941. On the other hand, the gold standard negative set was generated according to the assumption that proteins, located in the same subcellular localization and acting in the similar biological processes, are more likely to form a complex than proteins involved in different processes. This study applied the relative specificity similarity (RSS) [69] to measure the BP and CC similarities of PPIs based on the GO terms. According to 198,882 interactions in IntAct database, we considered a complex candidate is a negative case, if BP and CC RSS scores of any interface of the complex are less than 0.4 (Fig. 3-7). Here, the negative set consists of 1,960 complexes.

Precision, recall and F-measure were utilized to assess the reliability of the PCFamily server for searching homologous complexes. The F-measure is given as (2 × precision × recall) / (precision + recall) where the precision and recall using the gold standard positive and negative sets. Figure 3-6A shows the relationships between joint *Z*-value J_z and recall and precision using 941 complexes on the Integr8 database. The recall significantly decreases when joint *Z*-value ≥ 3; conversely, the precision increases slightly when joint *Z*-value is between 3

and 4. The recall and precision are 0.82 and 0.45, respectively, and the PCFamily server yields the highest F-measure value (0.55) if the threshold of joint *Z*-value is set to 3.

Figure 3-6B shows the relationships between *AR*s and the *CR*s of DCs, BP, CC, and MF. If the *CR* of DCs is greater than 0.6 (black), the *AR* between the query and their respective homologous complexes exceeds 0.95 (Equation 2). If the *CR* of GO terms (i.e. BP, CC, and MF) is greater than 0.6, the *AR*s are consistent larger than 0.74 for BP (0.77, green), CC (0.74, yellow), and MF (0.75, red). These experimental results demonstrate that this server achieves high agreements on DCs and GO terms between the query (i.e. template complexes) and their respective homologous complexes.

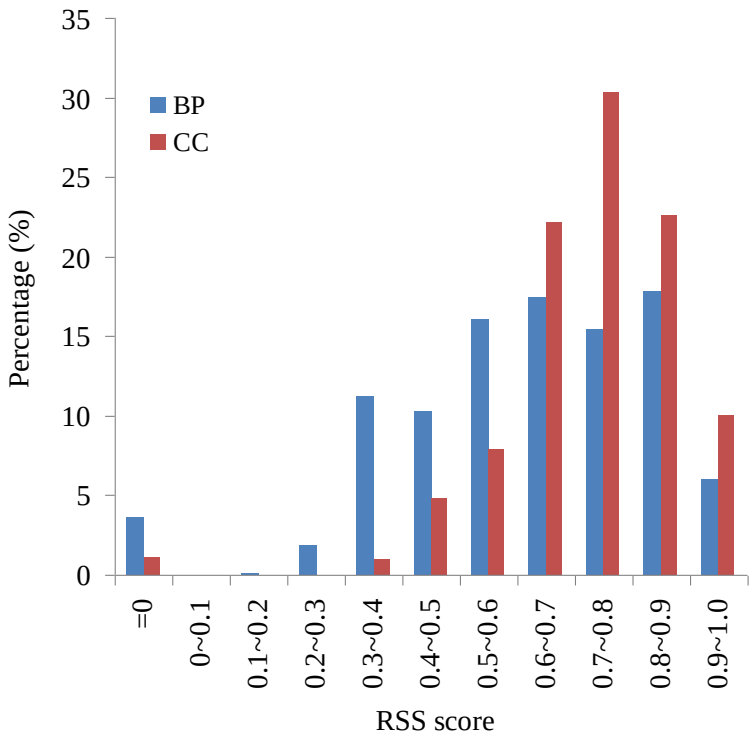
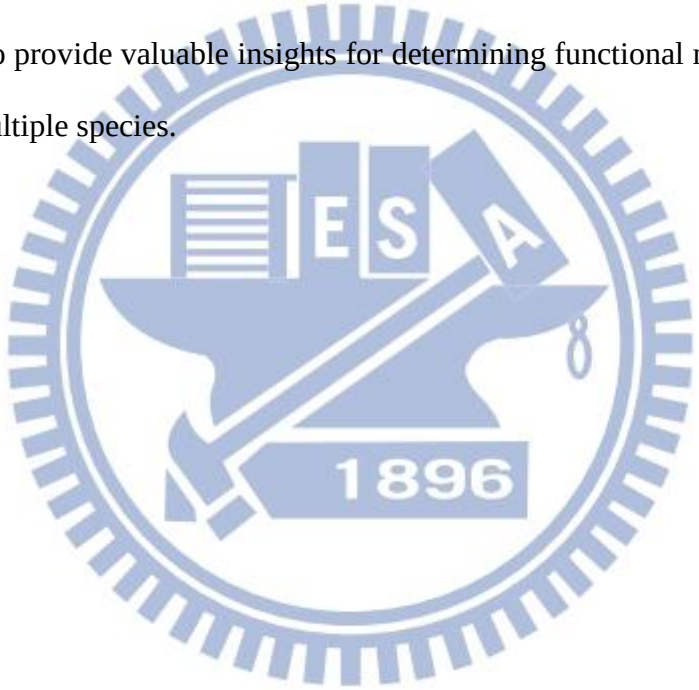


Figure 3-7. The distributions of the biological process (BP) and cellular component (CC) RSS scores on 84,082 protein-protein interactions selected from the IntAct database.

Among 198,882 interactions recorded in IntAct, 84,082 interactions can be calculated the BP and CC RSS scores. The BP and CC RSS scores of 14,188 (16.88%) and 1,742 (2.07%) interactions, respectively, are less than 0.4.

3-6. Conclusions

This study demonstrates the utility and feasibility of the PCFamily server in identifying homologous complexes and inferring conserved domains and GO terms from protein complex families. PCFamily is the first server to provide homologous complexes in multiple species; graphic visualization of the complex topology and detailed atomic residue-residue interactions; interface alignments; conservations of GO terms and domain compositions. Our experimental results demonstrate that the query and its homologous complexes achieve high agreements on domains and GO terms. We believe that PCFamily is a fast homologous complexes search server and is able to provide valuable insights for determining functional modules of biological networks across multiple species.



Chapter 4. Structural interactome of multiple vertebrate genomes though homologous protein-protein interactions

A crucial step toward understanding the spatiotemporal dynamics of a cellular system is to investigate protein-protein interaction (PPI) networks and biochemical progress. Currently, the large-scale methods are often unable to respond how a protein interacts with another one within a given PPI network and describe the relationship between the mutation of proteins and disease syndrome. To address this issue, we numerously enhanced and modified our previous PPI family search and 3D-domain interologs with template-based scoring function. Our method could efficiently enlarge the PPIs annotated with residue-based binding models in structure resolved networks in *H. sapiens*, *M. musculus*, and *D. rerio*. This work is the first to construct structure resolved PPI networks across multiple species, including *H. sapiens*, *M. musculus*, and *D. rerio*. The PPIs with atomic residue-based binding models in the derived structure resolved network achieved highly agreement with Gene Ontology similarities. Furthermore, the architecture of these networks is a scale-free network which is consistent with most of the cellular networks. In addition, our derived networks can be used to observe the consensus proteins and modules which are high conserved appearing in multiple organisms. These consensus proteins are often the essential genes and related to diseases recorded in OMIM. Experimental results also indicate that the mutations of interacting residues on the PPIs often related to diseases are often on. Our results demonstrate that the structure resolved PPI networks in vertebrates can provide valuable insights for understanding the mechanisms of biological processes.

4-1. Introduction

A crucial step toward understanding the spatiotemporal dynamics of a cellular system is to investigate protein-protein interaction (PPI) networks and biochemical progress [3,83,84]. Many high throughput experimental methods, such as high-throughput yeast two-hybrid screening [25,26] and co-affinity purification [27], and computational approaches have been proposed to construct the PPI network within an organism. These large-scale methods are often unable to respond how a protein interacts with another one and describe the relationship between the mutation of proteins and disease syndrome. Previous studies have combined protein structure information with experimental PPIs to investigate how mutations affect protein interactions in disease [14-16]. Based on experimental PPIs, a structurally resolved human protein interaction network has been reconstructed to examine the relationships between genes, mutations and associated disorders [16]. These experimental PPIs were distributed on several well-studied organisms (e.g. *S. cerevisiae*); conversely, the PPIs of most species were not complete. For example, the numbers of PPIs for *D. rerio* (227) and *Mus musculus* (7,736) recorded in five public databases [8-10,85,86] (e.g. BioGRID and IntAct).

To discover the sequence homologs of a known protein provides the clues for understanding the function of a newly sequenced gene. We have provided "protein-protein interaction family" to annotate genome-scale PPIs through the homologous PPIs [23] searching the complete genomic database (Integr8, containing 6,352,363 protein sequences in 2,274 species) [30]. Furthermore, a known three-dimensional (3D) structure complex could provide interacting domains, and atomic detailed binding models of interactions. Some methods have utilized template-based methods (i.e. comparative modeling [32] and fold recognition [33]) to predict the PPIs by accessing interface preference through the fitness of known template structures. However, these methods [32,33] are time-consuming to search all possible protein-protein pairs in a large genome-scale database across multiple species. Therefore, to

further utilize both "protein-protein interaction family" and 3D structure complexes, we are able to construct structure resolved PPI networks with binding mechanisms in multiple organisms.

To address this issue, we numerously enhanced and modified our previous PPI family search (sequence-based PPI search method [23]) and 3D-domain interologs with template-based scoring function (3D-template PPI prediction method [87]). Our method could efficiently enlarge the PPIs annotated with residue-based binding models in structure resolved networks in *H. sapiens*, *M. musculus*, and *D. rerio*. For each structure resolved network, we investigated the reliability by using the Gene Ontology and the network architecture (i.e. scale-free network). In addition, our method can identify the conserved proteins and network modules across multiple networks. These conserved proteins are highly related to the essential genes and diseases recorded in "Online Mendelian Inheritance in Man (OMIM [11])". Furthermore, we demonstrated that these disease-related mutations are more enrichment on the interacting residues, especially forming the hydrogen bonds. These results indicate that the structure resolved PPI networks can provide the insight for understanding the mechanisms of biological processes and interactomes.

4-2. Methods and Materials

Constructing the structural resolved PPI networks

A major challenge of systems biology is to understand the networks of interacting genes, proteins and small molecules that produce biological functions. For efficiently enlarging protein interactions annotated with residue-based binding models, we have proposed the concept "3D-domain interolog mapping [39,87]": for a known 3D-structure complex (template T with chains A and B), domain *a* (in chain A) interacts with domain *b* (in chain B) in one species.

The proteins of the homolog families A' and B' of A and B have the significant sequence similarity (i.e. BLASTP *E*-values $\leq 10^{-10}$) and contain interacting domains *a* and *b*, respectively. All possible protein pairs between these two homolog families are considered as protein-protein interaction candidates using the template T. Then, we utilize our previous scoring system [39,87] to evaluate the binding model similarity between candidates and template. According to this concept, protein sequence databases can be searched to annotated protein-protein interactions across multiple species efficiently.

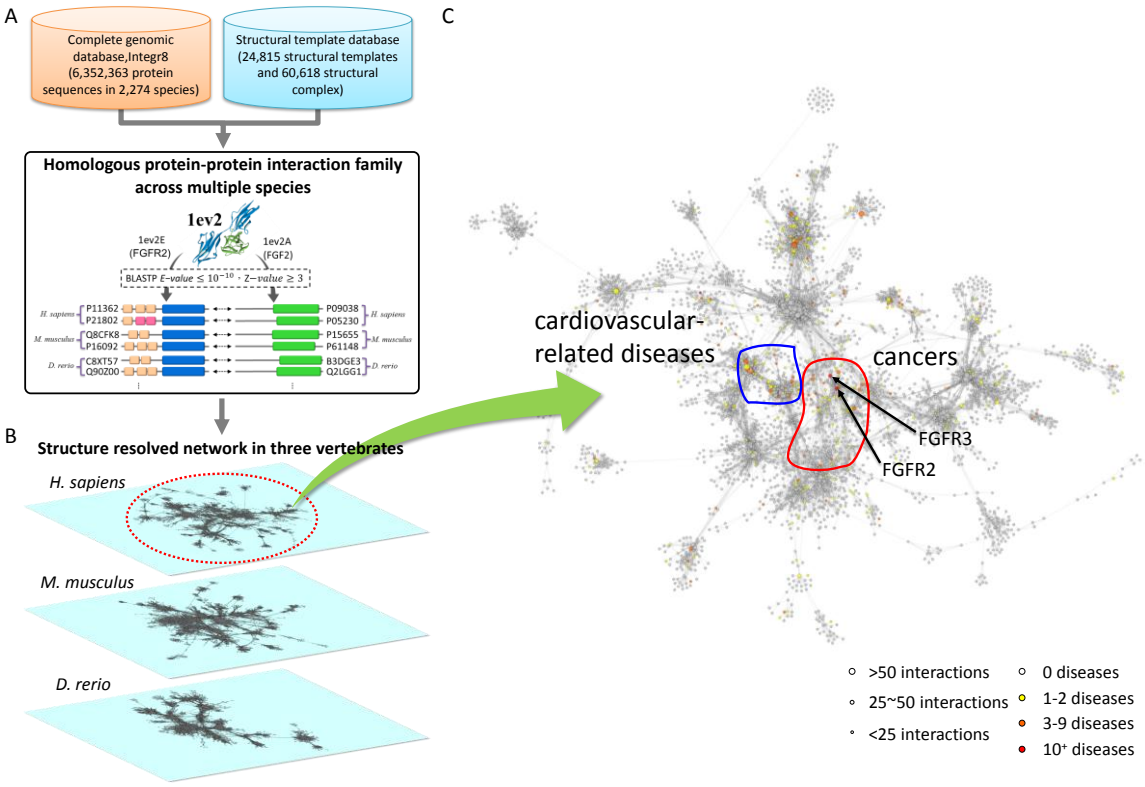


Figure 4-1. The overview of constructing structure resolved PPI networks in three vertebrates though "3D-domain interolog mapping"

(A) 3D-domain interolog mapping is used to infer the homologous PPIs through structural templates and complete genome databases. (B) The structure resolved PPI networks of *H. sapiens*, *M. musculus*, and *D. rerio*. (C)The human PPI network with the disease data derived from the OMIM. The size and color of a node (protein) denote the numbers of interactions and diseases, respectively.

Figure 4-1 illustrates the overview of constructing structure resolved PPI networks in three vertebrates though "3D-domain interolog mapping". First, a structure template library

comprising 60,618 3D-dimers involved in 24,815 complexes was selected from the protein data bank (PDB[88]) released in Sep 2, 2011 (Fig. 4-1A). The interacting residues and scoring functions defined by using our previous studies [39,87] were used to identify the similar binding interfaces of PPIs. After 3D-dimer template library and template profiles were built, we inferred homologous PPIs of each interface of the template with *Z*-value ≥ 3.0 from a complete genomic database (Integr8 [89]) (Fig. 4-1A). According to these homologous PPIs in *H. sapiens*, *M. musculus*, and *D. rerio*, we constructed and aligned these structure resolved PPI networks (Fig. 4-1B).

Multiple network alignment

The methods for network alignments can be roughly divided into global alignment and local alignment. By searching for a single comprehensive PPI network mapping of the whole set of proteins and protein interactions from different species[90], the global network alignment can answer interactome evolutions with conserved and specific proteins (and PPIs). Two basic issues should be addressed for a network alignment method. Firstly, an alignment method should provide the importance (such as hub and conservations) of proteins and PPIs in multiple networks across species. Second, for a selected protein and PPI, the score function of an alignment method should reflect the similarity of the aligned proteins and PPIs in the networks. Here, we described a new global network alignment method based on "3d-domain interologs mapping". According to the definition of the "3d-domain interologs mapping", the protein-protein interactions of the same family share the same interacting domains and have the similar binding models. Therefore, for a specific PPI, these PPIs could be considered as the corresponding PPI alignment candidate in other organisms.

Our global alignment method applied a greedy strategy which the PPI with highest

importance is aligned with the highest priority. Here, we evaluated the importance of a given PPI (I) within the network based on the degree, conservation, and PPI reliability. Two proteins, forming a PPI, with a higher degree are usually the hub in a network. The degree (DI) of a PPI forming by proteins a and b is defined as DI = Da + Db, where the Da and Db are the degrees of proteins a and b, respectively. The PPI involving in many organisms is usually the essential PPI and plays an important role for biological functions and processes. Therefore, the evaluation conservation (CI) of a PPI (I) is defined as CI = TaxI /11, where TaxI is the number of taxonomy divisions, defined by the NCBI taxonomy database [91], of the interacting proteins in the PPI I family. Here, the maximum number of taxonomy divisions is 11. Finally, the reliability (RI) of a PPI is defined as RI = (EI + TI) / 2, where EI is 1 if the PPI I was recorded in five public PPI databases (e.g. IntAct[8], DIP[9], MIPS[28], BioGRID[10], and MINT[29]); otherwise, EI is 0. The TI is set to 1 while the 3D-dimer template and the PPI I are in the same organism; otherwise, TI is 0. Final, the importance (S) of a given PPI (I) is calculated by S=CI + RI + TI.

The network alignment algorithm

Given three structural resolved PPI networks of *H. sapiens* (N_H), *M. musculus* (N_M), and *D. rerio* (N_D), we provided multiple network alignment by aligning N_M and N_D to N_H and the algorithm is summarized in Figure 4-2 and proceeds as follows:

(1) For each PPI of N_H, N_M and N_D, we calculate the importance (S) of the PPI by using the equation ($S=C_I + R_I + T_I$) describing in previous paragraph.

(2) After calculating the all importance of all PPIs among the N_H, each PPI gets the priority according to the value of importance. Then, Greedy picking the PPI *I* with the highest value of importance and its corresponding PPI I_H family (F_H).

(3) Selecting the most similar PPI I_M and I_D of *M. musculus* and *D. rerio*, respectively, in the F_H based on the significant joint sequence similarity between two pairs, *i.e.*, (A,

A_1') and (B, B_1'), of the I (A and B) and I_M (A_1' and B_1'). This work followed previous studies [23,40,44] to define joint sequence similarity as $J_E = \sqrt{E_A \times E_B}$. E_A is the E-value of proteins A and A_1'; and E_B is the E-value of proteins B and B_1'.

(4) If the I_M and I_D exist, the I is an alignable PPI of N_H and the summarized importance S' = S_H + S_M + S_D; otherwise, I is a human-specific PPI and S' is 0.

(5) Repeat the steps (4) and (5) until all PPIs of are N_H are assigned as alignable or human-specific PPIs with S'.

(6) Greedy aligning networks by choosing PPI alignment which I, I_M, and I_D have the highest summarized importance S'.

(7) Repeat the steps (4) and (5) to find the next PPI alignment with the highest summarized importance S'.

(8) Repeat step (7) until all PPIs of N_H, N_M, and N_D are aligned.

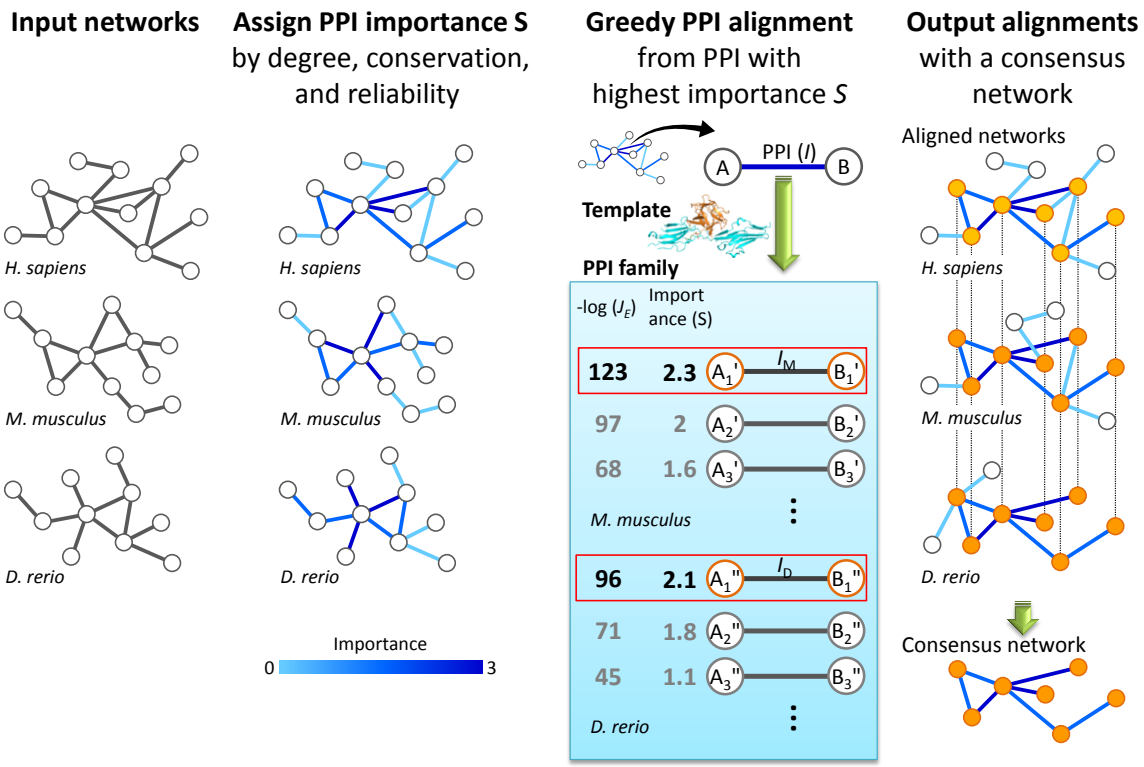


Figure 4-2. Conceptual overview of alignment procedure.

Finally, we identified 1,887 proteins and 5,845 PPIs which are consensus in structure resolved PPI networks of *H. sapiens*, *M. musculus*, and *D. rerio*.

Collecting the list of disease-associated genes, mutations, and diseases

To further investigate the relationship between disease-associated genes and mutations in the structure resolved human PPI networks, we collected the disease-related mutations from OMIM [11] database. The database of single nucleotide polymorphisms (dbSNP [92], build 132) is a public-domain archive for a broad collection of simple genetic polymorphisms. According to OMIM [11] database which contains the relationships between genes and diseases, we collected all "OMIM-curated-records" from the dbSNP database. We got 15,995 mutations including in-frame and truncating mutations in 1,949 genes. For the further analysis, we selected the 2,202 mutations in 137 genes to validate the structurally resolve human PPI network with annotations of mutations and diseases (Fig. 4-1C). Here, the sizes and color distributions of the nodes (proteins) denote the numbers of interactions and diseases, respectively. The larger node represents the protein with the more number of PPIs and the red node denotes the protein with the more number of diseases. There are two main disease hubs (No. of disease > 10): TGFR4 and TGFR3 with 14 and 13 diseases, respectively.

4-3. Results

Structure resolved PPI networks of *H. sapiens*, *M. musculus*, and *D. rerio*

For evaluating the structural PPI network annotated with residue-based binding models, we compared the numbers of the proteins and PPIs in our structure resolved PPI networks with ones of the human structural PPI network [16] which can only be applied on the well-studied

species. According to PPI recorded in five public databases, the number of PPIs in human (67,596 PPIs) is significantly more than the ones of mouse (7,735 PPIs) and zebrafish (221 PPIs) (Table 4-1). The method proposed by Wang, X. J. *et al.* [16] would not be useful to apply to the mouse and zebrafish because this methods considered both the experimental PPIs and protein structures. Conversely, our method using "3d-domain interologs mapping" and "PPI family is able to efficiently enlarge PPIs annotated with residue-based binding models, especially useful for seldom-studied organisms (e.g. zebrafish) or new sequencing organisms. Although most of the PPIs derived from our "3d-domain interologs mapping" are still not confirmed by experiments, our previous works have achieved the high annotating precision and high agreement with ddG of experimental binding energies and experimental PPIs [23,24,39,87].

Table 4-1. Statistics of proteins and PPIs derived from our result, public databases, and Wang, X. J. *et al.* on *H. sapiens*, *M. musculus*, and *D. rerio*

Species	No. proteins in genome[*1]	3D-domain interologs		Five public databases[*2]		Wang, X. J. *et al* [16].	
		No. proteins	No. PPIs	No. proteins	No. PPIs	No. proteins	No. PPIs
H. sapiens	56,006	9,493	39,058	12,206	67,596	2,816	4,222
M. musculus	36,379	7,689	33,125	4,177	7,735	-	-
D. rerio	21,601	5,084	21,236	137	221	-	-

[*1] The number of proteins in a specific genome is calculated by using the Integr8 database.

[*2] The experimental PPIs are derived from five public databases (IntAct, MIPS, DIP, MINT, and BioGRID)

To further verify the quality of our structure resolved PPI networks, we utilized the Gene Ontology (GO) [93] similarities, including biological process (BP), cellular component (CC), and molecular function (MF), between interacting protein pairs and all protein pairs in a structural PPI network. Here, we applied the relative specificity similarity (RSS) [69] to measure the GO similarities between two proteins. Figure 4-3 illustrates the RSS score distributions of BP, CC, and MF on interacting protein pairs and all protein pairs in the structural PPI network. GO annotations of BP, CC, and MF of are enrichment while the RSS scores are higher than 0.7 (Fig. 4-3). In addition, the RSS scores of interacting protein pairs are significantly greater than the ones of all pairs by using the Mann–Whitney U test (p-value < 10^{-40}) which is a non-parametric

statistical hypothesis test. The RSS score distributions of BP, CC, and MF on interacting protein pairs and all protein pairs within the mouse and zebrafish networks (Fig. 4-4) are similar to ones of the human network. These results illustrate the importance of structural resolution and imply that the PPIs in our structure resolved PPI networks significantly share the similar GO annotations.

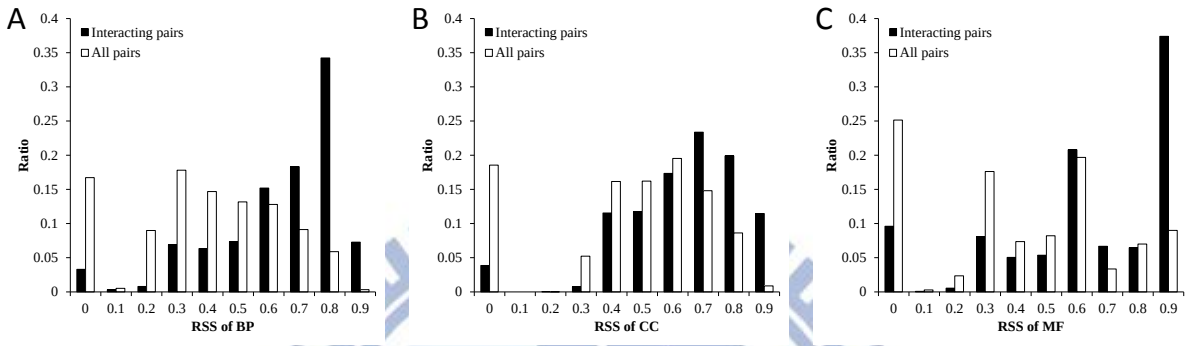


Figure 4-3. The distributions of relative specificity similarity (RSS) of BP, CC, and MF of the interacting protein pairs in the derived structural PPI networks

(A) The BP RSS distributions of 10,163 interacting protein pairs and all protein pairs (3,925,772 pairs). (B) The CC RSS distributions of 9,254 interacting protein pairs and all protein pairs (3,424,256 pairs). (C) The MF RSS distributions of 12,387 interacting protein pairs and all protein pairs (4,331,532 pairs). The protein pairs with BP (CC or MF) annotations are considered. The BP, CC, and MF RSS scores of interacting pairs have a significantly enrichment while RSS score >= 0.7. The interacting pairs have significantly higher RSS scores than the ones of random pairs in the networks according to the Mann–Whitney U test (p-value < 10^{-40}).

A network with a power degree distribution is called scale-free, a name that is rooted in statistical physics literature. An important finding of the cellular network architecture is that most networks within the cell approximate a scale-free topology [94]. Therefore, our structure resolved PPI networks of H. sapiens, M. musculus, and D. rerio were evaluated based on the characteristic of scale-free networks that the P(k), the probability of a node with k links, decreases as the node degree increases on a log-log plot (Fig. 4-5). Then, the degree exponent γ are 2.127, 2.088, and 1.958 in the structure resolved PPI networks of H. sapiens, M. musculus, and D. rerio, respectively. In general, the smaller the value of γ, the more important the role of the hubs is in the network. A scale-free network typically has degree exponents $2 \leq \gamma \leq 3$, but

can also exist with exponents less than 2 [94,95]. This result is consistent with the architecture (i.e. scale-free network property) of some cellular networks [95].

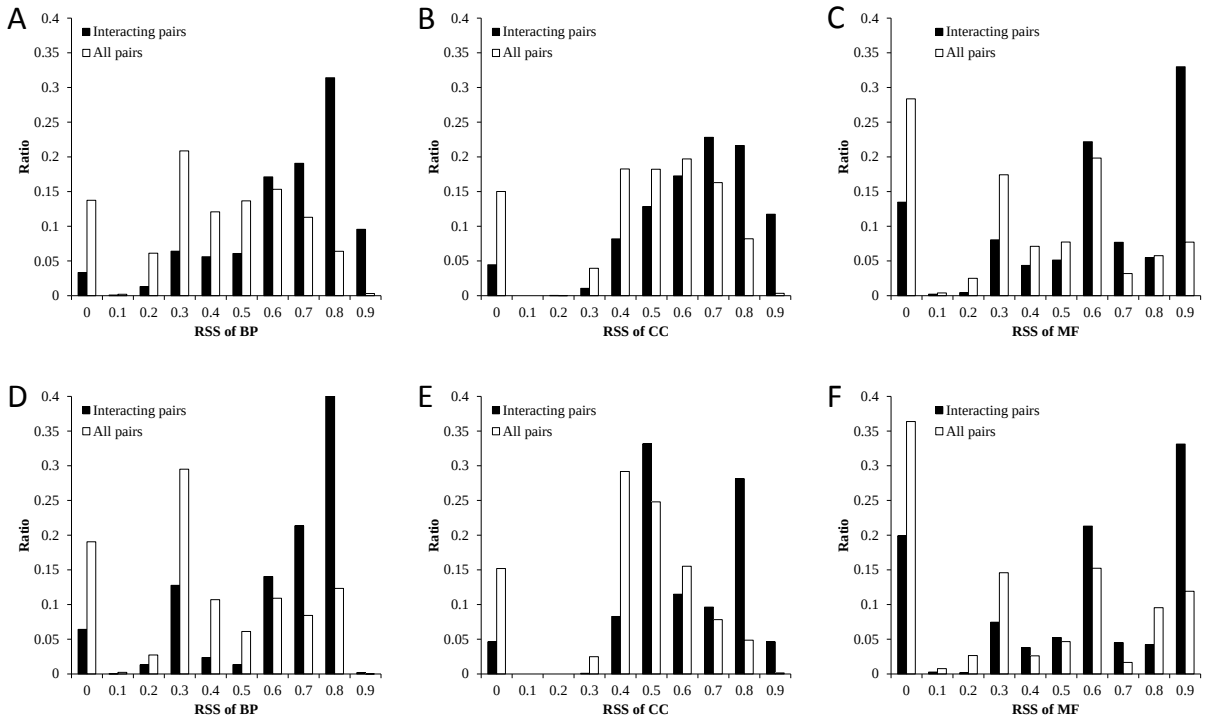


Figure 4-4. The distributions of BP, CC, and MF RSS scores on interacting protein pairs and all protein pairs within the mouse and zebrafish networks

The BP, CC, and MF RSS scores have a significantly enrichment while RSS score >= 0.7 in both mouse and zebrafish networks.

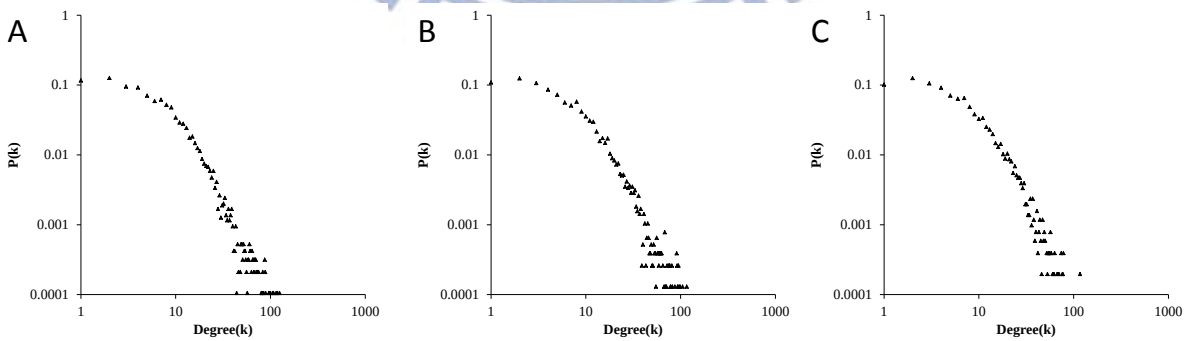


Figure 4-5. The node degree distributions of three structure resolved PPI networks: (A) *H. sapiens*, (B) *M. musculus*, and (C) *D. rerio*

The degree exponent γ are 2.127, 2.088, and 1.958 in the structure resolved PPI networks of *H. sapiens*, *M. musculus*, and *D. rerio*, respectively. A scale-free network typically has degree exponents $2 \leq \gamma \leq 3$, but can also exist with exponents less than 2. These three structural networks are scale-free networks.

The structure resolved PPI networks analysis

To further investigate the biological meaning of our networks, we analyzed the grouping property of human network by using the Gene Ontology annotations. Here, we defined the grouping property of a network as that the proteins which are involved in similar process and located on similar cellular component would be the neighbors in a network. We identified four cellular components (i.e. nucleus, intracellular, membrane, and others) for each protein based on the CC annotations (Fig. 4-6A). We also identified six biological processes (i.e. immune response, transport, signal transduction, protein metabolic, nucleic acid metabolic process, and others) for each protein based on the BP annotations (Fig. 4-6B).

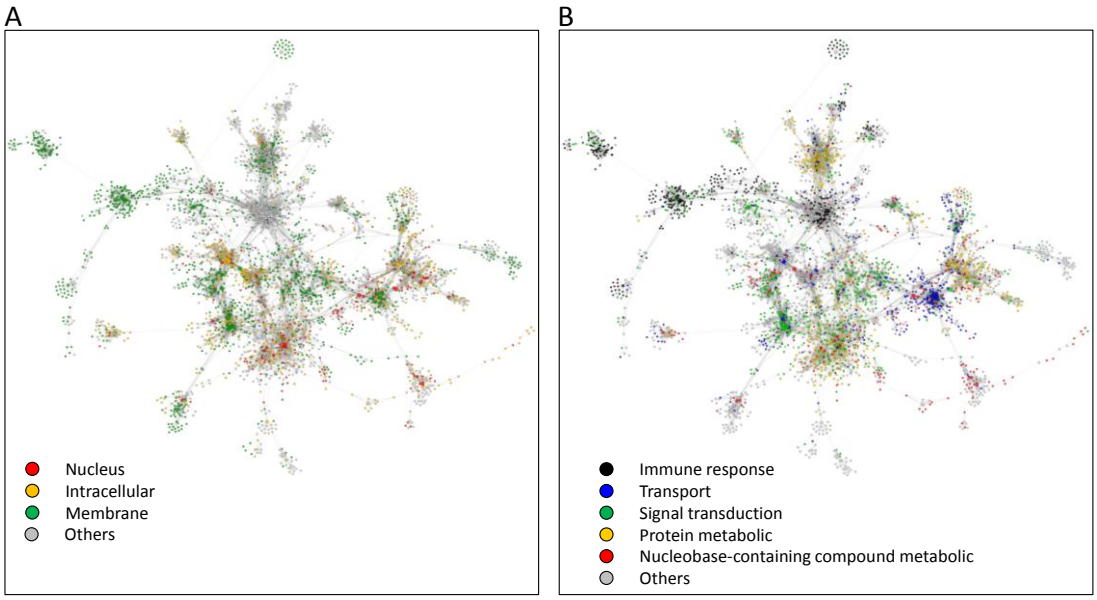


Figure 4-6. Characteristics of the structure resolved protein network in H. sapiens using GO annotations.
(A) According to GO cellular component (CC) annotations, proteins in structure resolved protein network can be annotated into four CC terms (groups), including 218 proteins in nuclear part (GO:0044428, red), 829 proteins in intracellular (GO:0005622, yellow), 1265 proteins in membrane (GO:0016020, green), and others (gray). (B) Based on biological processes, 281 proteins are annotated with nucleobase-containing compound metabolic process (e.g., transcription) (GO:0006139, red); 613 proteins are annotated with protein metabolic process (e.g., translation) (GO:0019538, yellow); 710 proteins are with signal transduction (GO:0007165, green); 364 proteins are with transport (GO:0006810, blue); and 436 proteins are with immune response (GO:0006810, black).

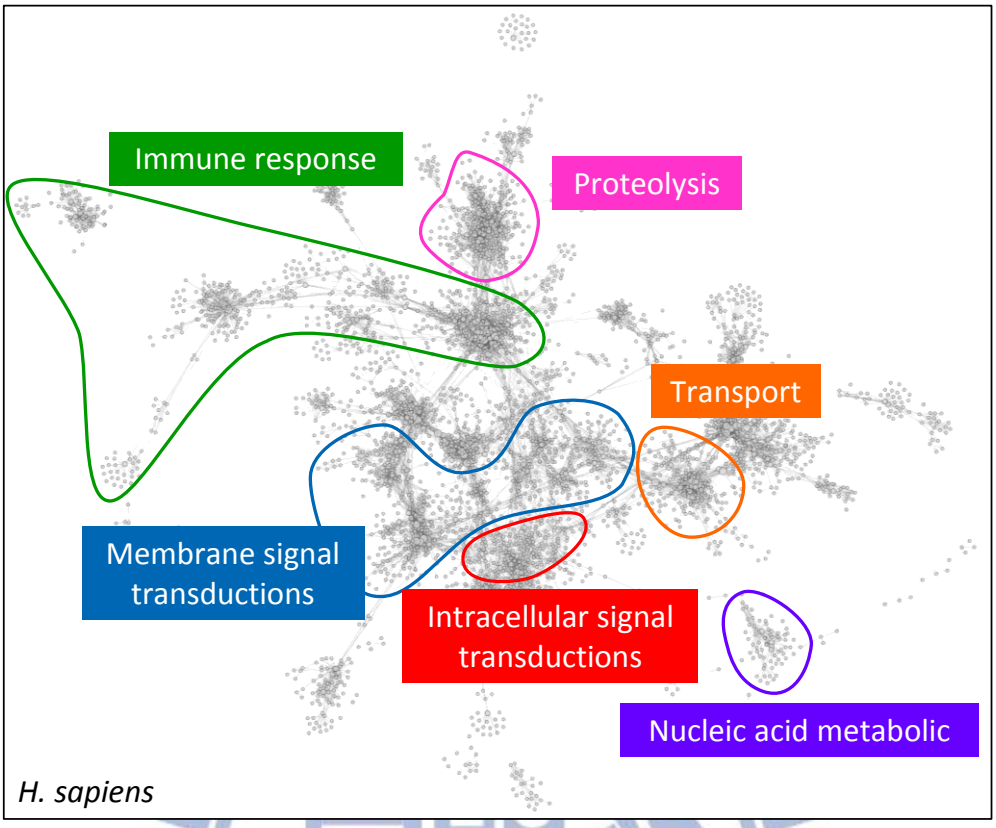


Figure 4-7. Six major cellular processes in our derived network of *H. sapiens*. According to the GO annotations (Fig. 4-6), our derived structure resolved PPI network could be grouped into six major cellular processes, including nucleic acid metabolic process (e.g., transcription); protein metabolic process (e.g., translation); intracellular signal transduction process; membrane signal transduction process; transport process; proteolysis process (e.g. proteasome); and immune responses.

According to the GO annotations, our derived structure resolved PPI network could be grouped into six major cellular processes, including nucleic acid metabolic process (e.g., transcription); protein metabolic process (e.g., translation); intracellular signal transduction process; membrane signal transduction process; transport process; proteolysis process (e.g. proteasome); and immune responses. In addition, our PPI network can also reflect the communication of six major cellular processes (Fig. 4-7). The intracellular signal transduction plays an important role in our network. This process receives the signals which are provided from the membrane signal transduction (e.g. EGFR, FGFR, and other membrane receptors) and the immune response (e.g. T-cell receptor). In addition, the intracellular signal transduction also communicates with the transport process which locates in cell membrane and cytoplasm and is

the peripheral portion of our derived network. The nucleic acid metabolic processes are the kernel processes of a living cell and could be regulated by the signal transduction. In our derived network, the nucleic acid metabolic process only communicates with the intracellular signal transduction and transport process. The results imply that the biological behavior of our derived network is consistence with our knowledge for a living cell.

The consensus proteins, processes, and organism-specific processes

According to "3d-domain interologs mapping" and the multiple network alignment described in Methods, we were able to compare these three vertebrate protein interaction networks (i.e. *H. sapiens*, *M. musculus*, and *D. rerio*) and identify the consensus proteins and protein-protein interactions. Here, we identified 1,887 consensus proteins and 5,845 consensus PPIs from 4,135 proteins and 21,648 PPIs of structure resolved human network. To further evaluate the biological meanings and network topologies of the consensus and non-consensus proteins, we investigated these consensus proteins according to the three dimensions, including the essential proteins; involving in diseases; and locating in the central part (e.g. hub) within the protein interaction network.

Essential genes usually involve in the fundamental cellular processes which required for the survival of an organism. As a result, the essential genes are often highly conserved across multiple organisms [96]. We collected the annotations of essential proteins from the Database of Essential Gene (DEG [97]). Because few vertebrate proteins, especially in *Homo sapiens*, were recorded as are essential genes recorded in DEG, we identified the essential proteins (genes) of the *Homo sapiens*, *Mus musculus*, and *Danio rerio* by using BLAST to search orthologs of essential genes recorded in DEG from Integr8 [30]. To investigate the reliability of the orthologs mapping, we collected the orthologs protein data set (named ORT) from the COG database [98]

and evaluated the relationship between the sequence similarity (i.e. BLASTP *E*-value) and orthologs protein pairs. The ORT set consists of 3,050,847 orthologs protein pairs and 112,920 proteins. Figure 4-8 illustrates the sequence similarity distribution of these orthologs protein pairs. When sequence similarity (BLASTP E-value) ≥ 10-70, the number of all protein pairs significantly increase to cause the decreasing of the precision (No. orthologs protein pairs / No. all protein pairs); moreover, the number of orthologs protein pairs decrease more gradually than all protein pairs at JE ≤ 10-70. While the threshold of sequence similarity (BLASTP E-value) is set to 10-70, the precision is higher than 0.7 and 431,062 orthologs pairs can be annotated. As a result, we could be able to enrich the number of essential proteins with a reliable accuracy. Finally, we annotated 1,557 essential proteins in the structurally resolved human protein interaction network through the DEG and orthologs annotation.

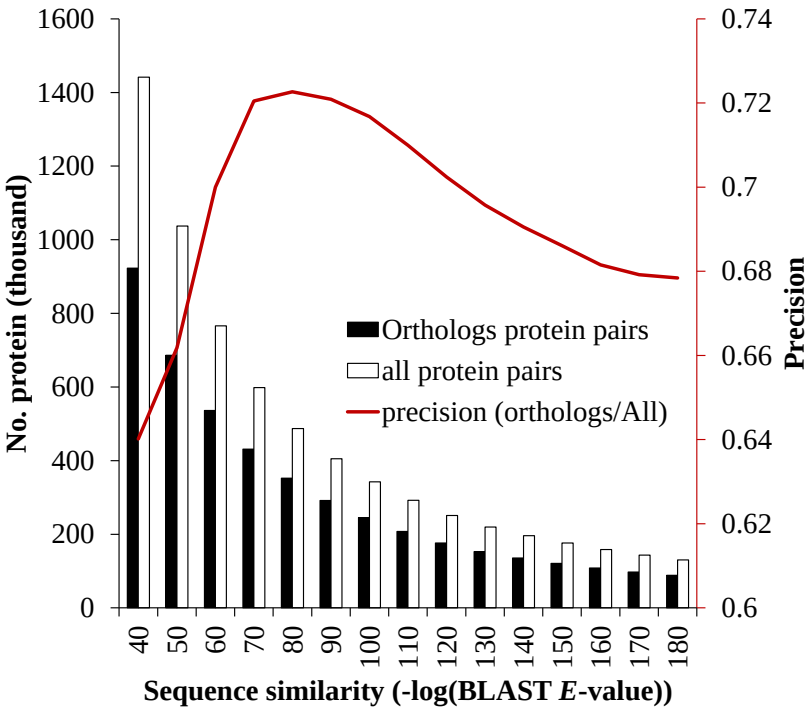


Figure 4-8. The distribution of orthologs protein pairs under different sequence similarities

The precisions of orthologs annotation (No. orthologs pairs / No. all pairs) are higher than 0.72, while the cut-off of sequence similarity (-log (BLASTP E-value)) are set to 70, 80, and 90.

To further investigate the relationship between the essential proteins and consensus prions, we evaluated the ratios of essential proteins in consensus (i.e. the proteins are conserved in three vertebrates) and between non-consensus proteins of structurally resolved human protein interaction network. Table 4-2 shows the distribution of essential proteins in consensus and non-consensus proteins. As a result, 969 (51.4%, 969/1887) and 558 (26.2%) essential proteins were recognized as consensus and non-consensus proteins, respectively. Furthermore, the consensus proteins lead to the significant enrichment (*Z*-value=16.69) of essential proteins in the structural network. Here, the *Z*-value is calculated by:

$$Z\text{-value} = \frac{X - \overline{X}}{\sigma_X}$$

where X (=969) is the number of essential proteins identified as consensus proteins; $\overline{X}$ and σ_X are the average and the standard deviation of essential proteins among 1,000 random sets. Each set consists of 1,887 proteins randomly selected from 4,135 proteins in the human network. This result indicates that these consensus proteins are significant related to the essential genes.

Furthermore, we investigated the relationship between the diseases and consensus proteins. Although many of diseases are organism specific, there are still many disease involved in the essential biological pathways which is conserved in multiple organisms (e.g. cancer). If a disease is involved in conserved pathway, the scientists could utility the animal model to research and investigate the mechanism of human disease. Therefore, we collected the disease and mutation data from the OMIM [11]. There are 1,442 in-frame and 1,898 truncating mutations within 393 proteins of our structurally resolved human protein interaction network. Table 4-2 illustrates the distributions of the consensus and non-consensus proteins involving in diseases. In the structurally resolved human protein interaction network, 207 (11.0%) consensus proteins involved in diseases and 84 (3.7%) non-consensus proteins involved in diseases. Furthermore,

the proteins involved in diseases have the significance enrichment (*Z*-value=8.92) of essential proteins. These results suggest that the mutations with structure binding in OMIM are highly related to the consensus proteins across three vertebrate PPI networks.

Table 4-2. The ratios of essential proteins and disease related proteins in consensus and non-consensus proteins

Protein type	No. of proteins	Essential proteins			Proteins involved in disease			Essential proteins involved in disease		
		No. of proteins	Ratio	Z-value	No. of proteins	Ratio	Z-value	No. of proteins	Ratio	Z-value
All	4,135	1,557	0.377		291	0.070		187		
Consensus	1,887	969	0.514	16.69	207	0.110	8.92	154	0.082	10.79
Non-consensus	2,248	588	0.262	-16.66	84	0.037	-9.03	33	0.015	-10.49

To further investigate the consensus proteins and its corresponding cellular process, we compared these 1,887 consensus proteins and 5,845 consensus PPIs with original human network. The original human PPI network could be grouped into six major cellular processes based on the GO annotations. Five of these groups (e.g. proteolysis, transport, signal transductions, nucleic acid metabolic processes) are the foundational process to maintain a living cell. As a result, although the number of proteins and PPIs are difference between these two networks, the consensus networks still keep these groups to maintain the foundational processes (Fig. 4-9). However, the mechanism of immune response is the organism-specific response. It has much difference from the fish to the mammalian. There are only few proteins and PPIs in the immune response region of the consensus network (green dot line in Fig. 4-9). This result indicates that the group of immune response is not the consensus region in three vertebrate protein interaction networks and it is consistence with the biological behaviors of immune response.

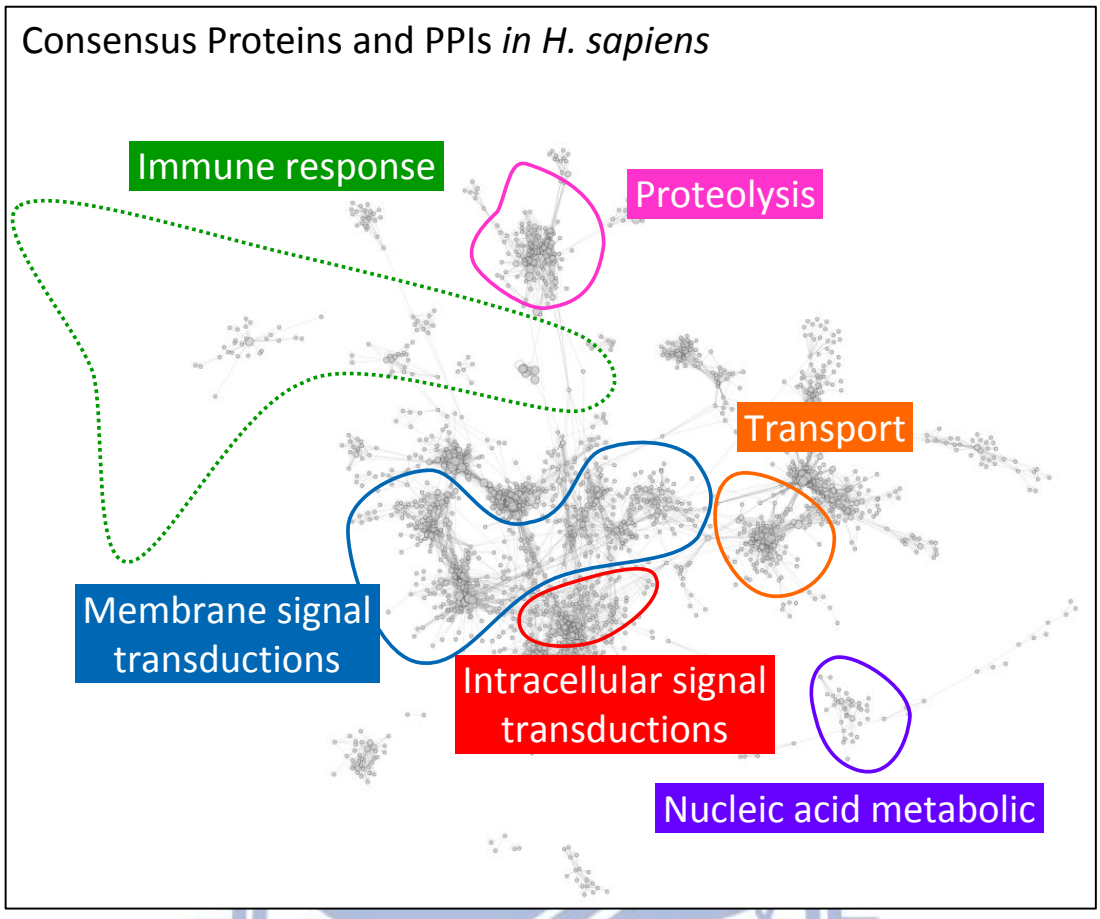


Figure 4-9. Six major cellular processes in consensus proteins and PPIs.

We identified 1,887 consensus proteins and 5,845 consensus PPIs from 4,135 proteins and 21,648 PPIs of structure resolved human network based on the "3d-domain interologs mapping". There are six major cellular processes in original human network based on the GO annotations (Figs. 4-6 and 4-7). The major difference of cellular process between original network and consensus PPIs is the immune response process. This result is consistence with that the mechanisms of immune response are difference from fish to the mammalian.

Finally, we investigated the relationship between the consensus proteins and network topology properties. In the scope of network analysis, there are various types of measures of the centrality of a protein (vertex) within a given network (graph) that determine the relative importance of a protein within the network. Because the consensus proteins often play an essential role in the biological processes, these proteins should have more relative importance among the network. Here, we used "*MS-matrix*" to evaluate the relative importance within a network based on the modularity structure property. The proteins with lower modularity structure properties are usually the hubs and locate on the central part of a network. This

modularity structure property is highly correlated to degree, betweenness centrality, and closeness centrality.

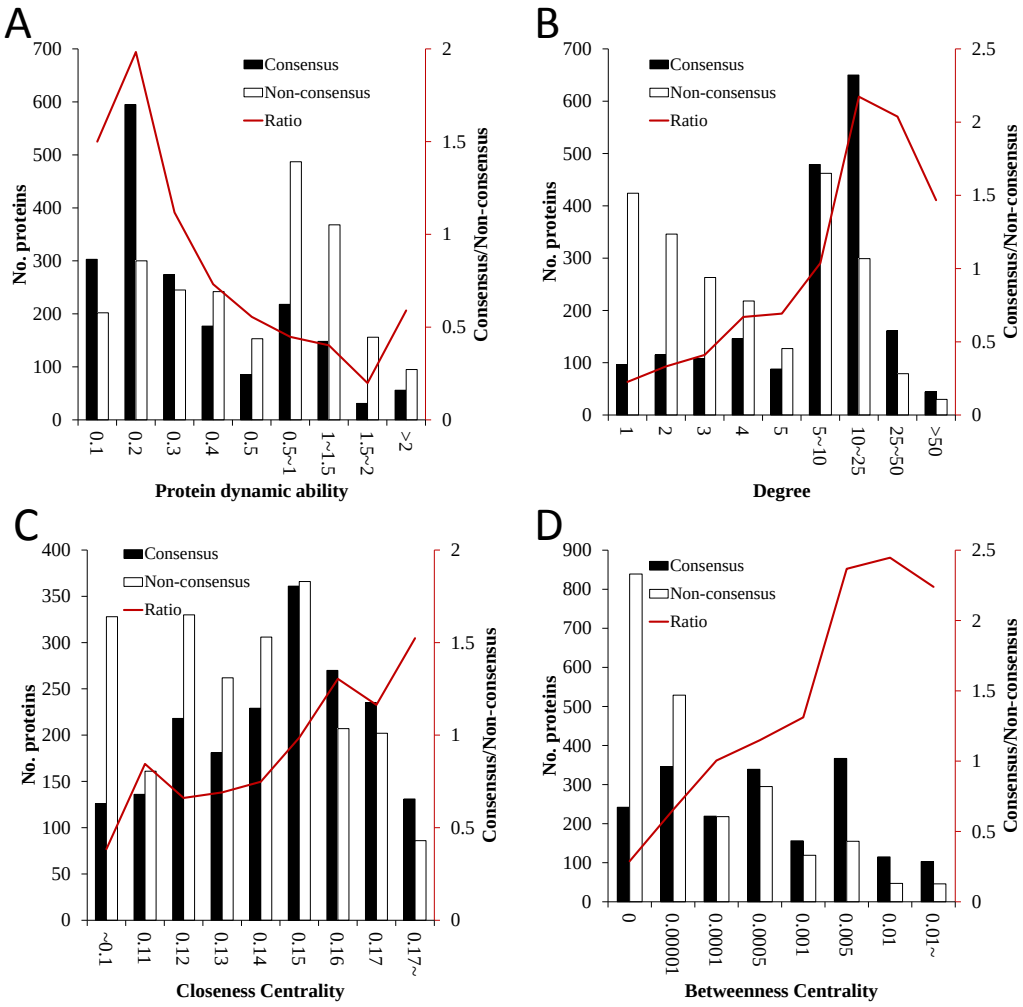


Figure 4-10. The distributions of protein dynamic ability (A), degree (B), closeness centrality (C), and betweenness centrality (D) in network of 1,887 consensus proteins

(A) The distributions of protein dynamic ability in protein interaction network of consensus and non-consensus proteins. The consensus proteins prefer to have lower dynamic abilities in network. On the contrary, the non-consensus proteins are not the central part of the network and have higher dynamic abilities. The proteins with higher degree (B), betweenness centralities (C), or closeness centralities (D) prefer to be conserved across three networks. However, the closeness centrality cannot distinguish consensus and non-consensus protein as well as degree or betweenness centrality.

Figure 4-10 shows the distributions of consensus and non-consensus protein dynamic ability in structurally resolved human protein interaction network. The consensus proteins prefer to have lower modularity structure properties in network. On the contrary, the

non-consensus proteins are not the central part of the network and have higher modularity structure properties. We also investigated the distributions of degree, closeness centrality, and betweenness centrality in consensus and non-consensus proteins (Fig. 4-10). The proteins with higher degree, betweenness centralities or closeness centralities prefer to conserve in these three networks. Our results imply that these consensus proteins among three vertebrate PPI networks usually have more relative importance with other proteins based on our "*MS-matrix*" and other graphic-based centrality properties. As a result, these consensus proteins are more relative importance according to that they are usually involved in several processes or are the regulated bridge between processes.

Disease related mutations in human network

Disease-related mutations can be roughly classified into two broad categories (i.e., in-frame and truncating mutations) [16]. Here, the in-frame mutations were considered as missense point mutations and the in-frame insertions or deletions are likely to produce full-length proteins with local defects. The truncating mutations including nonsense point mutations and frame-shift insertions or deletions often give rise to incomplete fragments. We collected 1,898 in-frame mutations and 304 truncating mutations on 124 and 35 proteins, respectively, in the structure resolved human network.

Previous studies have shown that the in-frame mutations can lead to loss of interactions [15]. To further evaluate the relationships between mutations and their associated disorders, we identified the positions of the disease-associated in-frame and truncating mutations on the corresponding proteins. Among the 1898 in-frame mutations, 427 mutations position on the contact residues that are important for PPIs. The disease-related mutations are significantly enriched with respect to the contact residues according to the odds ratio (Table 4-3 and Fig.

4-11A). Here, the odds ratio is calculated by [16]:

$$\text{Odds ratio} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

where p_1 is the fraction of observed mutations of the contact or non-contact residues; p_2 is the fraction of total sequence length. Here, the values of p_1 are 0.22 (427/1898; Table 4-3) and 0.77 (1471/1898) in the contact and non-contact residues, respectively. The values of p_2 are 0.087 (15,344/183,730) and 0.92 (168,386/183,730) in the contact and non-contact residues, respectively. Therefore, the odds ratios of contact residue and non-contact residues are 3.19 ([0.22/(1-0.22)]/[0.087/(1-0.087)]) and 0.31, respectively.

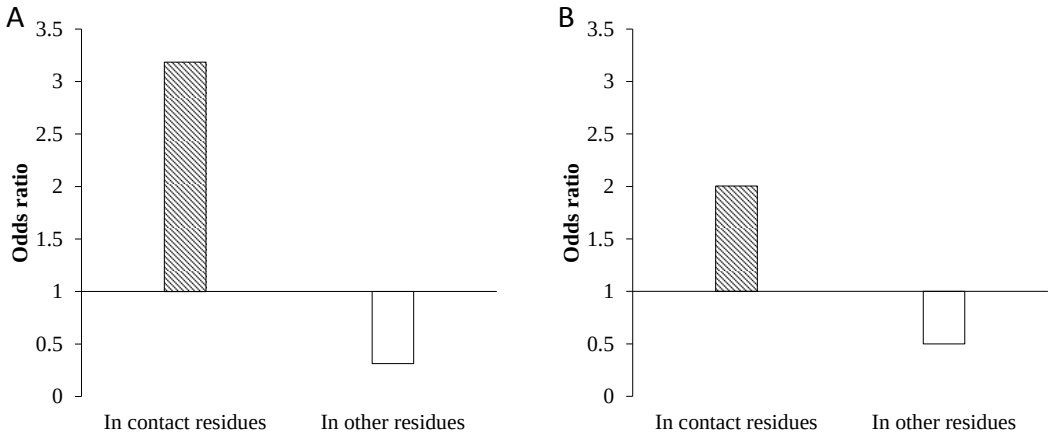


Figure 4-11. The odds ratios of in-frame and truncating mutations on the binding interface

The odds ratios for the distributions of (A) in-frame and (B) truncating mutations in. contact residues and non-contact residues on the protein interfaces in human protein interaction network. The in-frame mutations are enrichment in the contact residues with a high odds ratio (3.19).

Table 4-3. The distribution of in-frame and truncating mutations in human protein interaction network

	No. of protein	Sequence length	No. of contact residue	No. of non-contact residues	No. of mutation	No. mutations on the interfaces	No. of mutations not on the interface
In-frame mutation	124	183,730	15,344	168,386	1,898	427	1,471
truncating mutation	35	56,262	4,169	52,093	304	42	262

This result indicates that the contact residues of PPIs play an important role in diseases.

While a mutation is occurred and change the contact residue of the protein, it may influence the bind environment and lose the interaction to cause the corresponding disease. For example, the FGFR2 (821 amino acids) has 31 amino acids having the mutation data recorded in OMIM and has 14 disease syndromes most of which are related to the cancers. Then, the FGFR2 have 80 contacting residues based the PPIs derived from 3d-structures and "3d-domain interologs mapping". 13 mutations are contact residues and 18 mutations are non-contact residues. According to the definition of odd ratios, the odd ratios of contact residues is 6.69 and ([(13/31)/(1-13/31)]/[(80/821)/(1-80/821)]) is significant higher the non-contact residues (0.15). In addition, truncating mutations also have an enrichment on the contact residues (odds ratio = 2.0).

To further investigate the characteristic of disease-related mutations, we divided 427 in-frame mutations which locate on contact residues into four types, including the residues forming hydrogen-bond (H-bond), conserved residues, the conserved residues forming H-bond, the other residues. 335 mutations (78%) involved in forming hydrogen bonds on the interfaces; and 118 mutations (35%) are conserved and forming hydrogen binds. On the contrary, 57 (13%) mutations neither are conserved residues and nor involve in hydrogen bonds. For example, the 13 mutations involving in the PPI interface of FGFR2 have 7 mutations involving in H-bond, 2 mutations which are conserved residues, and 1 mutation both involving in H-bond and conserved residue. These results indicate that the disease-related mutations are usually located on the contact residues forming the hydrogen bonds within PPIs.

Disease-related consensus pathways

According to our structure resolved human network with mutations and diseases, two major groups of proteins are highly involved in cancers and cardiovascular-related diseases

(Figs. 4-1C, 12, and 13 and Table 4-4). The proteins, such as fibroblast growth factors (FGF), fibroblast growth factor receptors (FGFR) and protein kinases, involves in cancers. FGF and FGFR regulate some key biological processes, such as cell proliferation, survival, migration, and differentiation both during development and in the adult [99]. The FGFR2 and FGFR3 are the top-rank proteins with the numbers (i.e. 14 and 13 diseases recorded in OMIM; Fig. 4-1C) of annotated diseases.

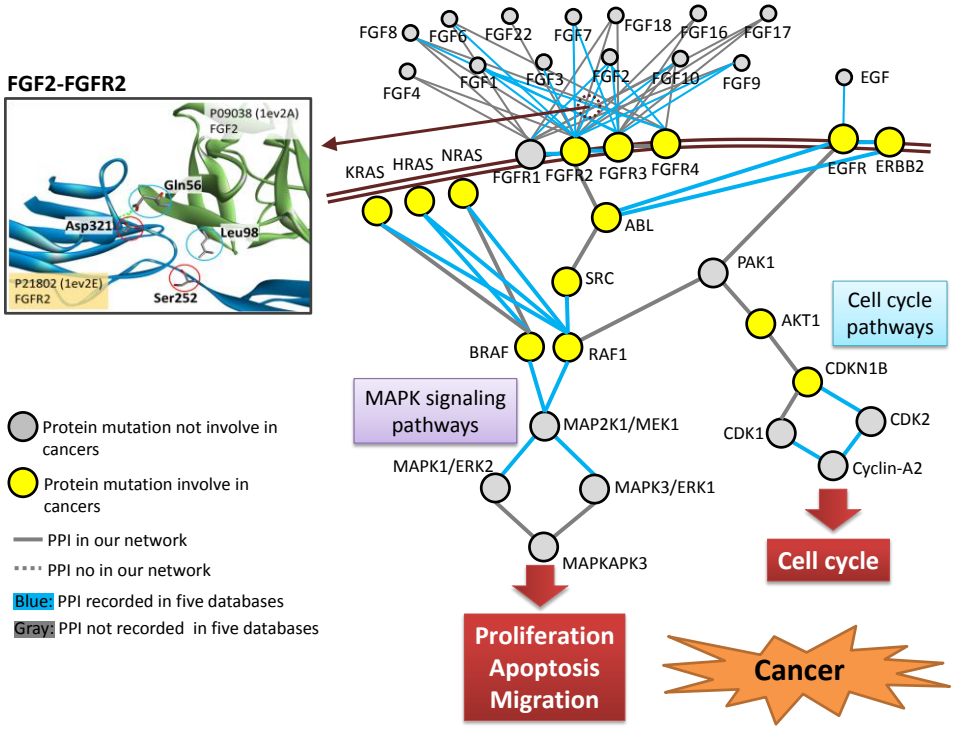


Figure 4-12. The pathways and proteins involved in a great amount of diseases, especially the cancers

The proteins colored with yellow are the proteins which have mutation data recorded in OMIM. The MAPK1/ERK2 and MAPK3/ERK1 play essential roles in several important pathways (e.g. proliferation and apoptosis) related to the cancer. FGFR2-FGF2 is one of the upstream regulating PPI of these pathways. The mutations of FGFR2 may influence the interaction of FGFR2-FGF2 and cause the defects of it corresponding pathways.

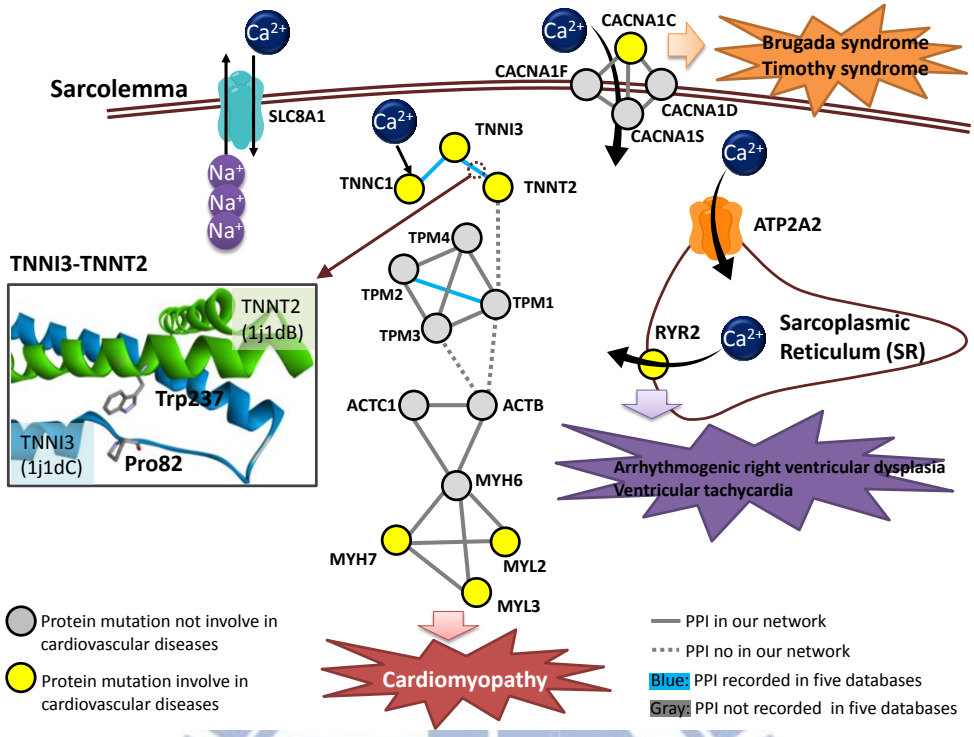


Figure 4-13. The pathways and proteins involved in a great amount of diseases, especially the cardiovascular-related diseases

The proteins colored with yellow are the proteins which have mutation data recorded in OMIM. The mechanisms of cardiovascular-related diseases are highly related to the regulation of calcium ion. There are three pathways to regulate the concentration of calcium ion. The first one is the sodium/calcium exchanger 1 (SLC8A1) which can rapidly transport Ca^{2+} during excitation-contraction coupling. The second one is voltage-dependent calcium channel. The third one regulates the calcium concentration by the transports Ca^{2+} between sarcoplasmic and endoplasmic reticulum (e.g. ATP2A2 and RYR2). Mutations of these Ca^{2+} transports could cause the ventricular tachycardia, Brugada syndrome, and Timothy syndrome. In addition, the mutations on the proteins of cardiac muscle contraction pathway could cause the cardiomyopathy.

Table 4-4. The diseases recorded in OMIM of each protein in FGF-FGFR and upstream proteins of MAPK1 and MAPK3

Gene name	UniProt AC	Diseases recorded in OMIM	Involved in cancer	Conserved in networks
BRAF	P15056	1. Adenocarcinoma of lung 2. Cardiofaciocutaneous syndrome 3. Colorectal cancer 4. LEOPARD syndrome 5. Melanoma, malignant 6. Nonsmall cell lung cancer 7. Noonan syndrome	ˇ	ˇ
EGFR	P00533	1. Adenocarcinoma of lung 2. Nonsmall cell lung cancer	ˇ	ˇ
EGF	P01133	1. Hypomagnesemia		ˇ
ERBB2	P04626	1. Adenocarcinoma of lung, somatic 2. Gastric cancer 3. Glioblastoma	ˇ	ˇ

		4. Ovarian cancer		
FGF10	O15520	1. Aplasia of lacrimal and salivary glands 2. LADD syndrome		˘
FGF3	P11487	1. Deafness, congenital with inner ear agenesis, microtia, and microdontia		˘
FGF8	P55075	1. Kallmann syndrome		˘
FGF9	P31371	1. Multiple synostoses syndrome		˘
FGFR1	P11362	1. Hypogonadotropic hypogonadism 2. Jackson-Weiss syndrome 3. Kallmann syndrome 4. Osteoglophonic dysplasia 5. Pfeiffer syndrome 6. Trigonocephaly		˘
FGFR2	P21802	1. Antley-Bixler syndrome without genital anomalies or disordered steroidogenesis 2. Apert syndrome 3. Beare-Stevenson cutis gyrata syndrome 4. Bent bone dysplasia syndrome 5. Craniofacial-skeletal-dermatologic dysplasia 6. Craniosynostosis 7. Crouzon syndrome 8. Gastric cancer 9. Jackson-Weiss syndrome 10. LADD syndrome 11. Pfeiffer syndrome 12. Saethre-Chotzen syndrome 13. Scaphocephaly and Axenfeld-Rieger nomaly 14. Scaphocephaly, maxillary retrusion, and mental retardation	˘	˘
FGFR3	P22607	1. Achondroplasia 2. Bladder cancer 3. CATSHL syndrome 4. Cervical cancer 5. Colorectal cancer 6. Crouzon syndrome with acanthosis 7. Igricans 8. Hypochondroplasia 9. LADD syndrome 10. Muenke syndrome 11. Nevus, keratinocytic, nonepidermolytic 12. Spermatocytic seminoma 13. Thanatophoric dysplasia, type I 14. Thanatophoric dysplasia, type II	˘	˘
FGFR4	P22455	1. Cancer progression/metastasis	˘	˘
HRAS	P01112	1. Costello syndrome 2. Bladder cancer 3. Thyroid carcinoma	˘	˘
KDR	P35968	1. Hemangioma	˘	˘
KRAS	P01116	1. Bladder cancer 2. Breast cancer 3. Cardiofaciocutaneous syndrome 4. Gastric cancer 5. Leukemia, acute myelogenous 6. Lung cancer 7. Noonan syndrome 8. Pancreatic carcinoma	˘	˘
MAP2K1	Q02750	1. Cardiofaciocutaneous syndrome		˘
NRAS	P01111	1. Autoimmune lymphoproliferative syndrome type IV 2. Colorectal cancer 3. Noonan syndrome	˘	˘

		4. Thyroid carcinoma, follicular		
RAF1	P04049	1. LEOPARD syndrome 2. Noonan syndrome		ˇ
RET	P07949	1. Central hypoventilation syndrome, congenital 2. Medullary thyroid carcinoma 3. Multiple endocrine neoplasia IIA 4. Multiple endocrine neoplasia IIB 5. Pheochromocytoma 6. Renal agenesis 7. Hirschsprung disease	ˇ	ˇ
SOS1	Q07889	1. Fibromatosis, gingival 2. Noonan syndrome	ˇ	ˇ
SRC	P12931	1. Colon cancer	ˇ	ˇ

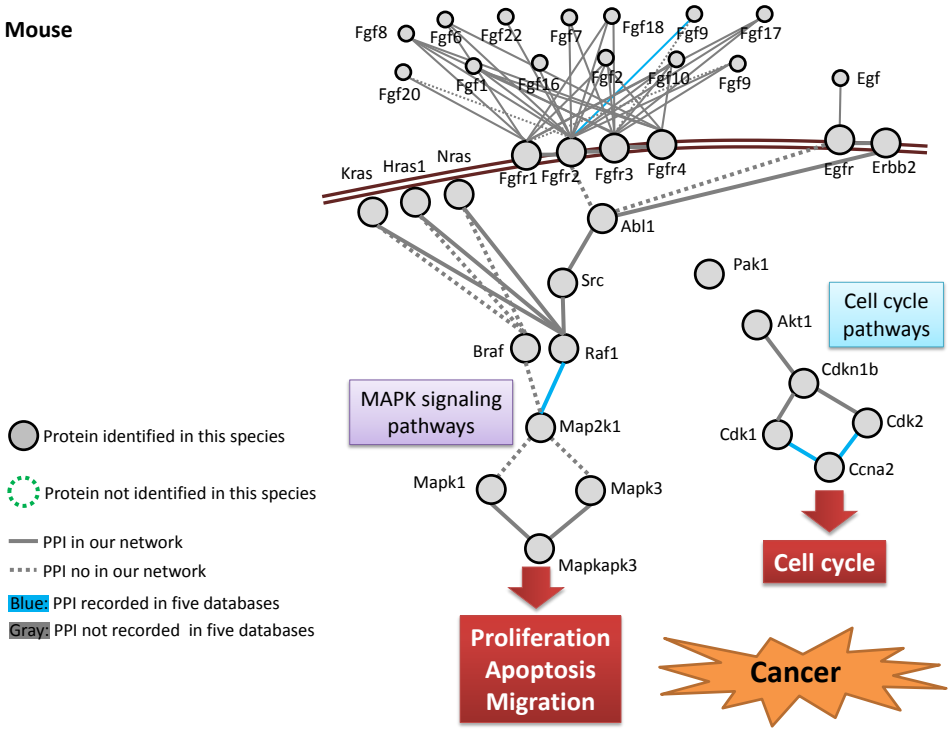


Figure 4-14. The mapping pathways and proteins which are related to the cancers of *M. musculus*.

Figures 4-12, 4-14, and 4-15 show the series of protein kinases and upstream of MAPK3/ERK1 and MAPK1/ERK2 in our derived networks of *H. sapiens*, *M. musculus*, and *D. rerio*, respectively. These two MAPKs are highly related to the cancer by involving diverse biological functions and critical pathways such as cell growth, adhesion, survival and differentiation [78,100]. In addition, the RAF and B-RAF, which regulate MAPKKK of ERK pathway, act as a regulatory link between the upstream signal proteins (e.g. membrane-associated Ras GTPases (i.e. KRAS, NRAS, and HRAS) and non-receptor protein

tyrosine kinase (e.g. SRC)) and the MAPK/ERK cascade. SRC can be activated by the EGFR and ERBB2 in the ERBB signal pathway for adhesion and migration [78]. The PPIs colored blue are also recorded in five public PPI databases (e.g. IntAct[8], DIP[9], MIPS[28], BioGRID[10], and MINT[29]). The PPIs with the dot line are not identified by structure template and "3D-domain interolog mapping" but have been recorded in PPI databases or KEGG [78] database.

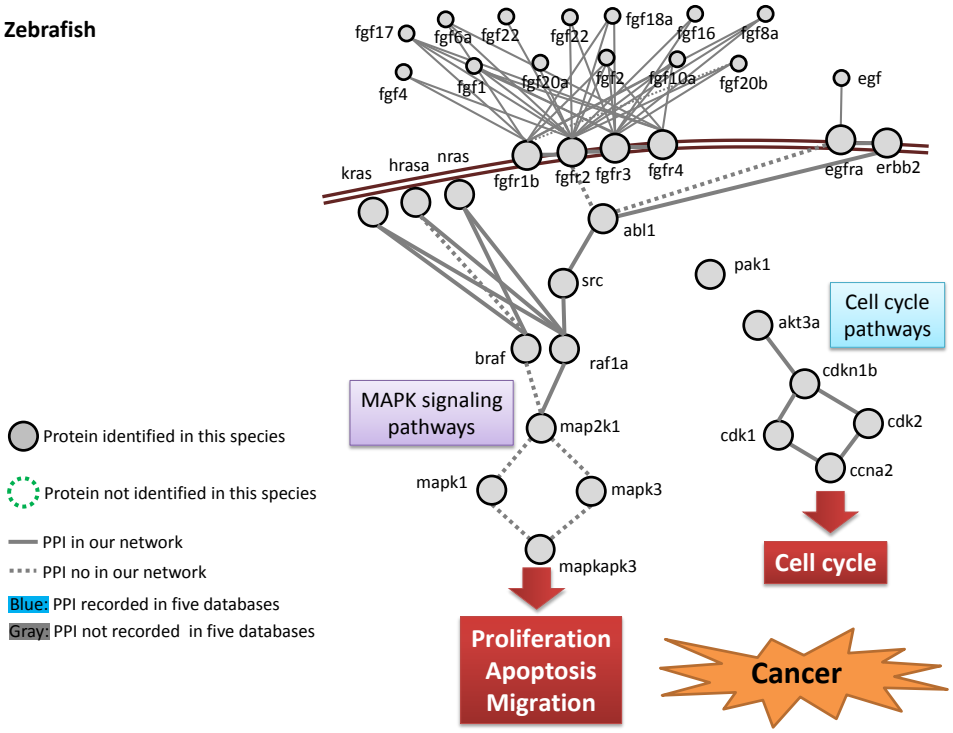


Figure 4-15. The mapping pathways and proteins which are related to the cancers of *D. rerio*.

There are 65 PPIs among the pathway of human network. 32 of 65 PPIs are also recorded in 5 public PPI databases. Although 33 PPIs are not recorded in databases, previous studies [101,102] have been indicated several protein pairs (e.g. FGF4-FGFR1, FGF4-FGFR2, FGF6-FGFR1, and FGF6-FGFR4) should be interacting protein pairs. In addition, there are only 3 PPIs and non PPIs recorded in public PPI databases among the pathways of *M. musculus* and *D. rerio*, respectively. This result implied that our structural networks can annotate and infer the cell behaviours of a new determined (or seldom-studied) species (e.g. zebrafish), by mapping some well-studied species.

Furthermore, the mechanisms of cardiovascular-related diseases are highly related to the regulation of calcium ion. There are three major pathways to regulate the concentration of calcium ion (Figs. 4-13, 4-16, and 4-17). One is the sodium/calcium exchanger 1 (SLC8A1) which can rapidly transport Ca^{2+} during excitation-contraction coupling. Another one is voltage-dependent calcium channel which transport Ca^{2+} without Na^{+} exchanging. The other one regulates the calcium concentration by the transports Ca^{2+} between sarcoplasmic and endoplasmic reticulum (e.g. ATP2A2 and RYR2). Mutations of these Ca^{2+} transports could cause the ventricular tachycardia, Brugada syndrome, and Timothy syndrome. Then, the myosins and actins are the major proteins of cardiac muscle contraction pathway. The mutations of these proteins could cause the cardiomyopathy.

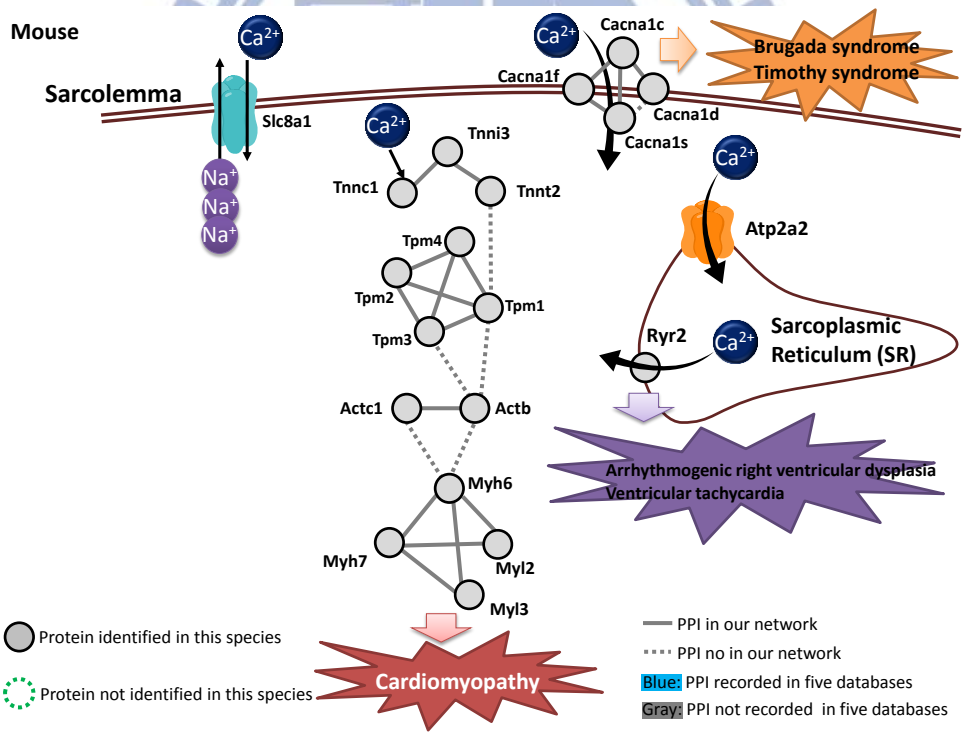


Figure 4-16. The mapping pathways and proteins which are involved in the cardiovascular-related diseases of M. musculus.

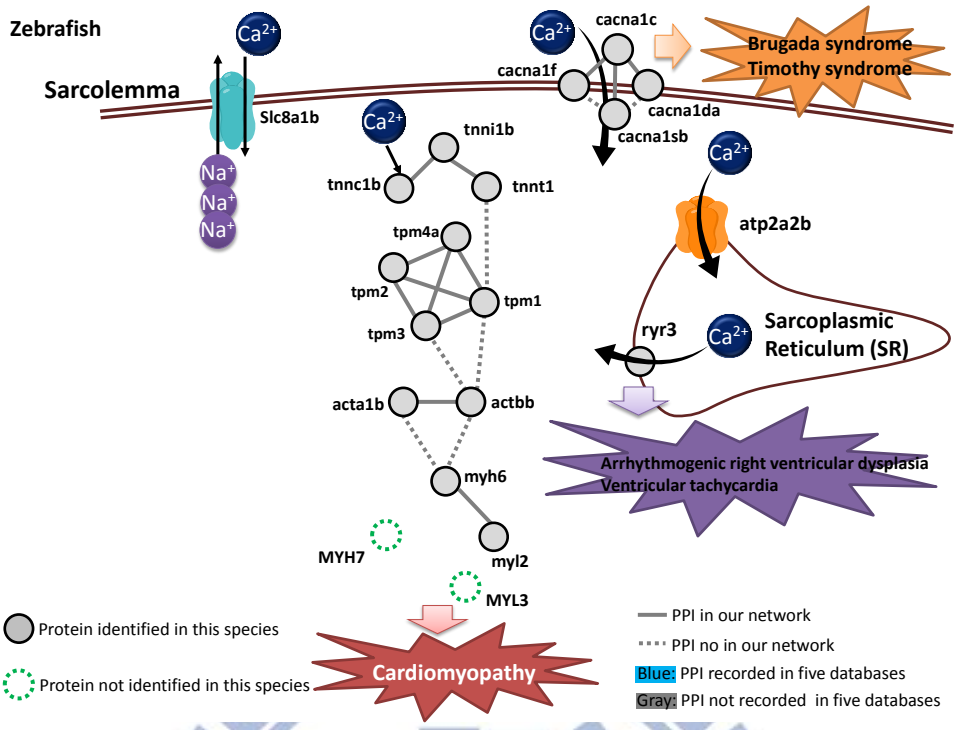


Figure 4-17. The mapping pathways and proteins which are involved in the cardiovascular-related diseases of *D. rerio*.

These proteins involving in cancers and cardiovascular-related diseases are conserved in three vertebrate PPI networks. Here, we used the FGF2-FGFR2, TNNI3-TNNT2, and F2-SERPINA5 as examples to explain the relationship between mutations, corresponding consensus pathways among three vertebrate PPI networks. In addition, our methods also provided the organism-specific proteins and its corresponding pathways. Here, we used the F2-SERPINA5 which only appear in human and mouse networks and involve in complement and coagulation pathway as an example to describe the organism-specific pathway among our derived networks.

Disease related mutations in the binding interface of FGF2-FGFR2

A mutated Fibroblast growth factor receptor 2 (FGFR2) could cause endometrial, gastric cancer, or pfeiffer syndrome [103,104]. Among 187 samples of endometrial carcinoma, previous work shows that seven somatic S252W mutations (the most common FGFR2 mutation) were

the endometrioid subtype and one S252W mutation was the serous subtype [105]. Ibrahimi et al. demonstrated that the D321A mutation increased the binding affinity between FGFR2c and the FGFs expressed in the cranial suture [106].

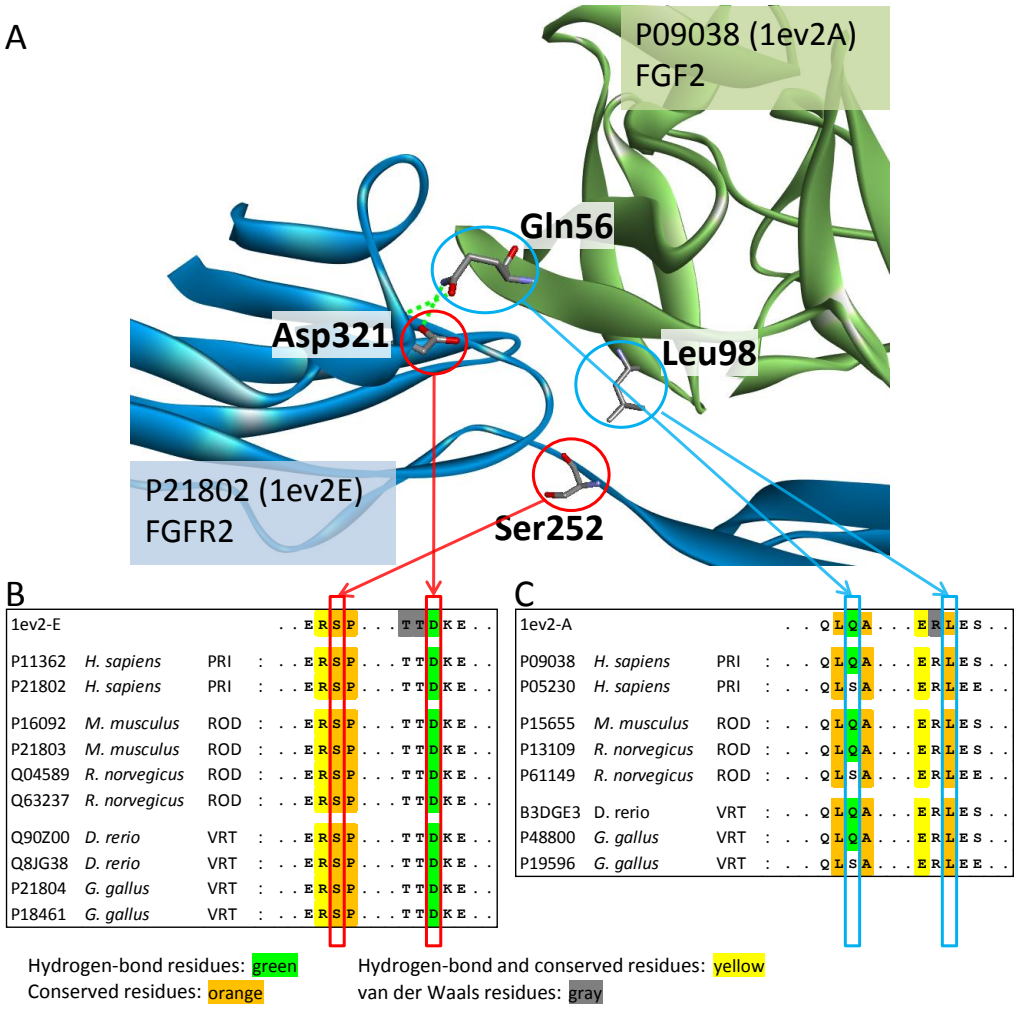


Figure 4-18. Binding models and multiple sequence alignments of PPI family derived from FGF2-FGFR2 heterodimer (PDB code: 1ev2)

(A) The atomic binding model with the highlight van der Waals and hydrogen-bond interaction of Asp321-Gln56 and Ser252-Leu98, respectively. (B) Multiple sequence alignments of PPI family of the interface E (FGFR2)-A (FGF2).

According to the FGFR2-FGF2 binding interface of the structure template (PDB code: 1ev2 [107]) , the Ser252 and Asp321 are the contact residues of FGFR2 on the FGF2-FGFR2 binding interface (Fig. 4-18). The Ser252 forms a conserved van der waals interaction to the Leu98 of FGF2 according to the PPI family of this template (Fig. 4-18B), and the Asp321

forms a hydrogen-bond interaction with the Gln56 (Fig 4-18C). Because the FGF2-FGFR2 is the upstream interactions of MAPK3/ERK1 and MAPK1/ERK2, the S252W mutation influences the cell proliferation and apoptosis in the ERK pathway for the endometrial cancer (Fig. 4-12). In addition, the interaction residues Ser252 and Leu98 are conserved on the PPI (FGF2-FGFR2) of three vertebrate PPI networks.

Cardiovascular-related diseases and its corresponding pathways

Figure 4-13 shows the cardiovascular-related diseases and its corresponding pathways. The proteins colored with yellow are the proteins which have mutation data recorded in OMIM. All of three type regulations for the concentration of calcium ion could be identified in our derived networks. In addition, the pathway for cardiac muscle contraction pathway could also be identified in our derived networks. The mutations of these proteins (e.g. TNNC1 [108] and MYH7 [109,110]) could cause the cardiomyopathy. According to KEGG database [78], there are three complexes, including cardiac troponin complex, TPM complex, and Actin-Myosin complex, involved in the muscle contraction pathway. All of these complexes of human have the experimental data (solid line in Fig. 4-13) recorded in five public databases. In addition, our method could provide the binding mechanism by the 3d-strucutre and "3d-domain interologs mapping". For example, the cardiac troponin (cTn) has an important function for cardiac muscle contraction, which is a complex of three subunits, including cardiac troponin C (cTnC, TNNC1), troponin I (cTnI, TNNI3) and troponin T (cTnT, TNNT2) [111]. According to binding interface of the structure template (PDB code: 1j1d), the Pro82 is a contact residue of TNNI3 on the TNNI3-TNNT2 binding interface. The Pro82 of TNNI3 can form a conserved van der waals interaction to the Trp237 of TNNT2 in the PPI family of this template (Fig. 4-19).

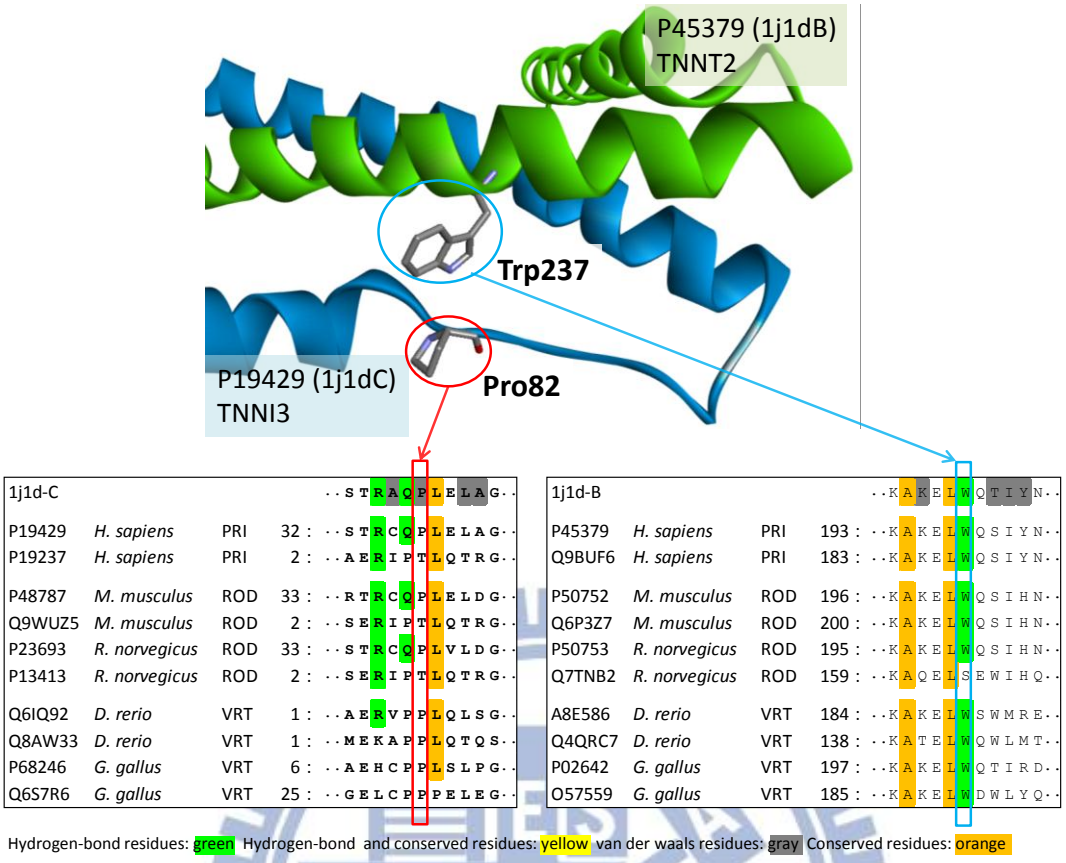


Figure 4-19. Binding models and multiple sequence alignments of PPI family derived from TNNT2-TNNI3 heterodimer (PDB code: 1j1d)

(A) The atomic binding model with a highlight van der Waals interaction of Trp237-Pro82. (B) Multiple sequence alignments of PPI family of the interface B(TNNT2)-C(TNNI3).

However, there are still some experimental PPIs (dot line in Fig. 4-13) cannot be annotated the binding mechanism by currently 3d-strucutre recorded in PDB. The construction of our derived PPI networks largely relies on the availability of 3D-crystal structures, which limits the coverage of our networks. However, there are no PPI within these complexes of *Mus musculus*, and *Danio rerio* recorded in five public databases (Figs. 4-16 and 4-17). Our methods are able to construct these complexes and pathways on the non-well-known organisms (e.g. zebrafish). It may provide a new insight for understanding the cardiovascular-related diseases based on using the animal model with our derived networks.

Complement and coagulation pathway in human

The complement system is a proteolytic cascade in blood plasma and a mediator of innate immunity [78]. According to the complement and coagulation pathway among our derived network (Fig. 4-20), there are several proteins (yellow node in Fig. 4-20) which are involved in blood coagulation related diseases. One of these proteins is prothrombin (F2), which is activated to the thrombin by coagulation factor X (F10). The activated thrombin plays important roles in hemostasis and thrombosis, and it converts fibrinogen to fibrin for blood clot formation; stimulates platelet aggregation, and activates coagulation factors V (F5), VIII (F8), and XIII (F13). A mutated F2 could case dysprothrombinemia, hypoprothrombinemia, or thrombophilia [112]. Poort et al. described a common genetic variation in the 3-prime untranslated region of the prothrombin gene that is associated with elevated plasma prothrombin levels and an increased risk of venous thrombosis [113].

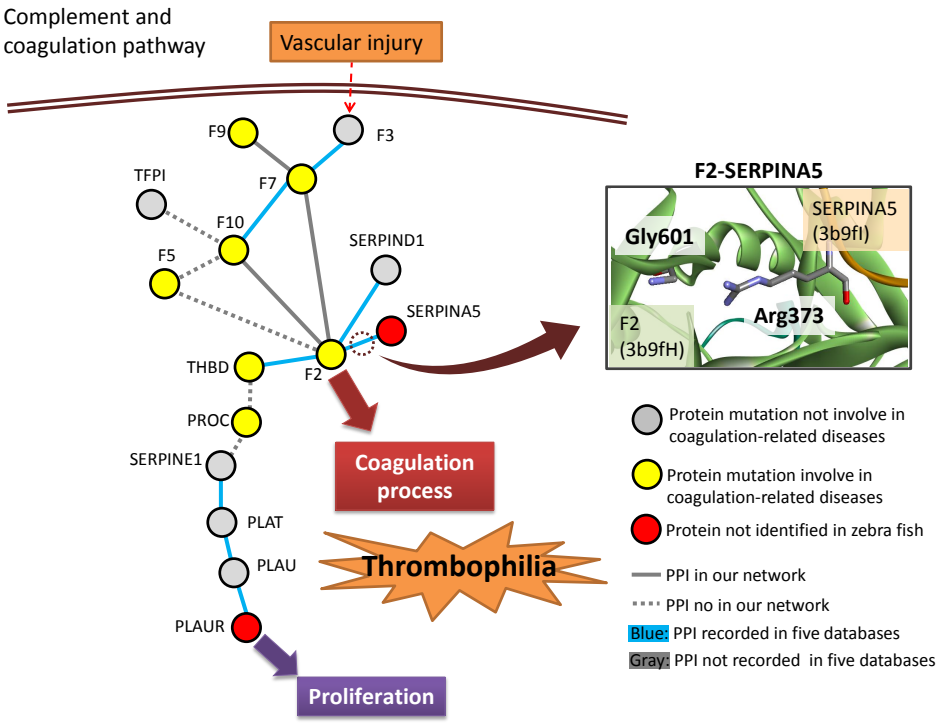


Figure 4-20. The specific proteins among the complement and coagulation pathway

The proteins colored with yellow are the proteins which have mutation data recorded in OMIM. The SERPINA5 and PLAUR are the specific proteins among the complement and coagulation pathway and are colored with red. SERPINA5 could inhibit the F2 which could be activated by coagulation factor X (F10) and plays important roles in hemostasis and thrombosis. A mutated F2 could case thrombophilia, dysprothrombinemia, or hypoprothrombinemia.

In addition, there are two specific proteins SERPINA5 and PLAUR both not identified in zebra fish among the complement and coagulation pathway. The PLAUR plays a central role in cell migration and tissue remodeling processes and has so far only been identified in mammals [112]. Moreover, SERPINA5 could inhibit the F2 and play hemostatic roles in the blood plasma, which appears to be a regulatory factor in blood coagulation and fibrinolysis [114]. According to binding mechanism derived from the interface of the structure template (PDB code: 3b9f), the Gly601 is a contact residue of F2 on the F2-SERPINA5 binding interface (Fig. 4-21). The Gly601 of F2 can form a conserved van der waals interaction to the Arg373 of SERPINA5 in the PPI family of this template (Fig. 4-21).

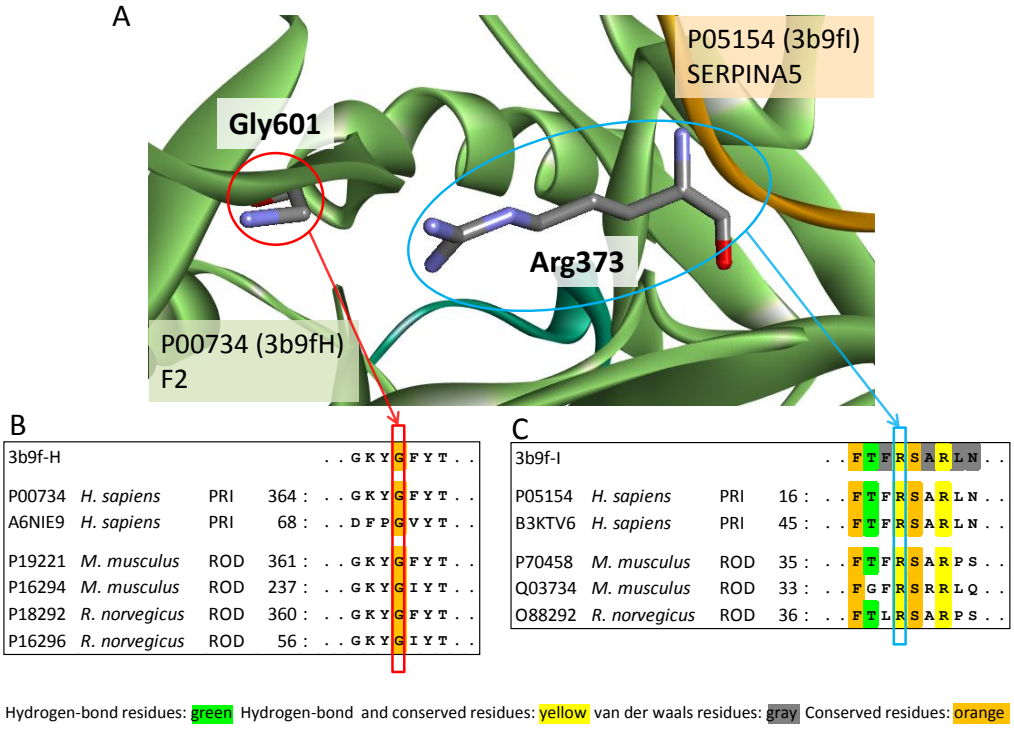


Figure 4-21. Binding models and multiple sequence alignments of PPI family derived from F2-SERPINA5 heterodimer (PDB code: 3b9f)

(A) The atomic binding model with a highlight van der Waals and conserved interaction of Gly601-Arg373. (B) Multiple sequence alignments of PPI family of the interface H(F2)-I(SERPINA5).

Therefore, the Gly601 mutation may cause the dysprothrombinemia by influencing the fibrin degradation process in this pathway (Fig. 4-20). We also suggested that the different

pathway of complement and coagulation was existed in zebra fish compared with human and mouse (Figs. 4-20 and 4-22); because SERPINA5 are not found in zebra fish (Fig. 4-23).

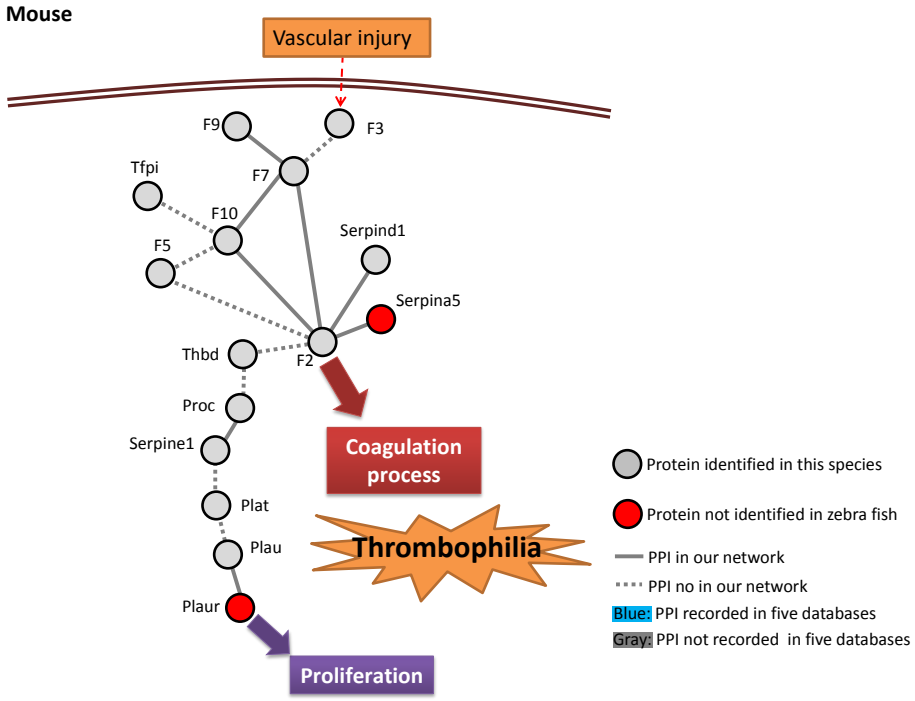


Figure 4-22. The mapping pathways and proteins involved in the complement and coagulation pathway of *M. musculus*.

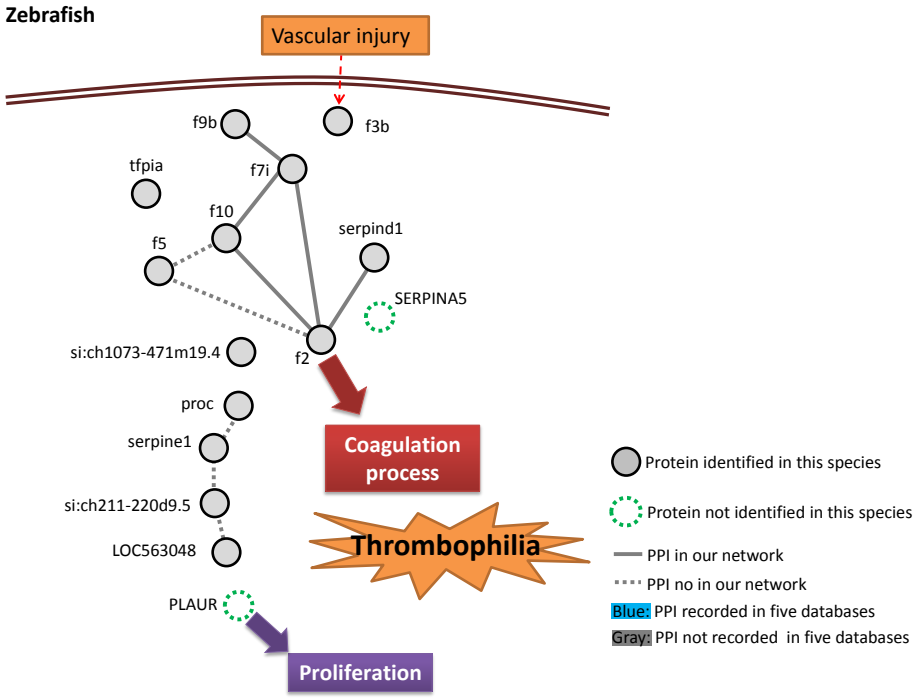


Figure 4-23. The mapping pathways and proteins involved in the complement and coagulation pathway of *D. rerio*.

4-4. Conclusions

This work is the first to construct structure resolved PPI networks across multiple species, including *H. sapiens*, *M. musculus*, and *D. rerio*. According to structure-based homologous PPIs in multiple species, the PPIs with atomic residue-based binding models in the derived structure resolved network achieved highly agreement with Gene Ontology (BP, CC, and MF terms) similarities. Furthermore, the architecture of these networks is a scale-free network which is consistent with most of the cellular networks. Experimental results also indicate that the mutations of interacting residues on the PPIs often related to diseases are often on. Our results demonstrate that the structure resolved PPI networks can provide valuable insights for understanding the mechanisms of biological processes.

The construction of our structurally resolved PPI networks largely relies on the availability of 3D-crystal structures, which limits the coverage of our network. However with the rapid growth of PDB, more 3D-crystal information will become available and our methods can be readily applied to uncover potential molecular mechanisms whose structural information is currently missing. More importantly, our structural networks can annotate and infer the cell behaviours of a new determined (or seldom-studied) species (e.g. zebrafish), by mapping some well-studied species. In addition, our methods can also be used to observe the consensus proteins and modules (a fundamental unit forming with highly connected proteins) which are high conserved appearing in multiple organisms. These consensus proteins are often the essential genes and related to diseases recorded in OMIM.

Chapter 5. Modularity structure matrix for investigating protein interaction network

A crucial step toward understanding cellular systems properties is to analyze the topology of biological networks and biochemical progress in cells. Many graphic features are purposed to measure the role of proteins and identify local modularity structures of high connectivity in a PPI network. Laplacian matrix is a matrix representation of a given network. Here, we proposed the modularity structure matrix (***MS-matrix***), which is the pseudoinverse of the Laplacian matrix for describing the kernels on a graph, to evaluate the modularity structure properties of a PPI network. According to our knowledge, the modularity structure property is the first property to identify both global important proteins and local modularity structures within a network. For a given PPI network of *S. cerevisiae*, our results demonstrate that the important proteins identified by the *MS-matrix* are related to the essential biological processes (i.e. essential genes) and highly consistence with the topology features (i.e. degree, closeness centrality, and betweenness centrality). Then, the relationship between proteins derived from the *MS-matrix* could reflect the similarity of Gene Ontology and could be useful for the module identification. Furthermore, biological characterization (e.g. Gene Onotology) of the modules derived from the *MS-matrix* is similar to the modules collected from the experiment database (e.g. MIPS). Our results demonstrate that the *MS-matrix* would provide the insight for investigating a PPI network through important proteins and local modularity structures.

5-1. Introduction

A crucial step toward understanding cellular systems properties is to analyze the topology of biological networks and biochemical progress in cells [3,83]. To construct the protein-protein

interaction (PPI) network as completely as possible, genome-scale interaction discovery approaches, such as high-throughput yeast two-hybrid screening [25,26] and coaffinity purification [27] , have been proposed. Because of the complexity of a PPI network, many graphic features (e.g. degree, closeness centrality, and betweenness centrality) are purposed to measure the role of proteins in a PPI network [115]. In addition, several agglomerative algorithmic approaches [116,117] have been developed to identify local modularity structures of high connectivity with relatively low connectivity to the rest of network. These dense sub-graphs are treated as potential functional modules.

In the mathematical and computational field of graph theory, the Laplacian matrix (or Kirchhoff matrix) is a matrix representation of a graph. In addition, the pseudoinverse of the Laplacian matrix plays a key role, has a nice interpretation in terms of random walk on a graph, and defines the kernels on a graph [118]. Its application on biological field, the Gaussian network model has succeeded in describing the local modularity structures (e.g. flexible/rigid regions and domains of proteins) and the important residues of a given protein [119,120]. However, a PPI network, which has the functional local modularity structures (i.e. module and complex) and the important hubs, is similar to the behaviors of a protein.

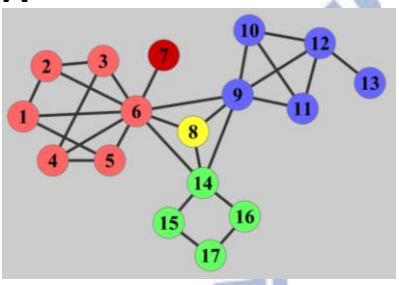
To address these issues, we proposed the ***MS-matrix*** to evaluate the modularity structure property within a PPI network. According to our knowledge, the *MS-matrix* is the first property to identify both global important proteins and local modularity structures within a network. For a given PPI network of *S. cerevisiae*, our results demonstrate that the important proteins identified by the *MS-matrix* are related to the essential biological processes (i.e. essential genes). In addition, the important proteins derived from *MS-matrix* are highly consistence with the topology features (i.e. degree, closeness centrality, and betweenness centrality). Then, the relationship between proteins derived from the *MS-matrix* could reflect the similarity of Gene Ontology and could be useful for the module identification. Furthermore, biological

characterization (e.g. Gene Onotology) of the modules derived from the *MS-matrix* is similar to the modules collected from the experiment database (e.g. MIPS). Our results demonstrate that the *MS-matrix* would provide the insight for investigating a PPI network through important proteins and local modularity structures.

5-2. Methods

Modularity structure matrix

A



B

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	3	-1	0	0	-1	-1	0	0	0	0	0	0	0	0	0	0	0
2	-1	3	-1	0	0	-1	0	0	0	0	0	0	0	0	0	0	0
3	0	-1	3	-1	0	-1	0	0	0	0	0	0	0	0	0	0	0
4	0	0	-1	3	-1	-1	0	0	0	0	0	0	0	0	0	0	0
5	-1	0	0	-1	3	-1	0	0	0	0	0	0	0	0	0	0	0
6	-1	-1	-1	-1	-1	9	-1	-1	-1	0	0	0	0	-1	0	0	0
7	0	0	0	0	0	-1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	-1	0	3	-1	0	0	0	0	-1	0	0	0
9	0	0	0	0	0	-1	0	-1	6	-1	-1	-1	0	-1	0	0	0
10	0	0	0	0	0	0	0	0	-1	3	-1	-1	0	0	0	0	0
11	0	0	0	0	0	0	0	0	-1	-1	3	-1	0	0	0	0	0
12	0	0	0	0	0	0	0	0	-1	-1	-1	4	-1	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	-1	1	0	0	0	0
14	0	0	0	0	0	-1	0	-1	-1	0	0	0	0	5	-1	-1	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	2	0	-1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	2	-1
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-1	2

C

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0.5	0.2	0.2	0.2	0.2	0.1	0.1	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0
2	0.2	0.5	0.2	0.2	0.2	0.1	0.1	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0
3	0.2	0.2	0.5	0.2	0.2	0.1	0.1	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0
4	0.2	0.2	0.2	0.5	0.2	0.1	0.1	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0
5	0.2	0.2	0.2	0.2	0.5	0.1	0.1	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0
6	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0	-0	-0	-0	-0	-0	-0	-0	-0	-0
7	0.1	0.1	0.1	0.1	0.1	0.1	1.1	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0
8	-0	-0	-0	-0	-0	0	-0	0.4	0.1	-0	-0	-0	-0	0.1	-0	-0	-0
9	-0	-0	-0	-0	-0	-0	-0	0.1	0.2	0.2	0.2	0.2	0.1	0	-0	-0	-0
10	-0	-0	-0	-0	-0	-0	-0	-0	0.2	0.6	0.3	0.3	0.3	-0	-0	-0	-0
11	-0	-0	-0	-0	-0	-0	-0	-0	0.2	0.3	0.6	0.3	0.3	-0	-0	-0	-0
12	-0	-0	-0	-0	-0	-0	-0	-0	0.2	0.3	0.3	0.6	0.5	-0	-0	-0	-0
13	-0	-0	-0	-0	-0	-0	-0	-0	0.1	0.3	0.3	0.5	1.4	-0	-0	-0	-0
14	-0	-0	-0	-0	-0	-0	-0	0.1	0	-0	-0	-0	-0	0.3	0.2	0.2	0.2
15	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	0.2	0.8	0.3	0.6
16	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	0.2	0.3	0.8	0.6
17	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	0.2	0.6	0.6	1

Figure 5-1. The overview of the evaluating the importance of each node in a simple network through the "*MS-matrix*"

(A) A simple network with three local density regions (red, blue and green nodes). (B) *Laplacian matrix* of the simple network. (C) *MS-matrix* is derived from the pseudo-inverse of *Laplacian matrix*.

Here, we consider a PPI network as an undirected graph. The Laplacian matrix is a matrix representation of a graph. Here, we use a simple network (Fig. 5-1A) with 17 proteins to construct the Laplacian matrix and introduce the ***MS-matrix***. First, we construct the *Laplacian*

matrix M (Fig. 5-1B) for the network. The M_{ij} is given as

$$M_{ij} = \begin{cases} -1, & \text{if } i \neq j \text{ and protein i interacts with protein j} \\ 0, & \text{if } i \neq j \text{ and protein i not interact with protein j} \\ k, & \text{if } i = j, k \text{ is the degree of protein i} \end{cases} \quad (1)$$

For example, the degree of node 8 is 3 (interacting with node 6, 9, and 14); and the $M_{8\,8}$, $M_{6\,8}$, $M_{8\,9}$, and $M_{8\,14}$ are 4, -1, -1, and -1, respectively. Then, the ***MS-matrix*** (*MS*) (Fig. 5-1C) is the pseudoinverse of *Laplacian matrix M*. Here, we got the pseudoinverse of *Laplacian matrix* based on the Scientific Tools for Python (SciPY).

According to the local modularity structure (MS_{ij}), these 17 proteins in this matrix ***MS*** can be clustered into three local modularity structures matching with the original network (red, blue and green regions). Additionally, the three lowest diagonal values (nodes 6, 9 and 14) of ***MS-matrix*** (MS_{ii}) are the centrality nodes; conversely, two highest values (nodes 7 and 13) of MS_{ii} are the peripheral nodes. These results are highly consistent with the graphic features, such as degree, closeness and betweenness centrality (Table 5-1).

Table 5-1. The degree, clustering coefficient, closeness centrality, betweenness centrality, and dynamic property of each node in the simple network (Fig. 5-1)

ID	Degree	clustering coefficient	closeness centrality	betweenness centrality	Qii
1	3	0.667	0.421	0.004	0.52
2	3	0.667	0.421	0.004	0.52
3	3	0.667	0.421	0.004	0.52
4	3	0.667	0.421	0.004	0.52
5	3	0.667	0.421	0.004	0.52
6	9	0.222	0.64	0.563	0.183
7	1	0	0.4	0	1.066
8	3	1	0.516	0	0.36
9	6	0.4	0.593	0.4	0.242
10	3	1	0.41	0	0.595
11	3	1	0.41	0	0.595
12	4	0.5	0.421	0.125	0.566
13	1	0	0.302	0	1.448
14	5	0.3	0.571	0.329	0.272
15	2	0	0.39	0.058	0.845
16	2	0	0.39	0.058	0.845
17	2	0	0.296	0.004	1.036

Centrality properties

Here, we introduce two measures of centrality determining the relative importance of a node within a network. The betweenness centrality $C_b(i)$ measures the node centrality in a network by computing the number of the shortest paths from all nodes to all others that pass through the node *i*. $C_b(i)$ is defined as follows:

$$\mathrm{C_b(i)} = \sum_{\mathrm{s\neq i\neq t}} (\sigma_{\mathrm{st}}(\mathrm{i})/\sigma_{\mathrm{st}}) \qquad (2)$$

where *s* and *t* are nodes different from *i*, σ_{st} denotes the number of shortest paths from *s* to *t*, and σ_{st} (*i*) is the number of the shortest paths from *s* to *t* that *i* lies on. The betweenness value of the node *i* is normalized by dividing by the number of node pairs excluding *i*: (*N*-1)(*N*-2)*/2*, where *N* is the total number of nodes in the paths that *i* belongs to.

The closeness centrality $C_c(i)$ of a node *i* is defined as the reciprocal of the average shortest path length and is computed as follows:

$$\mathrm{C_c(i)} = 1/\mathrm{avg}\big(\mathrm{L(i,m)}\big) \qquad (3)$$

where $L(i,m)$ is the length of the shortest path between two nodes *n* and *m*. The closeness centrality of each node is a value between 0 and 1.

The modular similarity between protein pair

The non-diagonal value of *MS-matrix* (MS_{ij}) could provide the relationship between related modularity properties of protein *i* and *j*. For a given protein A, we could identify the overall MS_{Ai} of A and all proteins to evaluate overall modularity relationships. Therefore, we are able to identify the similarity between a protein pair (A and B) based on the overall MS_{Ai} and MS_{Bi}. Here, the similarity is evaluated by the Pearson correlation coefficient (*r*) and

computed as follows:

$$r(A,B) = \frac{\sum_{k=1}^{n}(MS_{Ak}-\overline{MS_A})(MS_{Bk}-\overline{MS_B})}{\sqrt{\sum_{k=1}^{n}(MS_{Ak}-\overline{MS_A})^2}\sqrt{\sum_{k=1}^{n}(MS_{Bk}-\overline{MS_B})^2}} \quad (4)$$

where $\overline{MS_A}$ and $\overline{MS_B}$ are the averages of MS_{Ak} and MS_{Bk}, respectively.

For example, the $r(5,6)$ between nodes 5 and 6 located in the same region (red part in Fig. 5-1A) is 0.88. On the contrary, the *r* between nodes 6 and 9 which are in the different region (red and blue) is -0.53.

The protein-protein interaction network of *S. cerevisiae*

The high-through put data usually have the non-reliable protein-protein interactions. To construct a high-quality protein interaction yeast, we collected protein-protein interaction data from the core subset (named DIPc) of the DIP database [9] which consists of 1,882 proteins and 4,104 protein-protein interactions (the version dated 10 October 2010). Here, the DIPc consists of only the most reliable interactions [121].

Data set of module of *S. cerevisiae*

To evaluate reliability of modules which are identified through the *MS-matrix*, we collected a positive set of yeast module derived MIPS [85]. For 193 modules derived MIPS, we selected 160 modules which have more than a half of proteins in the network constructed by DIPc. According to the definitions of module from the previous studies [84,122,123], a module should have a higher connectivity. Here, the connectivity is defined by previous study [124] and calculated as follow:

$$\text{connectivity} = \frac{\text{No.of PPI within a module}}{k\times(k-1)} \quad (11)$$

where, *k* is the number of protein within a module. Finally, we defined a golden positive dataset which includes 69 MIPS modules, which connectivity is more than 0.6.

5-3. Results

The diagonal value of *MS-matrix* infers essential genes in PPI network of *S. cerevisiae*

Essential genes usually involve in the fundamental cellular processes which required for the survival of an organism [96,97,125]. As a result, the proteins which are products of essential genes should play an important role in the protein-protein interaction network of an organism. To further investigate the relationship between essential genes and important proteins detected by the diagonal values of *MS-matrix* (MS_{ii}), we constructed the yeast protein interaction network by using the high-quality protein-protein interaction data extracting from the core sub set in DIP database (named DIPc). Figure 5-2 displays the progressive ration of essential protein for MS_{ii} from 0 to corresponding value. There are approximately one-half of the proteins recorded as essential proteins while whose MS_{ii} values are less than 0.2; and the proportion of essential protein decreases with the increasing value of MS_{ii}. Furthermore, YBR160W (main cell cycle cyclin-dependent kinase [126]) and YJR045C (Hsp70 family ATPase [127]) are the proteins with lowest value of MS_{ii}, are recorded as essential genes, and play a key role in the important biological processes (e.g. cell cycle and protein folding). These two proteins have enriched interactions and locate on the center of the network. On the contrary, YGL001C and YLR100W, which are related to a non-essential process (ERGosterol biosynthesis), have highest value and only one interaction in the network. These results suggest that those proteins with lower MS_{ii} are located within the steadier regions among the network and more critical for the survival of an organism.

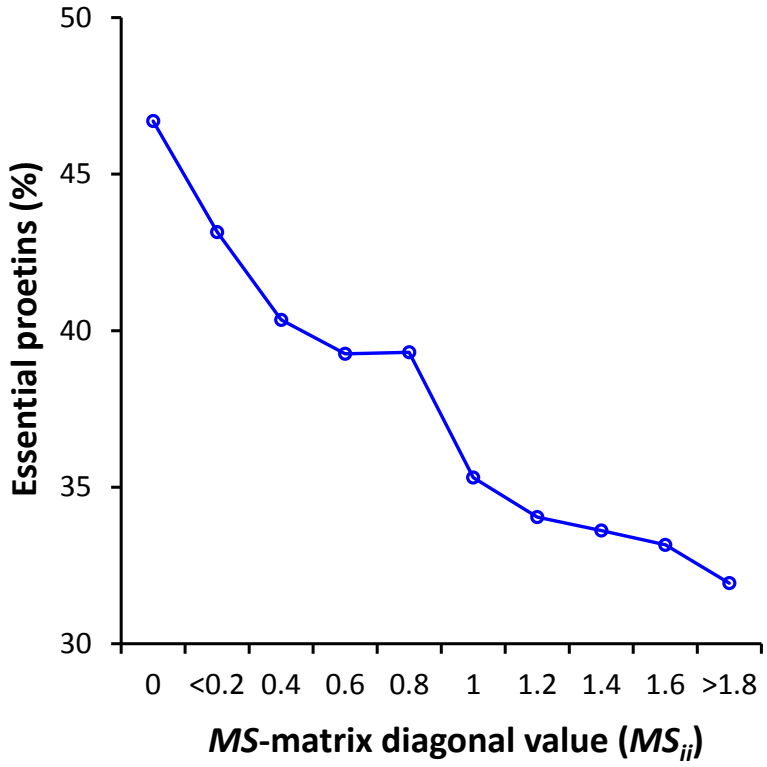


Figure 5-2. The relationship between importance of protein and essential proteins

The importance of protein is calculated by *MS-matrix* diagonal value (MS_{ii}). The interval of MS_{ii} denotes the progressive ratios of essential proteins; the lower MS_{ii} value, more essential proteins are among the network.

The characterization and quantification of network topology derived from the diagonal value of *MS-matrix*

For a given network, there are various types of measurement for determining the relative importance of a node (protein) within a network. For example, degree (degree centrality) is defined as the number of links incident upon a node. According to the degree distribution, *P(k)*, a network could be identified as a scale-free network, which is the architecture of many cellular networks [94]. Closeness centrality is defined as the inverse of the average shortest paths of a given node. The average shortest paths can be regarded as a measure of how fast it will take to spread information from a node to all other nodes sequentially [128]. The betweenness represents the fraction of all of the shortest paths between all nodes in a network that pass through a given node [115].

Our experimental result confirms that the MS_{ii} could represent the essential gene within the yeast PPI network. Next, we evaluated the relationship between MS_{ii} and relative importance (i.e. degree, closeness centrality, and betweenness centrality) of protein within a PPI network (Fig. 5-3). Although the Pearson's correlation coefficient (r) between degree and MS_{ii} is only -0.50, the Spearman correlation (s) is -0.85. This result implicates that the relative importance detected by the older of MS_{ii} is related to the older of relative importance detected by the degree. For example, the protein with the lowest MS_{ii}, YBR160W (main cell cycle cyclin-dependent kinase [126]), is also the node with highest degree (58). Furthermore, the r between closeness centrality and MS_{ii} is -0.78. For example, according to the network described in Figure 5-1, the node 8 is relative important by closeness centrality (0.52; top 4) and could also be identified by using MS_{ii}. In addition, the MS_{ii} is slightly similar (r=-0.3 and s=-0.70) to the betweennes centrality.

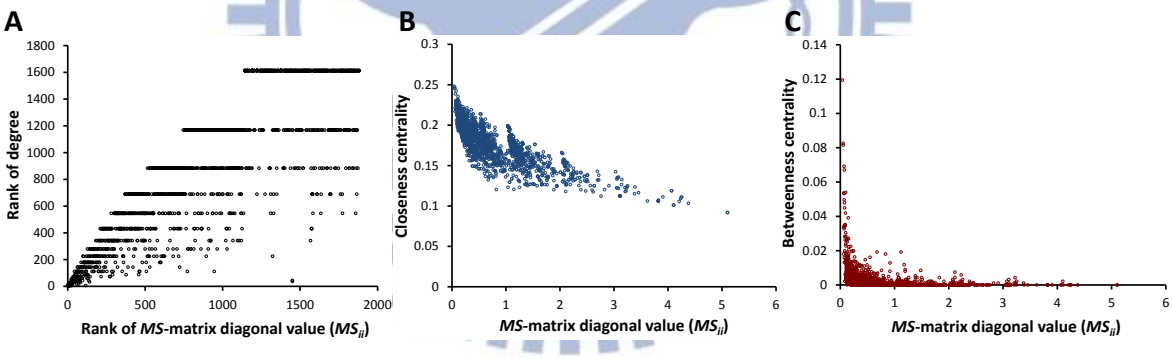


Figure 5-3. Evaluation importance of protein by (A) Degree centrality (B) Closeness centrality (C) Between centrality

(A) The Spearman correlation between degree centrality and MS_{ii} is -0.85. (B) The Pearson correlation between closeness centrality and MS_{ii} is -0.78. (C) The Spearman correlation between betweenness centrality and MS_{ii} is -0.70.

The non-diagonal value of *MS-matrix* reflects the relationship between proteins in yeast PPI network

We have introduced that the MS_{ii} could infer the relative importance of protein i among the whole network. Here, to further investigate the biological meaning of MS_{ij} within a given

network, we utilize the similarity of Gene Ontology[36] and distance of a given protein pair (*i* and *j*) to evaluate the MS_{ij}. The similarity of Gene Ontology is detected by the relative specificity similarity (RSS), proposed by Wu *et al.* [122], to measure the biological process, molecular function, and cellular component similarities.

For a given protein A, we could identify the overall MS_{Ai} of A and all proteins to evaluate overall modularity structure relationships. Therefore, we are able to identify the similarity of overall modularity relationships between protein pair (A and B) based on the MS_{Ai} and MS_{Bi}. Here, the similarity between A and B is evaluated by the Pearson correlation coefficient and derived from the equation (4).

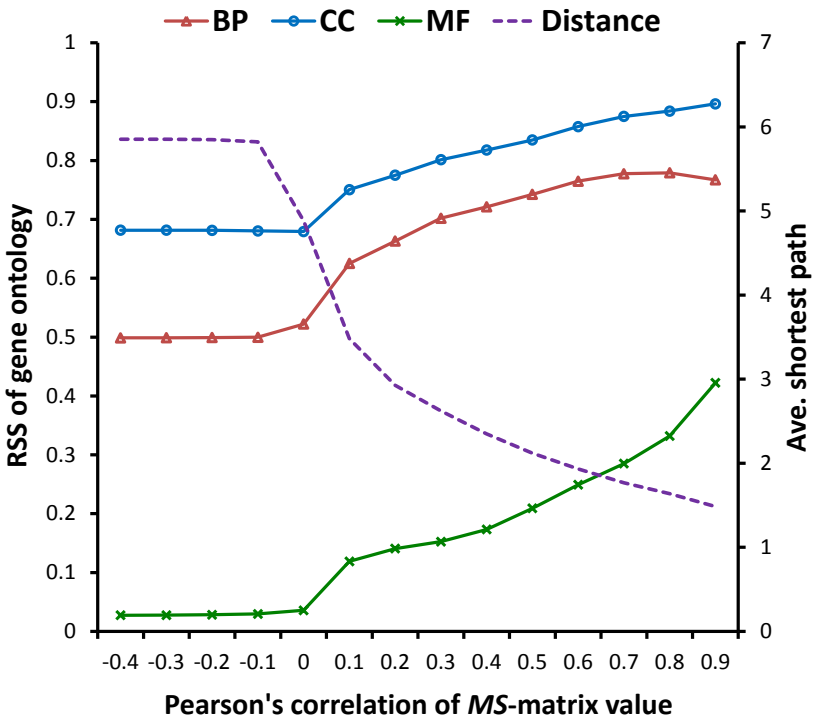


Figure 5-4. The distribution of gene ontology similarities (i.e. RSS of BP, CC, and MF) and the shortest path between protein pairs under different modular similarity

The RSS-BP and RSS-MF have the highest value while modular similarity is more than 0.9; moreover, the average distance is lower than 2. The RSS-CC are higher than 0.7 while modular similarities are higher than 0.4.

Figure 5-4 illustrates the distribution of gene ontology similarities and the shortest path between protein pairs. While the protein pairs have ≥0.1 modular similarity, the average of

their distance has an obvious decrease (from 4.88 to 3.48) and all of the average RSS have an obvious increase. In addition, the average of protein pair's distance would be less than 2.5 and share the higher biological process and cellular component annotation (RSS-BP > 0.7 and RSS-CC > 0.8), while these protein pairs have more than 0.4 modular similarity. The RSS-BP and RSS-MF have the highest value while modular similarity is more than 0.9; moreover, the average distance is lower than 2. This result implies that a protein pair with a highly modular similarity would share a significant similarity of Gene Ontology, especially BP and CC, and are neighboring proteins (e.g. an interaction protein pair) in the PPI network.

Identification of modules based on the non-diagonal value of *MS-matrix*

According to the definitions of module from the previous studies [84,122,123], the proteins of a module should locate on the same component, join a same biological process, carry out similar or related function, and have relatively autonomous of the whole network. We have introduced that the modular similarity of protein pair (A and B) derived from the Pearson correlation coefficient of MS_{Ai} and MS_{Bi}, could infer the similarity of Gene Ontology and the relationship between A and B within the PPI network. Therefore, we believe that the MS-matrix could be useful for identifying modules of a give PPI network. Here, we utilize the hierarchical clustering method to identify the modules and the distance between protein pair (A and B) is calculated by using the modular similarity (i.e. Pearson correlation coefficient of MS_{Ai} and MS_{Bi}). Then, we identified 126 modules including 724 proteins derived from the *MS-matrix*. To further investigate the reliability of modules, we compare our modules with the modules recorded in MIPS and analysis the Gene Ontology and connectivity of our modules.

For 193 modules derived MIPS, we selected 160 modules which have more than a half of proteins in the network constructed by DIPc. According to the definitions of module from the

previous studies [84,122,123], a module should have a higher connectivity. Finally, we defined a golden positive dataset which includes 69 MIPS modules, which connectivity is more than 0.6. The overlap between a reference MIPS module *R* and a predicted module *M* can be quantified by Jaccard index [129]. The Jaccard index is calculated as follow:

$$\text{Jaccard index} = \frac{|R \cap M|}{|R \cup M|} \quad (12)$$

where, the $|R \cap M|$ is the number of protein which is the intersection of R and M; the $|R \cup M|$ is the number of protein which is the union of R and M.

For each reference module, we find the prediction that has the highest Jaccard index. Total 47 modules are related to our modules (Jaccard index > 0). If a module with Jaccard index ≥ 0.5 is considered as a hit module, our method has 36 (52%) hits of golden positive dataset.

Next, because modules have relatively autonomous of the whole network, the connectivity of modules should be higher than the proteins which include the module and the proteins connecting to the module (named "extent 1 layer"). Table 5-2 shows the connectivity of our module, 160 MIPS module, and golden positive dataset. Because the 160 MIPS modules are only filtered by number of protein within PPI network, these 160 MIPS have a lower connectivity. In addition, both of our modules and the golden positive dataset have a higher average connectivity (i.e. 0.73 and 0.84, respectively). The average connectivity of all set would have an obvious decreasing from modules to the extent 1 layer. In addition, all modules derived from *MS-matrix* and golden positive dataset have a higher connectivity than the extent 1 layer (Table 5-2).

Table 5-2. Connectivity of module and proteins which include the module and the proteins connecting to the module (named "extent 1 layer")

Module Set	No. of Module	Average connectivity	Average connectivity of extent 1 layer	No. of module which connectivity > connectivity of extent 1 layer
Our	126	0.73	0.32	126
MIPS	160	0.49	0.18	150

Golden positive	69	0.84	0.28	69

Furthermore, we annotated modules by utilizing the consensus GO terms within a given module. To annotate a module with *Y* proteins, we define a consensus ratio (*CRM*) of GO term *i* as $CRM=Y_i/Y$, where Y_i is the number of proteins with GO term *i* in a module. Next, the enrichment for each module in each GO term was determined by the *p*-value of the hypergeometric distribution and then this *p*-value was adjusted based on Bonferroni correction [130,131]. Here, a GO term is considered as a representative GO term of a module if $CRM > 0.6$ and adjusted *p*-value of GO term ≤ 0.05 [130,131] based on statistically analysis. Figure 5-5 illustrates the distribution of the number of representative GO term within a given module derived from *MS-matrix* and MIPS. Then, we applied the two-tailed *T*-test to further investigate the difference between *MS-matrix* and MIPS. However, all of the *P*-values (0.18, 0.30, and 0.13) imply that the number of representative GO term within a given module do not have a significant different between *MS-matrix* and MIPS. In addition, we also investigate the representative GO terms which have the top 5 ratio in our modules or MIPS modules. The Jaccard index of BP, CC, and MF are 0.67, 0.67, and 1, respectively. This result implies that the biological characterization (i.e. No. of representative GO terms in a module and top 5 terms) of our module derived from the *MS-matrix* is similar to the MIPS modules which are identified by the experiments.

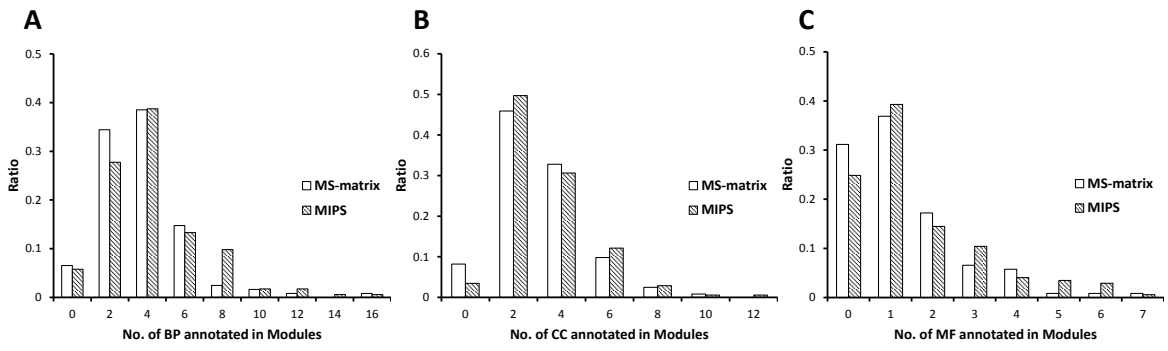


Figure 5-5. The distribution of the number of gene ontology annotations (i.e. (A)BP, (B)CC, and C(MF) within a given module derived from *MS-matrix* and MIPS

Based on the two-tailed *T*-test between *MS-matrix* and MIPS, all of gene ontology annotations (i.e. BP, CC and BF) do not have significant different (i.e. *P*-values are 0.18, 0.30, and 0.13 respectively).

Example of modules derived from the *MS-matrix*

According to 126 modules including 724 proteins derived from the *MS-matrix*, Figures 6A and 6B illustrate the 9 modules, which sizes are greater than 10, on the network and their density region on the *MS-matrix*. Two modules with the lowest average MS_{ii} values (0.1 and 0.16) are the 19S proteasome and U4/U6 x U5 tri-snRNP complex (purple and light blue regions in Fig. 5A). The proteasome is a protease that controls diverse processes in eukaryotic cells; and snRNPs are large RNA-protein molecular complexes upon which splicing of pre-mRNA occur. Both of two modules are play essential roles in a yeast PPI network. In addition, two largest modules (19 and 17 proteins) are the F1-F0 ATP synthase and peroxisomes. Then, we use two modules (i.e. anaphase-promoting complex/cyclosome (APC/C) and peroxisomes) as examples to further introduce the module identification derived from the *MS-matrix*.

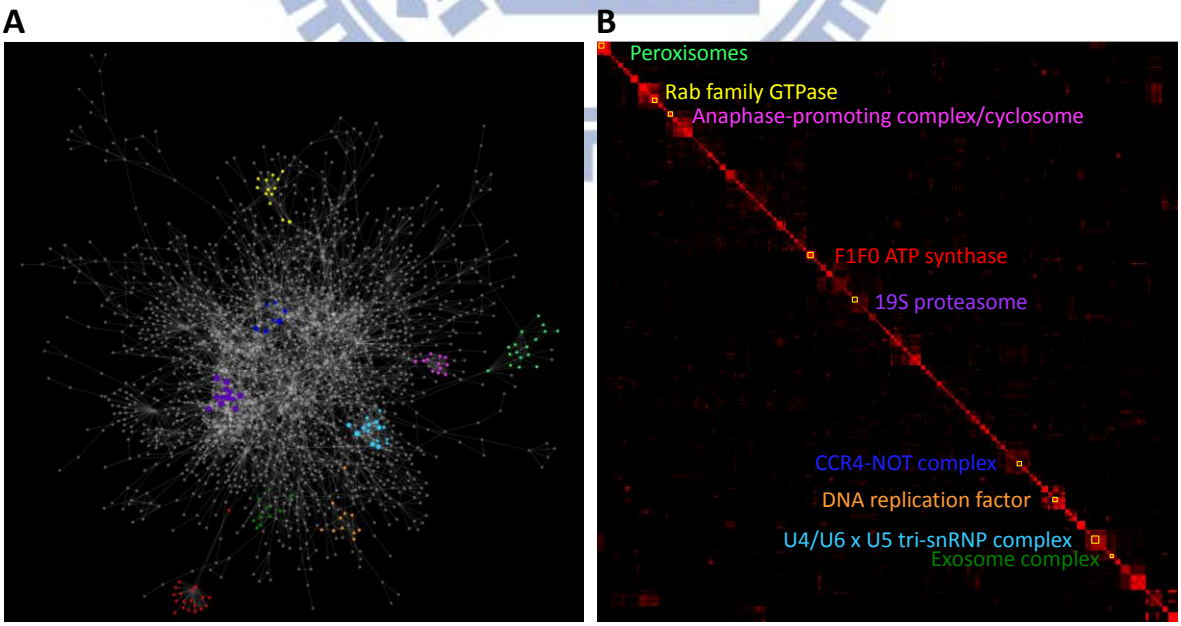


Figure 5-6. The modules derived from the *MS-matrix*

(A) Yeast protein interaction network with 9 colored modules (e.g. F1-F0 ATP synthase (red), 19S proteasome (purple), anaphase-promoting complex/cyclosome (pink), and peroxisome (light green)). (B) The MS-matrix of

PPI network with the 9 modules which map to the 9 colored regions on the network.

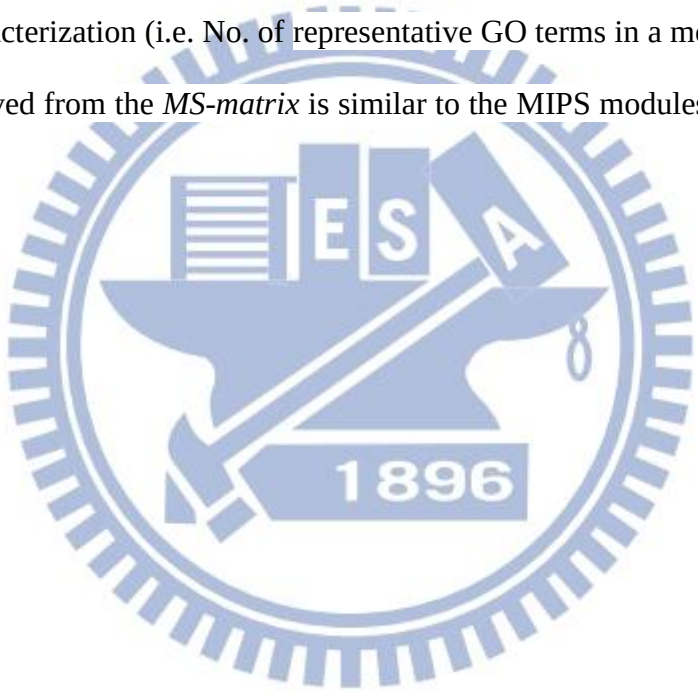
The anaphase-promoting complex/cyclosome (APC/C) mediates cell cycle–regulated ubiquitination, and thereby degradation, of proteins containing sequences called destruction boxes [132]. There are 11 proteins are defined as anaphase-promoting complex derived from the MIPS database. In addition, 6 of these proteins are also the products of essential genes derived from the DEG database. According to *MS-matrix*, we identified a region which is local density region (pink region in Fig. 5-6A) and has 12 proteins sharing the similar behaviours of non-diagonal values (pink in Fig. 5-6B). The protein not recorded in MIPS is YGL003C, a cell-cycle regulated activator of the APC/C [133,134]. Although YGL003C is not a member of APC, YGL003C is highly related to the APC and share the same Gene Ontology annotation (i.e. anaphase-promoting complex) with other APC/C proteins based on the Saccharomyces Genome Database (SGD) [135]. This result indicates that the 12 APC/C related proteins derived from the *MS-matrix* could be considered as a reasonable module.

In addition, we also identified novel modules which are not recorded in the MIPS. Peroxisomal proteins are synthesized on free polyribosomes and imported posttranslationally. The biogenesis of peroxisomes requires a group of protein factors referred to as peroxins which are encoded by the PEX genes [136]. According to *MS-matrix*, we identified a region which is local density region (light green region in Fig. 5-6A) and has 17 proteins sharing the similar behaviours of non-diagonal values (light green in Fig. 5-6B). There are 14 proteins are recorded as PEX genes in SGD. Two proteins (i.e. YML042W and YIL160C) are also involved in the same cellular component (i.e. peroxisome) based on SGD. Therefore, this module derived from the *MS-matrix* may be a reasonable module.

5-4. Conclusions

For a given PPI network of *S. cerevisiae*, our results demonstrate that the important

proteins identified by the *MS-matrix* are related to the essential biological processes (i.e. essential genes). In addition, the important proteins derived from *MS-matrix* are highly consistence with the topology features (i.e. degree, closeness centrality, and betweenness centrality). Then, the relationship between proteins derived from the *MS-matrix* could reflect the similarity of Gene Ontology and could be useful for the module identification. For 69 reference modules of golden positive dataset, there are 47 modules are related to our modules (Jaccard index > 0). If a module with Jaccard index ≥ 0.5 is considered as a hit module, our method has 36 (52%) hits of golden positive dataset. Furthermore, our results also imply that the biological characterization (i.e. No. of representative GO terms in a module and top 5 terms) of our module derived from the *MS-matrix* is similar to the MIPS modules which are identified by the experiments.



Chapter 6. Conclusion

6-1. Summary

In this thesis, we presented the "3D-domain interologs mapping" and "protein complex family" to construct the structure resolved PPI networks across multiple organisms. "3d-domain interolos mapping" is a concept for efficiently enlarging protein interactions annotated through the homologous PPIs with residue-based binding models. We verified the structure resolved PPI networks on Gene Ontology annotations[36] and the architecture of topology (i.e. scale-free network properties). In addition, we also provide the consensus proteins across three networks based on "3D-domain interologs mapping". These consensus proteins are highly related to the essential genes and disease related proteins. We believe that structure resolved PPI networks would provide the insight for understanding the mechanism of biological processes within a given PPI network. In summary, the major contributions of this study are listed as the following:

1. We proposed several new concepts, including "3D-domain interologs mapping" and "protein complex family", to study the evolution of PPIs and protein complexes across multiple species. A group of PPIs are regarded as a PPI family when they meet the following criteria: (1) The proteins of the PPIs are homologous proteins, respectively; (2) The interactions of PPIs share the similar binding model based on the structure templates. In addition, a group of protein complexes are regarded as a protein complex family when they meet the two criteria and an additional criterion: the protein complexes share the similar complex similarity. More importantly, these two concepts provide a new way to efficiently enlarge the PPIs and protein complexes annotated with residue-based binding models.

2. We developed a database, namely 3D-interologs, records the evolution of protein-protein interactions database across multiple species derived from “3D-domain interolog mapping” and a template-based scoring function. We have inferred 173,294 homologous protein-protein interactions by using 1,895 three-dimensional (3D) structure heterodimers to search the UniProt database (4,826,134 protein sequences). For a protein-protein interaction, the 3D-interologs database shows interacting domains and binding models derived from structure template. More importantly, this database provides the evolution of PPI by exploring its PPI family across multiple species.

3. We developed a web server, namely PCFamily, for identifying homologous complexes and inferring conserved domains and GO terms from protein complex families. PCFamily is the first server to provide homologous complexes in multiple species; graphic visualization of the complex topology and detailed atomic residue-residue interactions; interface alignments; conservations of GO terms and domain compositions. We believe that the server is able to provide valuable insights for determining functional modules of biological networks across multiple species.

4. Based on the two concepts, we were able to construct the structure resolved PPI networks in *H. sapiens*, *M. musculus*, and *D. rerio*. In each structure resolved network, the PPIs with atomic residue-based binding models in the derived structure resolved network achieved highly agreement with Gene Ontology similarities. In addition, our derived networks can be used to observe the consensus proteins and modules derived from the multiple network alignment of *H. sapiens*, *M. musculus*, and *D. rerio*. These consensus proteins are often the essential genes and play key roles in the architecture of these networks. More importantly, our results demonstrate that the structure resolved PPI networks would provide valuable insights into understanding the mechanism of biological processes (e.g. cancer, cardiovascular-related diseases, and complement and coagulation

pathway) across multiple organisms.

6-2. Discussion and future work

According to the characteristics of "3D-domain interologs mapping" and "protein complex family", the interactome behavior, we discussed here, is focused on the conserved proteins and PPIs which are the members of the PPI and protein complex families. In this thesis, we used our concept to studying the evolution of these PPIs and protein complexes. Therefore, we only discuss the conservation and difference of these consensus pathways across multiple organisms. However, the organism-specific proteins and PPIs usually play an important role during the organism evolution. This issue should be the next important issue for our studies.

Our structural networks can annotate and infer the cell behaviors of a new determined (or seldom-studied) species (e.g. zebrafish), by mapping some well-studied species. However, the construction of our structurally resolved PPI networks largely relies on the availability of 3D-crystal structures, which limits the coverage of our network. But, we believe that the rapid growth of PDB providing more 3D-crystal information and our methods can be readily applied to uncover potential molecular mechanisms whose structural information is currently missing.

In addition, our methods should also be considering these high-quality experimental PPIs with possible domain annotations. Prof. Yang Lab has already provided the sequence-based PPI family for annotating and studying PPIs across multiple organisms with non-structure information. Although the accuracy of method for PPI annotation is less than "3D-domain interologs mapping", the sequence-based PPI family has more coverage to explore the non-well-known organisms. In the future work, we could carefully utilize sequence-based PPI family with high-quality experimental PPIs to enlarging the coverage of PPIs and provide a

more complete PPI network for understanding the mechanism of cell behaviors.

However, dynamic architecture of the protein interaction network has an important role in the regulation of cell behavior. Understanding the functional organization of protein interaction networks is the most important issue for understanding the principles of cellular behavior. More importantly, it also provides a way for understanding the diseases where cellular behavior is miss-regulated. Currently, most of these studies have considered the protein interaction networks without taking into account the dynamic nature of protein expression, which is essential for a proper representation of biological networks.

In current state, we have already been able to construct the structure resolved PPI networks in multiple organisms. We also provide the consensus proteins and PPIs in these networks. According to our results, the structure resolved PPI networks derived from the PPI family would provide the insight for understanding the mechanism of biological processes within a given PPI network. To further investigate the behavior of PPI network within a given cell, gene expression data would provide an aspect of in-depth understanding of the dynamic organization of the PPI network and its role in the regulation of cellular processes. For example, the Connectivity Map (also known as cmap) provided by Lamb, J. *et al.* is a collection of genome-wide transcriptional expression data from cultured human cells treated with bioactive small molecules and simple pattern-matching algorithms that together enable the discovery of functional connections between drugs, genes and diseases through the transitory feature of common gene-expression changes [37].

Therefore, we will combine the gene expression data into the PPI network. We will try to illustrate the behavior of PPI networks under different cell types and different conditions. Because the Connectivity Map could provide the up-regulated and down-regulated proteins of given drugs and diseases, combining these data with our structure resolved PPI networks should be able to explain the mechanism of relationship between the drugs, genes and diseases.

List of publications

Journal paper

1. C.-Y Lin, Y.-W. Lin, S.-W. Yu, **Y.-S. Lo**, and J.-M. Yang*, "MoNetFamily: a web server to infer homologous modules and module-module interaction networks in vertebrates," *Nucleic Acids Research*, W263-W270, 2012.

2. S.-C. Hsu, C.-P. Chang, C.-Y. Tsai, S.-H. Hsieh, B.A. Wu-Hsieh, **Y.-S. Lo**, and J.-M. Yang, "Steric recognition of TCR contact residues is required to map mutant epitopes by immunoinformatical programmes," *Immunology*, 139-152, 2012.

3. I-H. Liu, **Y.-S. Lo**, and J.-M. Yang*, "Template-based Scoring Functions for Visualizing Biological Insights of H-2Kb–peptide–TCR Complexes," *International Journal of Data Mining and Bioinformatics*, 2012, in press.

4. I-H. Liu, **Y.-S. Lo**, and J.-M. Yang*, "PAComplex: a web server to infer peptide antigen families and binding models from TCR–pMHC complexes," *Nucleic Acids Research*, W254-W260, 2011.

5. **Y.-S. Lo**, C.-Y Lin, and J.-M. Yang*, "PCFamily: a web server for searching homologous protein complexes," *Nucleic Acids Research*, W516-W522, 2010.

6. **Y.-S. Lo**, Y.-C. Chen, and J.-M. Yang*, "3D-interologs: An evolution database of physical protein-protein interactions across multiple genomes," *BMC Genomics*, (Suppl 3):S7, 2010.

7. C.-C. Chen, C.-Y. Lin, **Y.-S. Lo**, and J.-M. Yang*, "PPISearch: a web server for searching homologous protein-protein interactions across multiple species," *Nucleic Acids Research*, W369-W375, 2009.

8. Y.-C. Chen, **Y.-S. Lo**, W.-C. Hsu, and J.-M. Yang*, "3D-partner: a web server to infer interacting partners and binding models, " *Nucleic Acids Research*, W561-W567, 2007.

REFERENCES

1 Stelzl, U. *et al.* A human protein-protein interaction network: A resource for annotating the proteome. *Cell* **122**, 957-968 (2005).

2 Tarassov, K. *et al.* An in vivo map of the yeast protein interactome. *Science* **320**, 1465-1470 (2008).

3 Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104-110 (2008).

4 Ihmels, J. *et al.* Revealing modular organization in the yeast transcriptional network. *Nat Genet* **31**, 370-377 (2002).

5 Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabasi, A. L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551-1555 (2002).

6 Han, J. D. J. *et al.* Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88-93 (2004).

7 Gavin, A. C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631-636 (2006).

8 Kerrien, S. *et al.* IntAct - open source resource for molecular interaction data. *Nucleic Acids Res* **35**, D561-D565 (2007).

9 Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* **32**, D449-451 (2004).

10 Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535-D539 (2006).

11 Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM (R)). *Nucleic Acids Res* **37**, D793-D796 (2009).

12 Wu, C. H. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* **34**, D187-D191 (2006).

13 Goh, K. I. *et al.* The human disease network. *P Natl Acad Sci USA* **104**, 8685-8690, (2007).

14 Schuster-Bockler, B. & Bateman, A. Protein interactions in human genetic diseases. *Genome Biol* **9**, (2008).

15 Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Mol Syst Biol* **5**, (2009).

16 Wang, X. J. *et al.* Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature Biotechnology* **30**, 159-164 (2012).

17 Evidence for network evolution in an Arabidopsis interactome map. *Science* **333**, 601-607 (2011).

18 Zhao, Y. Q. & Mooney, S. D. Functional organization and its implication in evolution of the human protein-protein interaction network. *Bmc Genomics* **13**, (2012).

19 Peterson, G. J., Presse, S., Peterson, K. S. & Dill, K. A. Simulated evolution of protein-protein interaction networks with realistic topology. *Plos One* **7**, e39052 (2012).

20 Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res* **36**, D281-D288 (2008).

21 Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res* **37**, D211-D215 (2009).

22 Andreeva, A. *et al.* SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* **32**, D226-D229 (2004).

23 Chen, C. C., Lin, C. Y., Lo, Y. S. & Yang, J. M. PPISearch: a web server for searching homologous protein-protein interactions across multiple species. *Nucleic Acids Res* **37**, W369-W375 (2009).

24 Lo, Y. S., Lin, C. Y. & Yang, J. M. PCFamily: a web server for searching homologous protein complexes. *Nucleic Acids Res* **38**, W516-W522 (2010).

25 Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-627 (2000).

26 Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *P Natl Acad Sci USA* **98**, 4569-4574 (2001).

27 Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**, 637-643 (2006).

28 Pagel, P. *et al.* The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**, 832-834 (2005).

29 Chatr-Aryamontri, A. *et al.* MINT: the molecular INTeraction database. *Nucleic Acids Res* **35**, D572-D574 (2007).

30 Kersey, P. *et al.* Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res* **33**, D297-D302 (2005).

31 Deshpande, N. *et al.* The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res* **33**, D233-D237 (2005).

32 Aloy, P. & Russell, R. B. Interrogating protein interaction networks through structural biology. *P Natl Acad Sci USA* **99**, 5896-5901 (2002).

33 Lu, L., Lu, H. & Skolnick, J. MULTIPROSPECTOR: An algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins-Structure Function and Genetics* **49**, 350-364 (2002).

34 Stein, A., Russell, R. B. & Aloy, P. 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res* **33**, D413-D417 (2005).

35 Finn, R. D., Marshall, M. & Bateman, A. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* **21**, 410-412 (2005).

36 Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource.

Nucleic Acids Res **32**, D258-D261 (2004).

37 Lamb, J. *et al.* The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929-1935 (2006).

38 Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Science of the USA* **96**, 4285-4288 (1999).

39 Chen, Y. C., Lo, Y. S., Hsu, W. C. & Yang, J. M. 3D-partner: a web server to infer interacting partners and binding models. *Nucleic Acids Res* **35**, W561-W567 (2007).

40 Yu, H. Y. *et al.* Annotation transfer between genomes: Protein-protein interologs and protein-DNA regulogs. *Genome Res* **14**, 1107-1118 (2004).

41 Shoemaker, B. A. & Panchenko, A. R. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput Biol.* **3**, 337-344 (2007).

42 Patil, A. & Nakamura, H. Filtering high-throughput protein-protein interaction data using a combination of genomic features. *Bmc Bioinformatics* **6**, 100-112 (2005).

43 Saeed, R. & Deane, C. An assessment of the uses of homologous interactions. *Bioinformatics* **24**, 689-695 (2008).

44 Matthews, L. R. *et al.* Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res* **11**, 2120-2126 (2001).

45 Aloy, P. *et al.* Structure-based assembly of protein complexes in yeast. *Science* **303**, 2026-2029 (2004).

46 Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405**, 837-846 (2000).

47 von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399-403 (2002).

48 Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449-453 (2003).

49 Wojcik, J. & Schachter, V. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics* **17**, S296-S305 (2001).

50 Cary, M. P., Bader, G. D. & Sander, C. Pathway information for systems biology. *FEBS Letters* **579**, 1815-1820 (2005).

51 Sun, J. *et al.* InPrePPI: an integrated evaluation method based on genomic context for predicting protein-protein interactions in prokaryotic genomes. *Bmc Bioinformatics* **8**, 414 (2008).

52 von Mering, C. *et al.* STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res* **33**, D433-D437 (2005).

53 Kelley, B. P. *et al.* Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *P Natl Acad Sci USA* **100**, 11394-11399 (2003).

54 Sharan, R. *et al.* Conserved patterns of protein interaction in multiple species. *P Natl Acad Sci USA* **102**, 1974-1979 (2005).

55 Ofran, Y. & Rost, B. Analysing six types of protein-protein interfaces. *Journal of Molecular Biology* **325**, 377-387 (2003).

56 Li, L., Stoeckert, C. J. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189 (2003).

57 Thorn, K. S. & Bogan, A. A. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* **17**, 284-285 (2001).

58 Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *P Natl Acad Sci USA* **89**, 10915-10919 (1992).

59 Glaser, F., Steinberg, D. M., Vakser, I. A. & Ben-Tal, N. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins: Structure, Function, and Bioinformatics* **43**, 89-102 (2001).

60 Black, B. E., Levesque, L., Holaska, J. M., Wood, T. C. & Paschal, B. M. Identification of an NTF2-related factor that binds Ran-GTP and regulates nuclear protein export. *Molecular and Cellular Biology* **19**, 8616-8624 (1999).

61 Rual, J. F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178 (2005).

62 Herold, A. *et al.* TAP (NXF1) belongs to a multigene family of putative RNA export factors with a conserved modular architecture. *Molecular and Cellular Biology* **20**, 8996-9008 (2000).

63 Fribourg, S., Braun, I. C., Izaurralde, E. & Conti, E. Structural basis for the recognition of a nucleoporin FG repeat by the NTF2-like domain of the TAP/p15 mRNA nuclear export factor. *Molecular cell* **8**, 645-656 (2001).

64 Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **28**, 10-14 (2000).

65 Bogan, A. A. & Thorn, K. S. Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology* **280**, 1-9 (1998).

66 Dickinson, C. D., Kelly, C. R. & Ruf, W. Identification of surface residues mediating tissue factor binding and catalytic function of the serine protease factor VIIa. *Proceedings of the National Academy of Science of the USA* **93**, 14379-14384 (1996).

67 Ruf, W. *et al.* Energetic contributions and topographical organization of ligand binding residues of tissue factor. *Biochemistry* **34**, 6310-6315 (1995).

68 Rhodes, D. R. *et al.* Probabilistic model of the human protein-protein interaction network. *Nature Biotechnology* **23**, 951-959 (2005).

69 Wu, X. M., Zhu, L., Guo, J., Zhang, D. Y. & Lin, K. Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Res* **34**, 2137-2150 (2006).

70 Edwards, A. M. *et al.* Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends in genetics : TIG* **18**, 529-536 (2002).

71 Kemmeren, P. *et al.* Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Molecular cell* **9**, 1133-1143 (2002).

72 Aloy, P. & Russell, R. B. Ten thousand interactions for the molecular biologist. *Nat Biotechnol* **22**, 1317-1321 (2004).

73 Levy, E. D., Pereira-Leal, J. B., Chothia, C. & Teichmann, S. A. 3D complex: a structural classification of protein complexes. *PLoS computational biology* **2**, e155 (2006).

74 Davis, F. P. *et al.* Protein complex compositions predicted by structural similarity. *Nucleic Acids Res* **34**, 2943-2952 (2006).

75 Herraez, A. Biomolecules in the computer - Jmol to the rescue. *Biochemistry and Molecular Biology Education* **34**, 255-261 (2006).

76 Hao, B. *et al.* Structural basis of the Cks1-dependent recognition of p27(Kip1) by the SCF(Skp2) ubiquitin ligase. *Molecular cell* **20**, 9-19 (2005).

77 Carrano, A. C., Eytan, E., Hershko, A. & Pagano, M. SKP2 is required for ubiquitin-mediated degradation of the CDK inhibitor p27. *Nature cell biology* **1**, 193-199 (1999).

78 Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**, D480-D484 (2008).

79 Masuda, S. *et al.* Functional erythropoietin receptor of the cells with neural characteristics. Comparison with receptor properties of erythroid cells. *J Biol Chem* **268**, 11208-11216 (1993).

80 Syed, R. S. *et al.* Efficiency of signalling through cytokine receptors depends critically on receptor orientation. *Nature* **395**, 511-516 (1998).

81 Damen, J. E. & Krystal, G. Early events in erythropoietin-induced signaling. *Experimental hematology* **24**, 1455-1459 (1996).

82 Yamaji, R. *et al.* The intron 5-inserted form of rat erythropoietin receptor is expressed as a membrane-bound form. *Biochimica et biophysica acta* **1403**, 169-178 (1998).

83 Buescher, J. M. *et al.* Global network reorganization during dynamic adaptations of Bacillus subtilis metabolism. *Science* **335**, 1099-1103 (2012).

84 Wagner, G. P., Pavlicev, M. & Cheverud, J. M. The road to modularity. *Nat Rev Genet* **8**, 921-931 (2007).

85 Mewes, H. W. *et al.* MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Research* **36**, D196-D201 (2008).

86 Licata, L. *et al.* MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research* **40**, D857-D861 (2012).

87 Lo, Y. S., Chen, Y. C. & Yang, J. M. 3D-interologs: an evolution database of physical protein- protein interactions across multiple genomes. *Bmc Genomics* **11**, (2010).

88 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).

89 Pruess, M., Kersey, P. & Apweiler, R. INTEGR8, a resource for proteomic and genomic data. *Molecular & Cellular Proteomics* **4**, S29-S29 (2005).

90 Ciriello, G., Mina, M., Guzzi, P. H., Cannataro, M. & Guerra, C. AlignNemo: A Local Network Alignment Method to Integrate Homology and Topology. *Plos One* **7**, e38107 (2012).

91 Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res* **40**, D136-143 (2012).

92 Smigielski, E. M., Sirotkin, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* **28**, 352-355 (2000).

93 Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25-29 (2000).

94 Barabasi, A. L. & Oltvai, Z. N. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics* **5**, 101-U115 (2004).

95 Li, S. S., Xu, K. & Wilkins, M. R. Visualization and Analysis of the Complexome Network of Saccharomyces cerevisiae. *J Proteome Res* **10**, 4744-4756 (2011).

96 Dotsch, A. *et al.* Evolutionary conservation of essential and highly expressed genes in Pseudomonas aeruginosa. *Bmc Genomics* **11**, (2010).

97 Zhang, R. & Lin, Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res* **37**, D455-458 (2009).

98 Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *Bmc Bioinformatics* **4,** (2003).

99 Haugsten, E. M., Wiedlocha, A., Olsnes, S. & Wesche, J. Roles of Fibroblast Growth Factor Receptors in Carcinogenesis. *Mol Cancer Res* **8**, 1439-1452 (2010).

100 Dhillon, A. S., Hagan, S., Rath, O. & Kolch, W. MAP kinase signalling pathways in cancer. *Oncogene* **26**, 3279-3290 (2007).

101 Zhang, X. Q. *et al.* Receptor specificity of the fibroblast growth factor family - The complete mammalian FGF family. *J Biol Chem* **281**, 15694-15700 (2006).

102 Ornitz, D. M. *et al.* Receptor specificity of the fibroblast growth factor family. *J Biol Chem* **271**, 15292-15297 (1996).

103 Jang, J. H., Shin, K. H. & Park, J. G. Mutations in fibroblast growth factor receptor 2 and fibroblast growth factor receptor 3 genes associated with human gastric and colorectal cancers. *Cancer Res* **61**, 3541-3543 (2001).

104 Lajeunie, E. *et al.* FGFR2 mutations in Pfeiffer syndrome. *Nat Genet* **9**, 108 (1995).

105 Pollock, P. M. *et al.* Frequent activating FGFR2 mutations in endometrial carcinomas parallel germline mutations associated with craniosynostosis and skeletal dysplasia syndromes. *Oncogene* **26**, 7158-7162 (2007).

106 Ibrahimi, O. A. *et al.* Biochemical analysis of pathogenic ligand-dependent FGFR2 mutations suggests distinct pathophysiological mechanisms for craniofacial and limb

abnormalities. *Hum Mol Genet* **13**, 2313-2324 (2004).

107 Plotnikov, A. N., Hubbard, S. R., Schlessinger, J. & Mohammadi, M. Crystal structures of two FGF-FGFR complexes reveal the determinants of ligand-receptor specificity. *Cell* **101**, 413-424 (2000).

108 Landstrom, A. P. *et al.* Molecular and functional characterization of novel hypertrophic cardiomyopathy susceptibility mutations in TNNC1-encoded troponin C. *J Mol Cell Cardiol* **45**, 281-288 (2008).

109 Laing, N. G. *et al.* Myosin storage myopathy: Slow skeletal myosin (MYH7) mutation in two isolated cases. *Neurology* **64**, 527-529 (2005).

110 Arad, M. *et al.* Gene mutations in apical hypertrophic cardiomyopathy. *Circulation* **112**, 2805-2811 (2005).

111 Takahashi-Yanaga, F. *et al.* Functional consequences of the mutations in human cardiac troponin I gene found in familial hypertrophic cardiomyopathy. *Journal of Molecular and Cellular Cardiology* **33**, 2095-2107 (2001).

112 Bager, R. *et al.* Urokinase-type Plasminogen Activator-like Proteases in Teleosts Lack Genuine Receptor-binding Epidermal Growth Factor-like Domains. *J Biol Chem* **287**, 27526-27536 (2012).

113 Poort, S. R., Michiels, J. J., Reitsma, P. H. & Bertina, R. M. Homozygosity for a novel missense mutation in the prothrombin gene causing a severe bleeding disorder. *Thromb Haemost* **72**, 819-824 (1994).

114 Suzuki, K., Nishioka, J., Kusumoto, H. & Hashimoto, S. Mechanism of Inhibition of Activated Protein-C by Protein-C Inhibitor. *J Biochem-Tokyo* **95**, 187-195 (1984).

115 Seebacher, J. & Gavin, A. C. SnapShot: Protein-protein interaction networks. *Cell* **144**, 1000, 1000 e1001 (2011).

116 Bader, G. D. & Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2 (2003).

117 Adamcsek, B., Palla, G., Farkas, I. J., Derenyi, I. & Vicsek, T. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* **22**, 1021-1023 (2006).

118 Fouss, F., Pirotte, A., Renders, J. M. & Saerens, M. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *Ieee T Knowl Data En* **19**, 355-369 (2007).

119 Rader, A. J. *et al.* Identification of core amino acids stabilizing rhodopsin. *Proc Natl Acad Sci U S A* **101**, 7246-7251 (2004).

120 Yang, L. W. & Bahar, I. Coupling between catalytic site and collective dynamics: A requirement for mechanochemical activity of enzymes. *Structure* **13**, 893-904 (2005).

121 Deane, C. M., Salwinski, L., Xenarios, I. & Eisenberg, D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics* **1**, 349-356 (2002).

122 Wu, X., Zhu, L., Guo, J., Zhang, D. Y. & Lin, K. Prediction of yeast protein-protein

interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Res* **34**, 2137-2150 (2006).

123 Espinosa-Soto, C. & Wagner, A. Specialization Can Drive the Evolution of Modularity. *Plos Comput Biol* **6** (2010).

124 Campillos, M., von Mering, C., Jensen, L. J. & Bork, P. Identification and analysis of evolutionarily cohesive functional modules in protein networks. *Genome Res* **16**, 374-382 (2006).

125 Kobayashi, K. *et al.* Essential Bacillus subtilis genes. *Proc Natl Acad Sci U S A* **100**, 4678-4683 (2003).

126 Enserink, J. M. & Kolodner, R. D. An overview of Cdk1-controlled targets and processes. *Cell Div* **5**, 11 (2010).

127 Liu, Q., Krzewska, J., Liberek, K. & Craig, E. A. Mitochondrial Hsp70 Ssc1: role in protein folding. *J Biol Chem* **276**, 6112-6118 (2001).

128 Newman, M. E. J. A measure of betweenness centrality based on random walks. *Soc Networks* **27**, 39-54 (2005).

129 Krause, R., von Mering, C. & Bork, P. A comprehensive set of protein complexes in yeast: mining large scale protein-protein interaction screens. *Bioinformatics* **19**, 1901-1908 (2003).

130 Medina, I. *et al.* Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic acids research* **38**, W210-213 (2010).

131 Boyle, E. I. *et al.* GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics (Oxford, England)* **20**, 3710-3715 (2004).

132 Zachariae, W. *et al.* Mass spectrometric analysis of the anaphase-promoting complex from yeast: Identification of a subunit related to cullins. *Science* **279**, 1216-1219 (1998).

133 Visintin, R., Prinz, S. & Amon, A. CDC20 and CDH1: A family of substrate-specific activators of APC-dependent proteolysis. *Science* **278**, 460-463 (1997).

134 Harper, J. W., Burton, J. L. & Solomon, M. J. The anaphase-promoting complex: it's not just for mitosis any more. *Genes Dev* **16**, 2179-2206 (2002).

135 Dwight, S. S. *et al.* Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Research* **30**, 69-72 (2002).

136 Sacksteder, K. A. & Gould, S. J. The genetics of peroxisome biogenesis. *Annu Rev Genet* **34**, 623-652 (2000).