國立交通大學

生物資訊及系統生物研究所

碩士論文

以模板導向方法建立蛋白質-蛋白質交互作用家族

Template-driven Approaches for Protein-protein Interaction Families

研 究 生：林峻宇

指導教授：楊進木　教授

中華民國九十九年五月

以模板導向方法建立蛋白質-蛋白質交互作用家族

Template-driven Approaches for Protein-protein Interaction Families

研究生：林峻宇　　Student：Chun-Yu Lin

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

碩士論文

A Thesis Submitted to Institute of Bioinformatics and Systems Biology

College of Biological Science and Technology

National Chiao Tung University in partial Fulfillment of the Requirements

for the Degree of Master in

Bioinformatics and Systems Biology

May 2009

Hsinchu, Taiwan, Republic of China

中華民國九十九年五月

以模板導向方法建立蛋白質-蛋白質交互作用家族

學生：林峻宇　　　　　　　　　　　　　　　　　　指導教授：楊進木

國立交通大學 生物資訊與系統生物所碩士班

摘　　要

將蛋白質分類成家族(family)可幫助研究者更深入瞭解蛋白質功能和彼此間的演化關係。同樣地，因應蛋白質-蛋白質交互作用(protein-protein interaction, 簡稱PPI)資料的快速增加(大部分來自高量高速篩選實驗)，研究者為了瞭解新辨識出來的蛋白質-蛋白質交互作用，迫切地需要快速且準確的方法將蛋白質-蛋白質交互作用分類成由同源蛋白質-蛋白質交互作用(homologous PPI)所組成的家族。針對這個議題，我們提出了一個新概念：蛋白質-蛋白質交互作用家族(PPI family)，並分別以之建立PPISearch以及SB-HomPPI兩種模板導向方法。PPISearch (http://gemdock.life.nctu.edu.tw/ppisearch)是一個可迅速搜尋蛋白質交互作用家族的工具，同時也能合理地註解未知性質的蛋白質交互作用，這些註解(annotation)的內容包括功能性區塊(domain)和生化功能(biochemical function)。本研究指出，當某蛋白質-蛋白質交互作用與其提問蛋白質對(query protein pair)間具有顯著的序列相似性(BLASTP *E*-values $\le 10^{-40}$)時，而且該交互作用也已被記錄在大型PPI資料庫(包含來自 576 個物種的 290,137 筆PPIs)中，則此交互作用為該提問蛋白質對之homologous PPI。我們的結果顯示，高達 88%和 69%的功能性區塊及生化功能註解可以合理地由homologous PPI轉移至其提問蛋白質對。

然而仍有兩個問題。其一，我們建立的大型 PPI 資料庫中每個物種的 PPI 數量並不平均，少數的物種佔據大量的 PPIs 紀錄，尤其以酵母菌(yeast)為最。其二，我們使用局部序列比對工具(如 BLASTP)尋找同源蛋白質，會偏向具有長序列的功能性區塊，但此區域並不一定參與交互作用。針對這樣的問題，我們結合 PPI family 的概念和本實驗室先前提出的「立體功能區域交互作用同源性對應(3D-domain interologs)」方法，建立一個新的方法「SB-HomPPI」。SB-HomPPI 以異二聚體結構(heterodimer structures)之交互作用界面(interface)做為模板，橫跨多個具有完整基因組的物種(如 Integr8 資料庫)來辨識一個以結構為基礎之蛋白質-蛋白質交互作用家族(structure-based PPI family, 簡稱 SB-PPI family)，此家族是由具有相似交互作用界面結構之同源蛋白質-蛋白質交互作用(SB-HomPPIs)所組成。此方法論使用 Integr8 資料庫，可避免 PPI 資料庫造成的限制；針對 interface，則可以修正 BLASTP 局部序列比對所造成的誤判。我們的結果顯示，SB-PPI family (94%)與 PPI family (86%)在交互作用功能區塊對(domain pair)呈現出很高的保留程度。類似的結果也出現在 Gene Ontology (GO)註解對的保留程度分析上，同時也發現 SB-PPI family 高出 PPI family 超過 30%。綜合以上所述，交互作用功能區塊對及 GO 註解對在蛋白質-蛋白質交互作用家族是高度保留的生物特性。

Template-driven Approaches for Protein-protein Interaction Families

Student: Chun-Yu Lin　　Advisor: Dr. Jinn-Moon Yang

Institute of Bioinformatics and System Biology
National Chiao Tung University

ABSTRACT

Classifying proteins to families provides a description of the functional and evolutionary relationships of proteins. Likewise, as an increasing number of protein-protein interactions (PPIs) become available and high-throughput experiments provide systematic identification of PPIs, there is a growing need for fast and accurate approaches to classify PPIs into families (*i.e.*, a group of homologous PPIs) to understand a newly determined PPI. To address this issue, we proposed a concept "PPI family" to construct new template-driven approaches "PPISearch" and "SB-HomPPI". PPISearch is a tool (http://gemdock.life.nctu.edu.tw/ppisearch) that rapidly identifies PPI family and infers transferability of interacting domains and functions of a query protein pair. We identified homologous PPIs when these protein pairs have significant joint sequence similarity (BLASTP *E*-values $\leq 10^{-40}$) with the query sequences and were in the annotated database (290,137 PPIs in 576 species). Our results demonstrated that the transferability of conserved domain-domain pairs and conserved function term pairs between query pairs and homologous PPIs are 88% and 69%, respectively.

However, we found that the annotated database is dominated by few species, especially yeast, and the method of searching homologs by local alignment (*i.e.*, BLASTP) has a bias in favor of the large domain but that may not involve in binding interface. For these questions, we combined the concept of "PPI family" and our previous study "3D-domain interologs" to construct the approach "SB-HomPPI". The SB-HomPPI identifies structure-based PPI family (SB-PPI family), which is composed of structure-based homologous PPIs (SB-HomPPIs), across multiple complete genomes (i.e., Integr8 database) by using the interfaces of heterodimer structures as templates. This approach uses the Integr8 database and emphasizes the interface to avoid the limitation of the annotated database and the bias of searching homologs by local alignment using BLASTP. Our results presented that SB-PPI family (94%) and PPI family (86%) are highly conserved in interacting domain pairs. Similarly, SB-PPI family was better (at least 30%) than PPI family in conservations of Gene Ontology (GO) term pairs. In conclusion, interacting domain pairs and GO term pairs are the highly conserved biological properties in family.

誌謝

這本論文的完成，背後是眾多良師益友的支持與協助，沒有你們的幫助，峻宇一人是無法完成的。請容我向大家致上最深的感謝。

首先，峻宇何其幸運能受教於恩師楊進木教授，您對研究的熱忱及嚴謹，讓峻宇學習到探索與發現的樂趣，以及正確的研究態度。此外，在生活及待人處事方面，您也讓峻宇了解到許多該糾正的缺點及不足之處。在您的教導下，學生也感受到科學領域的廣闊無邊，而您總是不厭其煩幫助學生找到正確的方向及目標，十分感謝您！峻宇真的獲益良多。

感謝實驗室的學長姐們、同儕們以及學弟妹們：俊辰學長及宇書學長，感謝你們在研究上帶我學習成長，更總是耐著性子聽我陳述，幫我歸納重點，在準備論文的時期，更是不斷給予協助，沒有你們，峻宇不可能順利完成這本論文，真的真的很感謝你們！其樺學長，感謝你在程式及網頁建置上不厭其煩的指導。阿甫學長、章維學長、PIKI學長、志達學長、彥修學長以及敬立學長，感謝你們在研究上的給予的建議與幫助。怡馨學姐感謝妳總是在大家煩悶時，默默為實驗室帶來歡笑與活力。還有我的同學們，力仁、超哥、偉帆，大夥總是彼此幫忙及成長。最後，怡瑋、御哲及伸融，在生物知識及程式方面也給了我很多的協助，感謝大家，你們都是我的良師益友。

在這兩年的研究所期間，家人一直是我最大的支柱。我親愛的父親林昆民先生，母親葉淑燕女士，以及總是特地選我回家時回來陪我的老弟家宏，感謝你們一直以來的支持與鼓勵，無論開心或難過，你們總是在背後支持著我，家永遠是我最溫暖的避風港。當然，如同家人般，我最親愛的女友佳慧，在我低潮時陪我伴我，聽我訴苦，一直以來陪我走過五年的歲月，在這條路上，有妳相伴，真好！

十分感謝熊昭教授、王雯靜教授以及黃鎮剛教授願意擔任峻宇的口試委員，給予指導與學習的機會。峻宇在這兩年的研究所期間，得到太多人的幫助與支持，真的很幸運來到BioXGEM這個大家庭，最後，致上峻宇最深切的感謝，感恩大家，謝謝！

Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Background

Classifying proteins to families provides a description of the functional and evolutionary relationships of proteins, such as Pfam[1] that classifies protein domains through multiple sequence alignments and profile hidden Markov models and PIRSF[2] in that members are homologous and homeomorphic that sharing common ancestry and full-length sequence similarity with common domain architecture. Additionally, SCOP[3] and CATH[4] classify protein structures to families through homologous relationship in evolution.

In recent studies, interactions between proteins are critical to most biological function. To identify and characterize PPIs and their networks, many high-throughput experimental approaches[5-6], such as yeast two-hybrid screening, mass spectroscopy, and tandem affinity purification, and computational methods (phylogenetic profiles[7], known 3D complexes[8], and interologs[9]) have been proposed[10]. Some PPI databases, such as IntAct[11], BioGRID[12], DIP[13], MIPS[14], and MINT[15], have accumulated PPIs submitted by biologists, and those from mining literature, high-throughput experiments, and other data sources.

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[16-17] applied homologous interactions to assess the reliability of PPIs identified by large-scale experiments. The discovery of sequence homologs to a known protein often provides clues for understanding the function of a newly sequenced gene. As these interaction databases continue growing in size, they become increasingly useful for analysis of newly identified interactions and classification of protein-protein interactions.

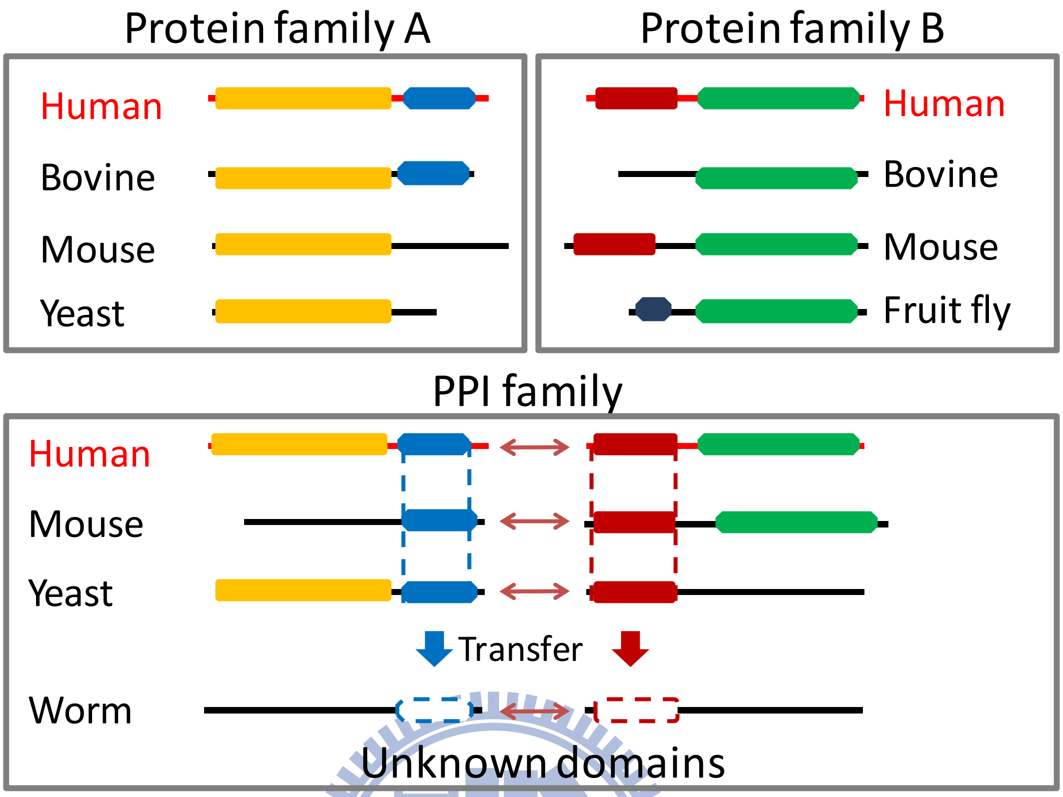

Figure 1. Comparison of protein family and PPI family. The PPI family may emphasize the properties that are involved in protein-protein interacting, such as domains and functions.

Therefore, how could we classify PPIs into families? A PPI family may consist of a group of homologous PPIs. Comparing with the protein family, the PPI family may highlight the characteristics that are involved in protein-protein interacting, such as domains and functions (Figure 1).

In sequence level, we proposed the PPISearch server[18] 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. PPISearch is an easy-to-use web server that allows users to input a pair of protein sequences. Then, this server finds homologous PPIs in multiple species from five public databases (IntAct, MIPS, DIP, MINT, and BioGRID) and annotates the query. Our results demonstrated that this server achieves high agreements on interacting domain-domain pairs and function pairs between query protein pairs and their corresponding homologous PPIs.

The structural information of interacting domains and atomic details for thousands of directly physical interactions between proteins are available[19-20]. There were many studies analyzed the PPIs on residue-based binding models and to derive domain-domain interaction (DDI) databases, such as 3did[21], iPfam[22], and DAPID[23] using three-dimensional (3D) dimer protein structures recorded in Protein Data Bank (PDB)[24]. For predicting a PPI by searching a 3D-complex library to identify homologous templates of this pair of query protein sequences by accessing interface preference according to how they fit the known template structures, some methods have proposed by utilizing template-based methods (*e.g.*, comparative modeling[8,19] and fold recognition[20]). Despite the diversity of the strategies and algorithms used in these methods, they all focus on predicting the protein-protein interactions and seem to produce comparable results. For a query protein pair (or 3D-dimer template), these methods lack PPI families (*i.e.*, paralogous PPIs in one species and orthologous PPIs across multiple genomes) for studying evolution of PPIs. In addition, it is time-consume (unfeasible) to query all protein pairs in multiple complete genomes (such as 6,352,363 proteins with 2,274 complete genomes in Integr8[25], having ~2.0×10^{10} possible pairs).

We combine two concepts that are protein-protein interaction family (PPI family) and 3D-domain interolog mapping[8] to identify structure-based homologous protein-protein interactions (SB-homPPIs) across multiple complete genome. The 3D-domain interologs is similar to "generalized interologs mapping". Our concept is defined as "Domain *a* (in chain A) interacts with domain *b* (in chain B) in a known 3D complex, meaning that their inferring protein pair A' (containing domain *a*) and B' (containing domain *b*) in the same species would be likely to interact with each other if two pairs, *i.e.*, (A, A') and (B, B'), are homologous". Based on the new approach of "SB-homPPI", we could infer the interacting domain-domain pairs, function pairs (Gene Ontology annotations[26]), the binding models (*e.g.*, hydrogen-bond interactions and conserved residues), couple-conserved residues and the evolution of the PPIs. In this study, we used 1,895 3D heterodimers and our scoring functions to infer 224,713 PPIs by searching on the Integr8 database.

1896

1.2 Motivation

In biological systems, interactions between proteins are critical. Currently, the description of the functional and evolutionary relationships of proteins could be provided by classifying proteins to families (*e.g.*, Pfam and SCOP). However, evaluation of the relationship between sequence homology and function is ambiguous, because no clear measure of functional similarity exists in interacting protein pairs[27-28]. As interaction databases continue growing in size, PPI families may offer biologists to understand newly identified interactions and evolutionary relationship of homologous PPIs in a family.

To address this issue, we proposed the PPISearch server for searching homologous PPIs (*i.e.*, PPI family) across multiple species and annotating consensus domain-domain pairs and molecular functions to the query protein pair. According to our knowledge, PPISearch is the first public approach that identifies homologous PPIs from annotated PPI databases and infers transferability of interacting domains and functions between homologous PPIs and the query. Our results demonstrated that this approach achieves high agreements on interacting domain-domain pairs and function pairs between query protein pairs and their corresponding homologous PPIs.

Table 1. The list of the numbers of proteins and PPIs in 11 common used organisms

NCBI Taxonomy ID	Organisms	No. Proteins in Integr8 database	No. PPIs in the annotated database
9606	*Homo sapiens*	56,006	45,963
10090	*Mus musculus*	36,379	4,367
3702	*Arabidopsis thaliana*	35,825	4,192
6239	*Caenorhabditis elegans*	23,154	10,117
7227	*Drosophila melanogaster*	15,155	43,166
7955	*Danio rerio*	21,601	44
10116	*Rattus norvegicus*	13,807	1,748
9913	*Bos taurus*	12,235	255
9031	*Gallus gallus*	6,279	45
36329	*Plasmodium falciparum*	5,353	2,737
4932	*Saccharomyces cerevisiae*	5,727	113,161
	Total	231,521	225,795

However, there are some constraints on the approach of PPISearch. Firstly, the annotated PPI database composed of five public databases is dominated by few species, especially yeast (Table 1). Secondly, the method of searching homologs by local sequence alignment using BLASTP has a bias in favor of the large domain which may be not involved in protein-protein interacting. To improve these problems, we used "3D-domain interolog mapping", which represents the interface of 3D heterodimers as the template, to identify SB-HomPPIs in Integr8 database (a database of multiple complete genomes). The structure-based protein-protein interaction family (SB-PPI family) consists of these SB-HomPPIs. Based on the binding models (*e.g.*, hydrogen-bond interactions and conserved residues) of interfaces, we could infer couple-conserved residues and the evolution of the PPIs through multiple sequence alignments. Our results showed that the SB-PPI family between the 3D heterodimers and their corresponding SB-HomPPIs according to our results and the PPI family are highly conserved in interacting domain pairs and Gene Ontology (GO) term pairs.

1.3 Thesis overview

The thesis consists of the two studies "PPI family" and "SB-PPI family". Their frameworks were shown in **Section 2.1** and **2.6**, respectively. We first proposed the concept of identifying homologous PPIs across species to construct a web server, PPISearch. Moreover, we used case studies to present the limitations and biases of the PPI family (**Section 3.8**), and the improvements that used on the new approach to identify the SB-PPI family. We proposed evidence to demonstrate the conservation of interacting domain pairs and GO term pairs and predicting ability for PPI family and SB-PPI family. The reliable PPI family will be identified through combining PPI family and SB-PPI family and also help us to verify the reliability of the members in a family.

Chapter 2 Methods and Materials

2.1 Overview of identifying PPI family

In this section, we present the concept and the approaches of identifying protein-protein interaction family through sequence similarity (*e.g.*, *E*-value and joint *E*-value). Figure 2 illustrates the concept of identifying protein-protein interaction family.

For this purpose, we define a group of homologous protein-protein interactions forming a PPI family. All homologous PPIs in the family are similar protein-protein interactions that share a common ancestry. The following steps show the details of the PPISearch server to search homologous PPIs of a query protein pair (A and B) (Figure 3A). This server first identifies the homologous families (A' and B') of A and B, respectively, with *E*-value $\leq 10^{-10}$ by using BLASTP to scan the annotated PPI databases (Figures 3B and C). All protein pairs of A' and B' are considered candidates of homologous PPIs. We selected homologous PPIs from these candidates, which are recorded in the annotated databases, and have significant joint sequence similarity [*E*-value $\leq 10^{-40}$, Equation (1)] between candidates and the query (Figure 3D). Then, we measured the conservation ratios of domain-domain pairs (DDPs; Pfam domains) and protein functions (Gene Ontology annotations) derived from these homologous PPIs of the query (Figure 3E).

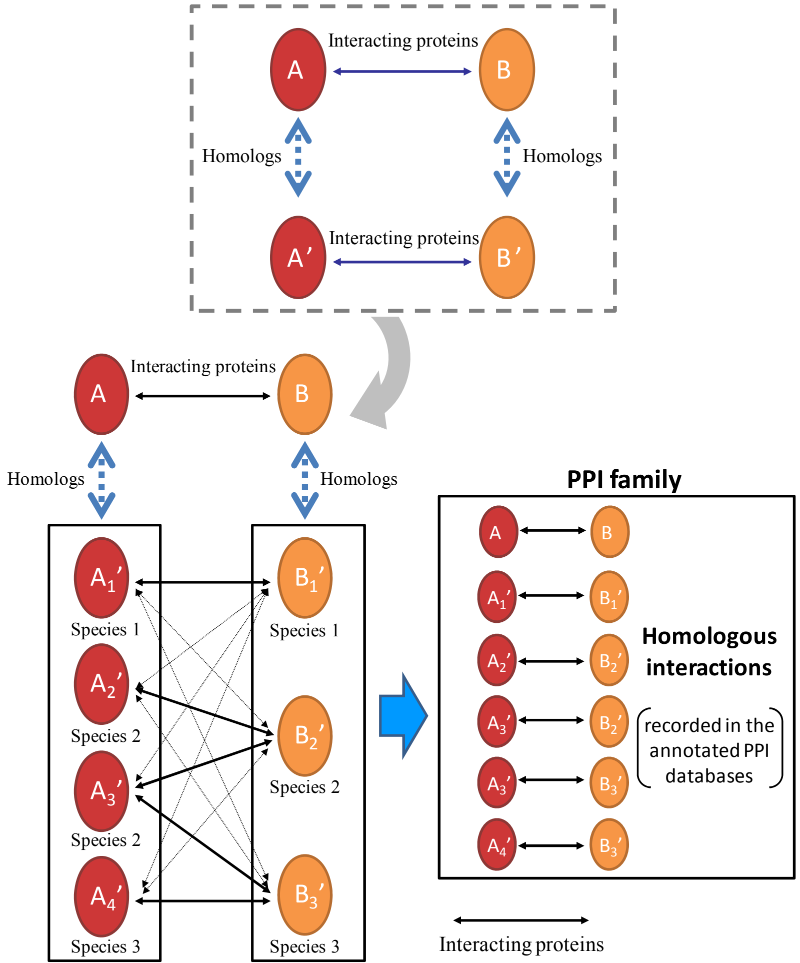

Figure 2. Illustration of identifying a protein-protein family. A protein-protein interaction (A-B) is the query protein pair given by users. A' and B' are the homologs of proteins A and B, respectively. Two pairs, (A, B) and (A', B'), are similar. The homolog pair A'-B' are considered homologous interaction when it is recorded in the annotated PPI database (290,137 PPIs). The protein-protein family consists of all homologous interactions (*e.g.*, A_1'-B_1' and A_2'-B_2') and the query pair.

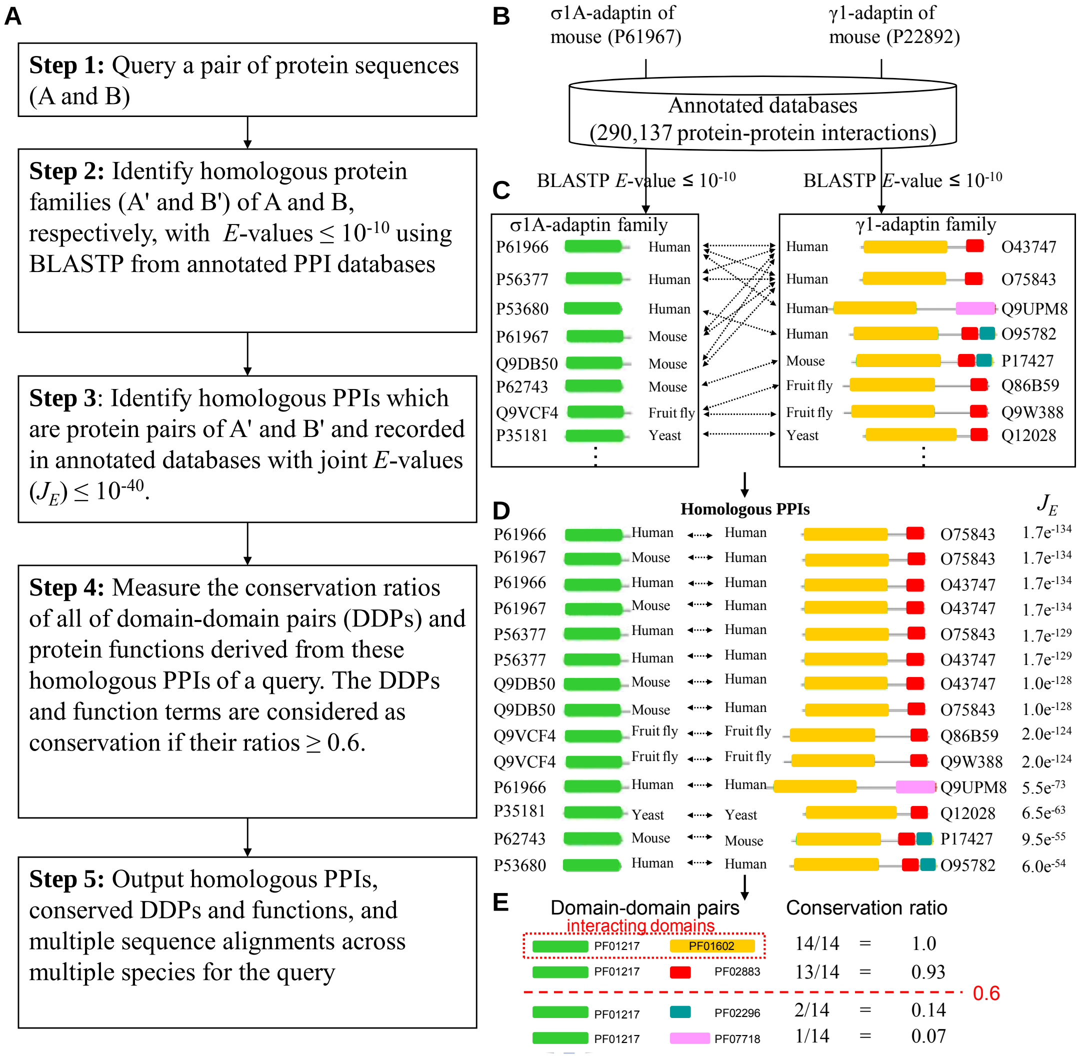
A
Step 1: Query a pair of protein sequences (A and B)
Step 2: Identify homologous protein families (A' and B') of A and B, respectively, with E-values ≤ 10-10 using BLASTP from annotated PPI databases
Step 3: Identify homologous PPIs which are protein pairs of A' and B' and recorded in annotated databases with joint E-values (JE) ≤ 10-40.
Step 4: Measure the conservation ratios of all of domain-domain pairs (DDPs) and protein functions derived from these homologous PPIs of a query. The DDPs and function terms are considered as conservation if their ratios ≥ 0.6.
Step 5: Output homologous PPIs, conserved DDPs and functions, and multiple sequence alignments across multiple species for the query
B
σ1A-adaptin of mouse (P61967)
γ1-adaptin of mouse (P22892)
Annotated databases (290,137 protein-protein interactions)
C
BLASTP E-value ≤ 10-10
BLASTP E-value ≤ 10-10
σ1A-adaptin family
P61966 Human
P56377 Human
P53680 Human
P61967 Mouse
Q9DB50 Mouse
P62743 Mouse
Q9VCF4 Fruit fly
P35181 Yeast
γ1-adaptin family
Human O43747
Human O75843
Human Q9UPM8
Human O95782
Mouse P17427
Fruit fly Q86B59
Fruit fly Q9W388
Yeast Q12028
D
Homologous PPIs
JE
P61966 Human ↔ Human O75843 1.7e-134
P61967 Mouse ↔ Human O75843 1.7e-134
P61966 Human ↔ Human O43747 1.7e-134
P61967 Mouse ↔ Human O43747 1.7e-134
P56377 Human ↔ Human O75843 1.7e-129
P56377 Human ↔ Human O43747 1.7e-129
Q9DB50 Mouse ↔ Human O43747 1.0e-128
Q9DB50 Mouse ↔ Human O75843 1.0e-128
Q9VCF4 Fruit fly ↔ Fruit fly Q86B59 2.0e-124
Q9VCF4 Fruit fly ↔ Fruit fly Q9W388 2.0e-124
P61966 Human ↔ Human Q9UPM8 5.5e-73
P35181 Yeast ↔ Yeast Q12028 6.5e-63
P62743 Mouse ↔ Mouse P17427 9.5e-55
P53680 Human ↔ Human O95782 6.0e-54
E
Domain-domain pairs
Conservation ratio
interacting domains
PF01217 PF01602 14/14 = 1.0
PF01217 PF02883 13/14 = 0.93
0.6
PF01217 PF02296 2/14 = 0.14
PF01217 PF07718 1/14 = 0.07

Figure 3. Overview of the PPISearch server for homologous protein-protein interaction search and conservation analysis using proteins σ1A-adaptin and γ1-adaptin as the query. (A) The main procedure. (B) Identify homologs of σ1A-adaptin and γ1-adaptin using BLASTP to scan the annotated PPI databases. (C) The homologous families of σ1A-adaptin and γ1-adaptin with E-values $\leq 10^{-10}$. (D) Homologous PPIs of the query. (E) Conservation ratios of domain-domain pairs derived from homologous PPIs.

2.2 Homologous PPIs of a PPI family

The concept of PPI family is the core of the PPISearch server to identify homologous PPIs and measure DDPs and functional conservations of a query protein pair (A and B). We define a homologous PPI as follows: (1) homologs of A and B are proteins with significant sequence similarity BLASTP E-values $\leq 10^{-10}$ [9,29]; (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 corresponding homologs (A_1' and B_1') recorded in annotated PPI databases. This work followed previous studies[9,29] to define joint sequence similarity as

$$J_E = \sqrt{E_A \times E_B} \tag{1}$$

where 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'. Here, $J_E \leq 10^{-40}$ is considered a significant similarity according to statistical analysis of 290,137 annotated PPIs and 6,597 orthologous PPI families collected from the PORC database[25].

2.3 The annotated PPI database

We totally collected 290,137 PPIs as the annotated PPI database in that duplications were removed by using UniProt accession numbers from five public databases (*e.g.*, 147,634 PPIs in IntAct, 18,529 PPIs in MIPS, 52,445 PPIs in DIP, 77,846 PPIs in MINT, and 150,827 PPIs in BioGRID). These PPIs were identified experimentally from 576 species.

2.4 Annotations of homologous PPI

A query protein pair and its homologous PPIs, which have significant sequence and joint sequence similarity, can be considered a PPI family. The concept of PPI family is derived from that of protein sequence family[1] and protein structure family[3]. We believed that PPI families can be applied widely in biological investigations. Here, we assumed that the members of a PPI family are conserved on specific functions and in interacting domain(s). Using these conservations of query' homologous PPIs, our server could be used to annotate the protein functions and DDPs of a query protein pair.

2.4.1 Transferability of domain-domain pairs

A query protein pair and its homologous PPIs can often agree on interacting DDPs. To measure the agreement of each DDP in a PPI family, we define the conservation ratio (CRD_p) of a DDP p in homologous PPIs of a query protein pair i as

$$CRD_p = \frac{\text{Number of homologous PPIs with a domain pair } p}{\text{Number of homologous PPIs of query } i} \tag{2}$$

Figures 3D and E show an example to calculate the *CRD* values of four DDPs. In addition, to statistically evaluate the transferability of DDPs between a query and its homologous PPIs, this study defines the shared ratio (*SRD*) of DDPs using CRD_p and 290,137 annotated PPIs as query protein pairs. The *SRD* of DDPs against different ratio c is given as

$$SRD = \frac{\sum_{i \in Q} d_i(CRD_p \geq c)}{\sum_{i \in Q} D_i(CRD_p \geq c)} \tag{3}$$

where Q is a set of annotated PPIs in databases (here, the total number of PPIs in Q is 290,137); i is a query protein pair; $d_i(CRD_p \geq c)$ is the number of DDPs with CRD_p values exceeding c; and these DDPs are shared by the query i and its homologous PPIs. $D_i(CRD_p \geq c)$ is the total number of the DDPs with $CRD_p \geq c$, where DDPs are derived from homologous PPIs of the query i. Here, this work used a statistical approach to determine the threshold c (here, c=0.6) of CRD_p to yield reliable DDP annotations with an acceptable level of D_i. Please note that CRD_p and *SRD* are computed from a query protein pair and a set of queries, respectively.

2.4.2 Transferability of Gene Ontology

We assumed that the members of a PPI family are usually conserved on specific molecular function, pathway, and cellular component. We utilize the Gene Ontology[26] to annotate the molecular function, biological process, and cellular component of a query protein pair. To statistically evaluate the shared ratio of GO terms between the query pair and its PPI family (with N homologous PPIs), we define the shared ratio (*SR*) using the conservation ratio ($CR=N_a/N$), where N_a is the number of homologous PPIs with the same GO term in a PPI family. The *SR* is given as

$$SR = \frac{\sum_{i \in Q} T_i(CR \geq k)}{\sum_{i \in Q} P_i(CR \geq k)} \qquad (4)$$

where Q is a set of query pairs; P_i ($CR \geq k$) is the total number of the GO terms of query pair i when $CR \geq k$; T_i ($CR \geq k$) is the number of the shared GO terms of query pair i when $CR \geq k$. The shared ratio of MFPs (*SRF*) is statistically derived from 290,137 annotated queries. Here, k is set to 0.6.

2.5 Data sets for evaluating the approach of identifying PPI family

To evaluate the usefulness of the PPISearch for the discovery of PPI family and for the annotations of a query protein pair, we selected two query protein sets, termed HOM and ORT. For searching homologous PPIs, HOM and ORT data sets are used to assess performance of PPI family and to determine the threshold of joint *E*-value J_E [Equation (1)].

The HOM set includes all of 290,137 PPIs and the ORT set has 6,597 orthologous PPI families (14,571 PPIs) derived from the annotated PPI database and PORC orthology database. PORC data (putative orthologous clusters) were defined as orthologous families from Integr8 and CluSTr[30] databases. These clusters contain all sequenced organisms (1,125 bacteria, 125 eukaryota and 50 archaea in the release 94). Each entry in PORC represents a cluster of genes grouped by the similarity of their longest protein product. According to the construction process of PORC, a gene cluster contains at most a single protein from a given species and a protein can be assigned to only a single cluster.

2.6 Overview of structure-based PPI family (SB-PPI family)

The structure-based protein-protein family (SB-PPI family) is extended from PPI family[18] and 3D-domain interologs with structural template-based scoring function[8]. The 3D-domain interologs is defined as "Domain *a* interacts with domain *b* in the protein pair (A-B) of a known 3D complex, their inferring homologous protein pair A' (containing domain *a*) and B' (containing domain *b*) in the same species would be likely to interact with each other." Figure 4 presents the approach of identifying SB-PPI family. We define a SB-PPI family consists of a group of structure-based homologous protein-protein interactions (SB-HomPPIs). All PPIs in the family are similar protein-protein interactions that sharing a common ancestry and having similar binding model (*i.e.*, significant interface similarity).

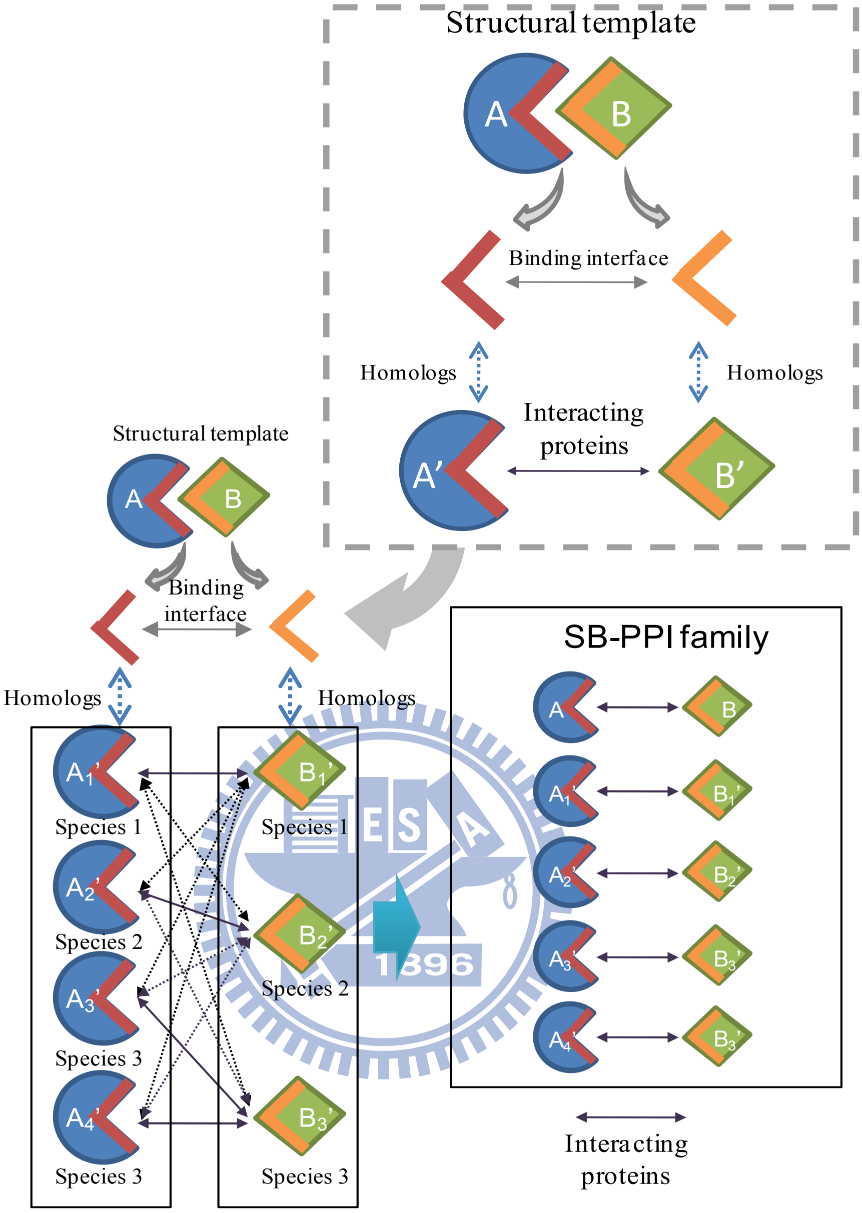

Figure 4. Illustration of identifying a structure-based protein-protein family. A 3D-dimer template (A-B) is the query protein pair given by users. The domain *a* and *b* are the binding interface of A-B. Protein A' and B' that containing domain *a* and *b* are the homologs of proteins A and B, respectively. The homologous pairs A'-B' have similar binding model with the query pair A-B. The structure-based protein-protein family consists of all homologous interactions (*e.g.*, A_1'-B_1' and A_2'-B_2') and the query pair.

Figure 5 shows the framework of identifying SB-HomPPIs of the 3D-dimer template utilizing 3D-domain interolog mapping. We first identifies the homologous families (A' and B') (Figure 5A) of a 3D-dimer template T (Figure 5B), which has proteins A and B (with interacting domains a and b), with significant sequence similarity (PSI-BLASTP E-values $\leq 10^{-10}$) searching in the Integr8 database (6,352,363 protein sequences in 2,274 species). All protein pairs of A' and B' (with interacting domains a and b) are considered the SB-HomPPIs (a SB-PPI family) of the template T if their interfaces have significant interface similarity with the template T (Figure 5C) (See **Section 2.7**). For a SB-PPI family, we can measure the conservation ratios of domain-domain pairs (DDPs) and Gene Ontology annotations, including molecular function (MF), biological process (BP) and cellular component (CC). Additionally, the multiple sequence alignments of interfaces further provide the analysis of co-evolution in contact residue pairs in interfaces (Figure 5D).

E S A 1896

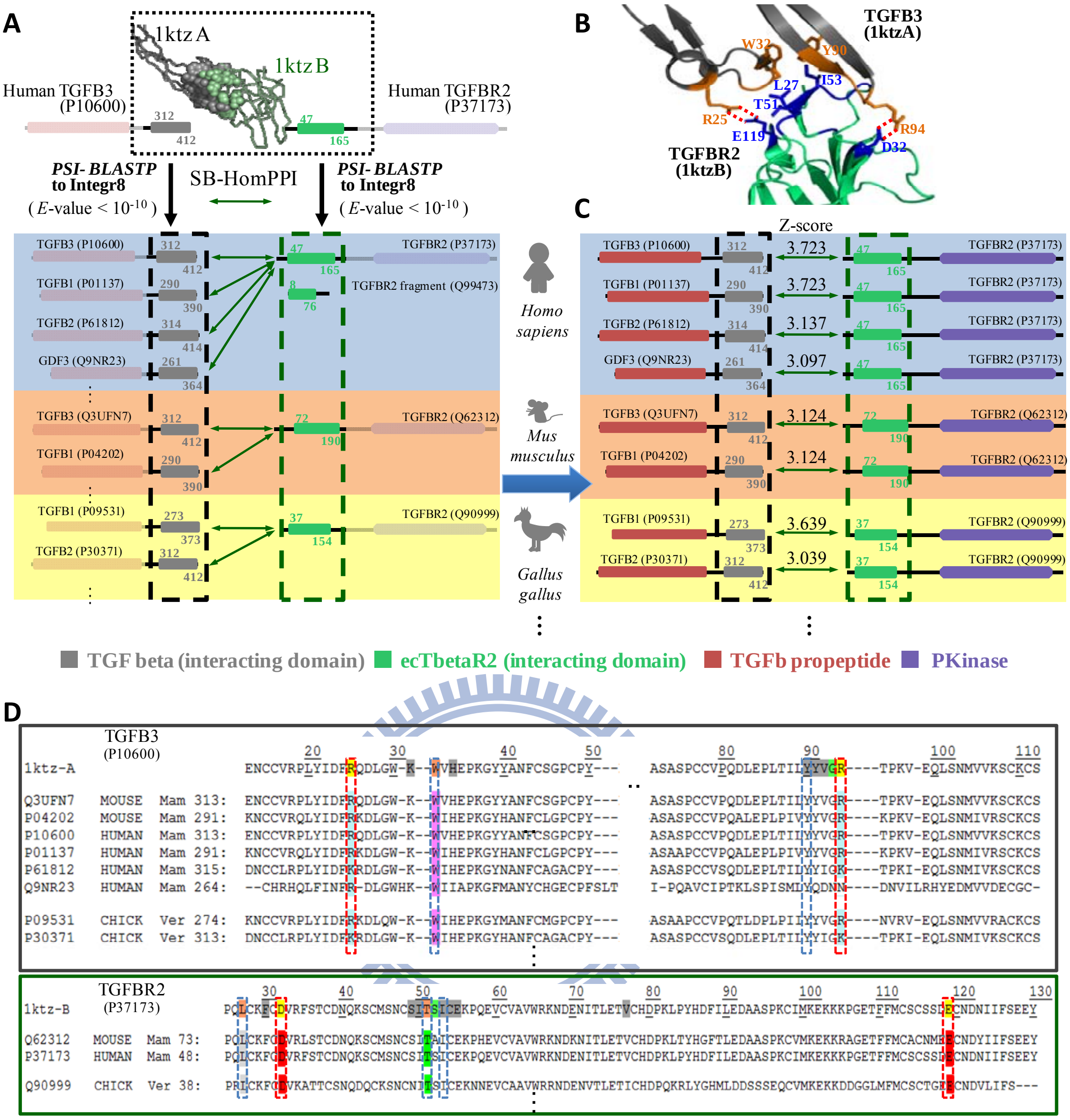
A
1ktz A
1ktz B
Human TGFB3 (P10600)
Human TGFBR2 (P37173)
312
412
47
165
PSI-BLASTP to Integr8 (E-value < 10-10)
SB-HomPPI
PSI-BLASTP to Integr8 (E-value < 10-10)
TGFB3 (P10600)
TGFB1 (P01137)
TGFB2 (P61812)
GDF3 (Q9NR23)
TGFBR2 (P37173)
TGFBR2 fragment (Q99473)
8
76
TGFB3 (Q3UFN7)
TGFB1 (P04202)
TGFBR2 (Q62312)
72
190
TGFB1 (P09531)
TGFB2 (P30371)
TGFBR2 (Q90999)
37
154
290
390
314
414
261
364
273
373
Homo sapiens
Mus musculus
Gallus gallus
B
TGFB3 (1ktzA)
TGFBR2 (1ktzB)
W32
Y90
L27
I53
T51
R25
E119
R94
D32
C
Z-score
3.723
3.723
3.137
3.097
3.124
3.124
3.639
3.039
TGF beta (interacting domain)
ecTbetaR2 (interacting domain)
TGFb propeptide
PKinase
D
TGFB3 (P10600)
1ktz-A ENCCVRPLYIDFRQDLGW-K--WVHEPKGYYANFCSGPCPY--- ASASPCCVPQDLEPLTILYYVGR----TPKV-EQLSNMVVKSCKCS
Q3UFN7 MOUSE Mam 313: ENCCVRPLYIDFRQDLGW-K--WVHEPKGYYANFCSGPCPY--- ASASPCCVPQDLEPLTILYYVGR----TPKV-EQLSNMVVKSCKCS
P04202 MOUSE Mam 291: KNCCVRQLYIDFRKDLGW-K--WIHEPKGYHANFCLGPCPY--- ASASPCCVPQALEPLPIVYYVGR----KPKV-EQLSNMIVRSCKCS
P10600 HUMAN Mam 313: ENCCVRPLYIDFRQDLGW-K--WVHEPKGYYANFCSGPCPY--- ASASPCCVPQDLEPLTILYYVGR----TPKV-EQLSNMVVKSCKCS
P01137 HUMAN Mam 291: KNCCVRQLYIDFRKDLGW-K--WIHEPKGYHANFCLGPCPY--- ASAAPCCVPQALEPLPIVYYVGR----KPKV-EQLSNMIVRSCKCS
P61812 HUMAN Mam 315: DNCCLRPLYIDFKRDLGW-K--WIHEPKGYNANFCAGACPY--- ASASPCCVSQDLEPLTILYYIGK----TPKI-EQLSNMIVKSCKCS
Q9NR23 HUMAN Mam 264: --CHRHQLFINFR-DLGWHK--WIIAPKGFMANYCHGECPFSLT I-PQAVCIPTKLSPISMLYQDNN----DNVILRHYEDMVVDECGC-
P09531 CHICK Ver 274: KNCCVRPLYIDFRKDLQW-K--WIHEPKGYMANFCMGPCPY--- ASAAPCCVPQTLDPLPIIYYVGR----NVRV-EQLSNMVVRACKCS
P30371 CHICK Ver 313: DNCCLRPLYIDFKRDLGW-K--WIHEPKGYHANFCAGACPY--- ASASPCCVSQDLEPLTILYYIGK----TPKI-EQLSNMIVKSCKCS
TGFBR2 (P37173)
1ktz-B PQLCKFCDVRFSTCDNQKSCMSNCSITSICEKPQEVCVAVWRKNDENITLETVCHDPKLPYHDFILEDAASPKCIMKEKKKPGETFFMCSCSSDECNDNIIFSEEY
Q62312 MOUSE Mam 73: PQLCKFCDVRLSTCDNQKSCMSNCSITAICEKPHEVCVAVWRKNDKNITLETVCHDPKLTYHGFTLEDAASPKCVMKEKKRAGETFFMCACNMEECNDYIIFSEEY
P37173 HUMAN Mam 48: PQLCKFCDVRFSTCDNQKSCMSNCSITSICEKPQEVCVAVWRKNDENITLETVCHDPKLPYHDFILEDAASPKCIMKEKKKPGETFFMCSCSSDECNDNIIFSEEY
Q90999 CHICK Ver 38: PRLCKFCDVKATTCSNQDQCKSNCNITSICEKNNEVCAAVWRRNDENVTLETICHDPQKRLYGHMLDDSSSEQCVMKEKKDDGGLMFMCSCTGEECNDVLIFS---
20
30
40
50
80
90
100
110
60
70
120
130

Figure 5. Framework of the SB-HomPPI approach. (A) A known 3D-dimer template structure (PDB code 1ktz) which identified their homologs (E-value ≤ 10^{-10}) of TGFB3 and TGFBR2 and 204 PPI candidates (B) The interface and some interactions of the template (PDB code: 1ktz). (C) SB-PPI family uses the structural template-based interface similarity scoring function to evaluate these 204 PPI candidates and 28 structure-based homologous PPIs with significantly interface similarity (interface sequence identity ≥ 25%, and contact residue identity ≥ 25%, Z-score ≥ 3) in six species are selected to comprise the SB-PPI family of the template 1ktz. Here, 8 PPIs are selected from *Homo sapiens*, *Mus musculus*, and *Gallus gallus*. (D) SB-HomPPI utilizes two multiple sequence alignments of structure-based homologous PPIs for interface evolution analysis. Hydrogen bond: red dotted box; Hydrophobic cavity: blue dotted box.

2.7 Structure-based homologous PPIs of a structure-based PPI family

We define a SB-HomPPI as follows: (1) Homologs of binding domains in proteins A and B are proteins with significant PSI-BLASTP *E*-value ≤ 10^{-10}; (2) An interacting candidate is regarded as a SB-HomPPI if its significant interface similarity (interface sequence identity ≥ 25%, contact residue identity ≥ 25%, and *Z*-score ≥ 3) and it ranks in the Top 25 in one species. This work followed our previous studies to define homologs and similar binding model[8,18]. The binding affinity Z-score of a SB-HomPPI (proteins A_1' and B_1') of a template *T* (proteins A and B) is defined as

$$Z = \frac{E_{A_1^{'}B_1^{'}} - < E_{AB} >}{\sigma_{AB}} \tag{5}$$

where $E_{A1'\,B1'}$ is the sequence-interface similarity score of the SB-HomPPI proteins A_1' and B_1'; $<E_{AB}>$ and σ_{AB} are the mean and standard deviation of sequence-interface similarity scores of 10,000 random interfaces in a decoy set of the template *T*, respectively. $E_{A1'\,B1'}$ is given as

$$E_{A_1^{'}B_1^{'}} = E_{vdw} + E_{SF} + E_{sim} + wE_{cons} \tag{6}$$

where E_{vdw} and E_{SF} are the interacting van der Waals energy and the special interacting bond energy (i.e. hydrogen-bond energy and electrostatic energy), respectively, using four knowledge-based matrices[8]. E_{sim} is the sequence similarity score and the E_{cons} is the couple-conserved residue score of the SB-PPI family of template *T*.

2.8 3D-dimer template library and data sets for evaluating the predicting ability between the PPI family and SB-PPI family

We used the data set of 1,894 heterodimers (*i.e.*, 3,788 protein sequences, called NR-1894) from the Protein Data Bank (PDB) released in Feb 24, 2006 as the 3D-dimer template library. This set was derived from our previous study[8]. Additionally, we selected a non-redundant set, called NR-563, to evaluate conservations of interacting domain pairs and GO term pairs and performance of prediction. This set consisted of 563 heterodimer complexes selected from the NR-1894 according to their SCOP interacting-domain pairs. At least one chain of these complexes has different SCOP family.

We used the annotated PPI database (290,137 PPIs) as the positive PPIs. Among 290,137 PPIs, 173,277 interactions can be used to calculate relative specificity similarity (RSS scores)[31] of GO terms of BP and CC. The BP and CC RSS scores of 23,929 (13.81%) and

3,294 (1.90%) interactions, respectively, are less than 0.4 (Figure 6). Here, we considered a protein pair as a negative PPI if its BP and CC RSS scores are less than 0.4.

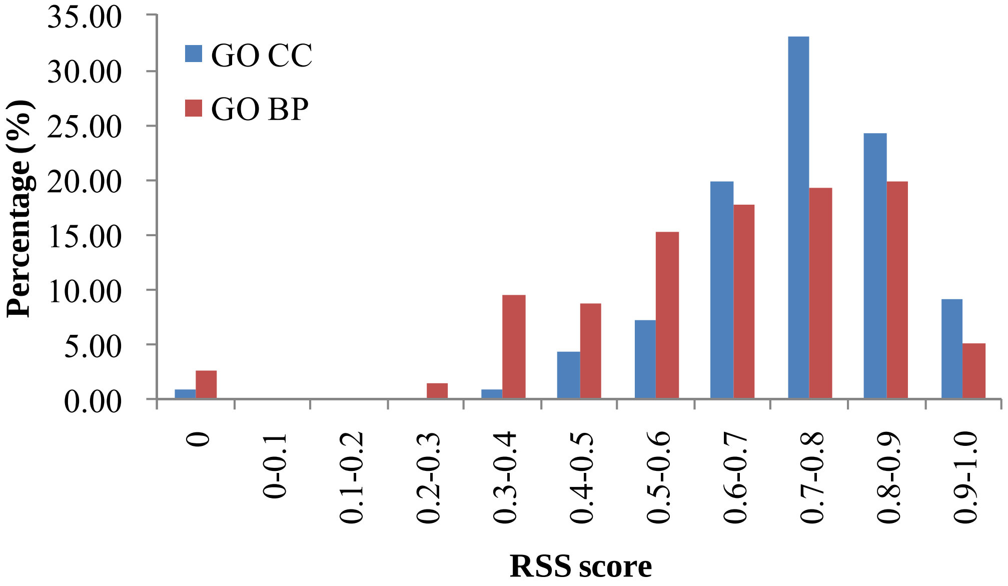

Figure 6. The RSS score distributions of the biological process (BP) and cellular component (CC). Among 290,137 PPIs recorded in the annotated database, 173,277 PPIs having both BP and CC annotations are used to calculate the BP and CC RSS scores. There are 23,929 (13.81%) and 3,294 (1.90%) PPIs with low BP and CC RSS scores (≤ 0.4), respectively.

2.9 Annotations of SB-HomPPI

A heterodimer structure and its SB-HomPPIs, which are significant in sequence and interface similarity (binding model similarity), can be considered a SB-PPI family. Here, we assume that the members of a SB-PPI family have similar binding models and are conserved on interacting domain pair(s) and biological properties.

2.9.1 Conservation of interacting domain pairs

A heterodimer structure and its SB-HomPPIs often possess the same interacting domain pair(s) (IDP). To measure the IDP conservation of a SB-PPI family, we defined the consensus ratio (*CRIDP*) of a SB-HomPPI *i* in the heterodimer structure *Q* as

$$CRIDP = \sum_{i \in Q} \frac{CIDP_i}{NIDP_i} \qquad (7)$$

where $CIDP_i$ is the number of IDPs that is consensus between the structure Q and the SB-HomPPI i; and $NIDP_i$ is the number of IDPs of the structure i. Based on 359 heterodimer structures and 14,095 PPIs (the annotated database) with IDPs annotated in iPfam database[22], we statistically evaluated the consensus of IDPs between SB-PPI families and PPI families.

2.9.2 Transferability of Gene Ontology

Similar to PPI family, we assumed that the members of a PPI family are usually conserved on specific molecular function, pathway, and cellular component. We measured the conservation ratio and the shared ratio (described in **Section 2.4.2**) of a GO term pair in a SB-PPI family of a heterodimer structure *i*. Here, the shared ratio of GO MF, BP, and CC term pairs, which are statistically derived from 281, 292, and 245 heterodimer structures with GO annotations, respectively, are utilized to estimate conserved GO term pairs shared by a SB-PPI family. In addition, we compared the transferability of Gene Ontology between PPI family and SB-PPI family.

2.10 Performance criteria

We used precision and recall to measure the predicting ability of our two approaches (*i.e.*, PPI family of PPISearch and SB-PPI family of SB-HomPPI). The precision and recall show as

$$precision = \frac{the\ number\ of\ True\ Positives}{the\ number\ of\ True\ Positives + the\ number\ of\ False\ Positives} \qquad (8)$$

$$recall = \frac{the\ number\ of\ True\ Positives}{the\ number\ of\ True\ Positives + the\ number\ of\ False\ \mathrm{N}egatives} \qquad (9)$$

SB-HomPPI used the distribution between precision and recall to decide the threshold of *Z*-score. This study also compared the distribution between the number of true positives and the precision to assess the quality of PPI family and SB-PPI family.

1896

Chapter 3. Results and Discussions

In this study, we proposed two approaches (*i.e.*, PPISearch and SB-HomPPI) of identifying PPI family and SB-PPI family. According to our knowledge, PPISearch is first server to identify PPI family through the sequence similarity. Moreover, the SB-PPI family, using structural template-based scoring function of 3D-domain interolog mapping, is more conserved than the PPI family in interacting domain pairs and GO annotations. Firstly, we provided evidence of classifying PPIs to families and more details were shown in our previous study. Some case studies were used to present the potential limitations and biases of the PPI family. Secondly, we analyzed conservations of biological properties in PPI family and SB-PPI family, such as conservation of interacting domain pairs and molecular function. Finally, protein complexes are the fundamental units of macromolecular organization[32]. We applied our concept of PPI family and the approach of SB-HomPPI to construct protein complex family for determining functional modules of biological networks across multiple species.

3.1 The criteria for identifying PPI family

HOM and ORT were used to assess the PPISearch server in identifying homologous PPIs and orthologous PPIs, respectively, by searching the annotated PPI database (290,137 PPIs with 54,422 proteins). Figure 7 shows the relationships between J_E values and number of orthologous PPIs (black) and homologous PPIs (red). The orthologous PPIs often have the same functions and domains. When $J_E \leq 10^{-40}$, the number of orthologous PPIs decreases significantly; conversely, the number of homologous PPIs decreases more gradually than that at $J_E \geq 10^{-40}$. This result showed that the proposed method is able to identify 98.2% orthologous PPIs with a reasonable number of homologous PPIs when $J_E \leq 10^{-40}$. Based on our definition and criteria, there are 90,715 PPI families excluding two proteins of the PPI from different species, and these families consist of 93,406 PPIs in 174 species.

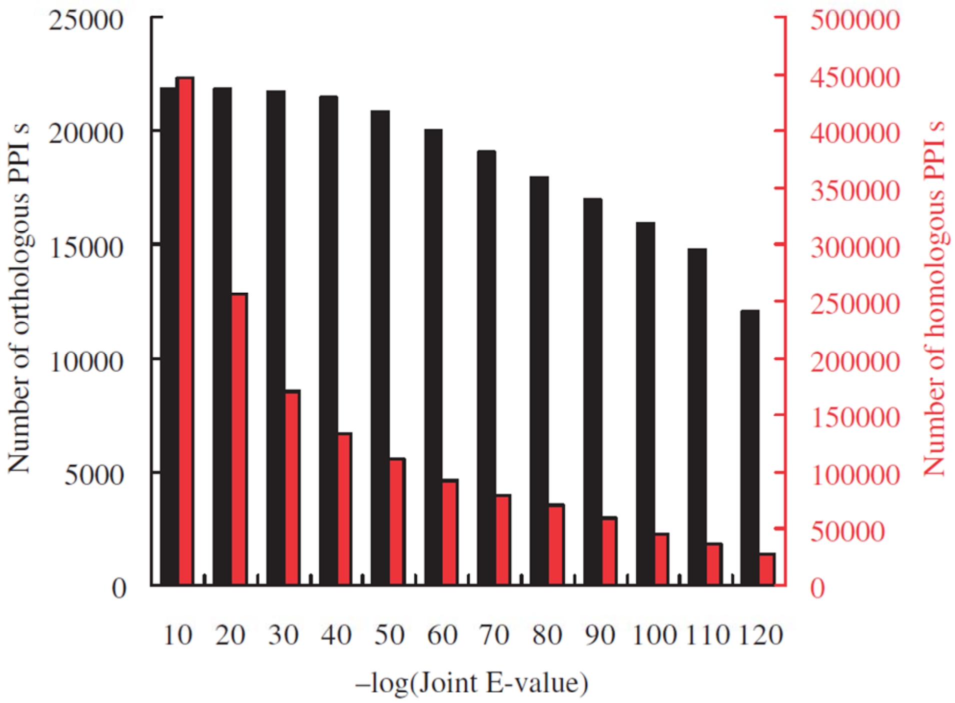

Figure 7. The relationships between joint *E*-value J_E and the numbers of orthologous PPIs (black) and homologous PPIs (red) derived from 290,137 annotated PPIs.

3.2 Conservation of domain-domain pairs and molecular function pairs in PPI family

To evaluate the transferability of DDPs and MFPs between a query and its homologous PPIs, we used the *SRD* [Equation (3)] and *SR* [Equation (4)]. The HOM set is used to evaluate the utility of the PPISearch server in annotating the query protein pair. By excluding proteins without domain annotations from the query set, 103,762 PPIs are used to evaluate the transferability (*SRD*) of conserved DDPs between these query PPIs and their corresponding homologous PPIs (Figure 8A). The transferability (*SRF*) of conserved functions between the 106,997 PPIs and their homologous PPIs is assessed by excluding proteins without molecular function terms of GO from the original query set (Figure 8B).

Figure 8A shows the relationship between conservation ratios (*CRD*) of DDPs and the *SRD* ratios. The *SRD* ratio increases significantly (solid lines) when the *CRD* increases and *CRD* ≤ 0.6. Conversely, the number of DDPs derived from 103,762 PPI families decreases (dotted lines) as *CRD* increases. If the *CRD* is set to 0.6 and the joint *E*-value is set to 10^{-40}

(green lines), the *SRD* is 0.88 (*i.e.*, 88%) and the number of DDPs is 252,728. This result demonstrated that members of a PPI family derived by PPISearch reliably share DDPs (or interacting domains). Additionally, similar results were obtained for transferability of conserved functions between homologous PPIs and the query (Figure 8B). The members of a PPI family have similar molecular functions, and *SRF* ratios are highly correlated with conservation ratios (*CRF*) of MFPs. When the *CRF* is 0.6 and the joint *E*-value is 10^{-40} (green lines), the *SRF* is 0.69 (*i.e.*, 69%) and the number of MFPs is 454,251.

These results revealed that the PPI family achieves a high *SRD* with a reasonable number of DDPs when the joint *E*-value is set to 10^{-40}. In summary, these experimental results demonstrated that this server achieves high agreement on DDPs and MFPs between the query and their corresponding homologous PPIs.

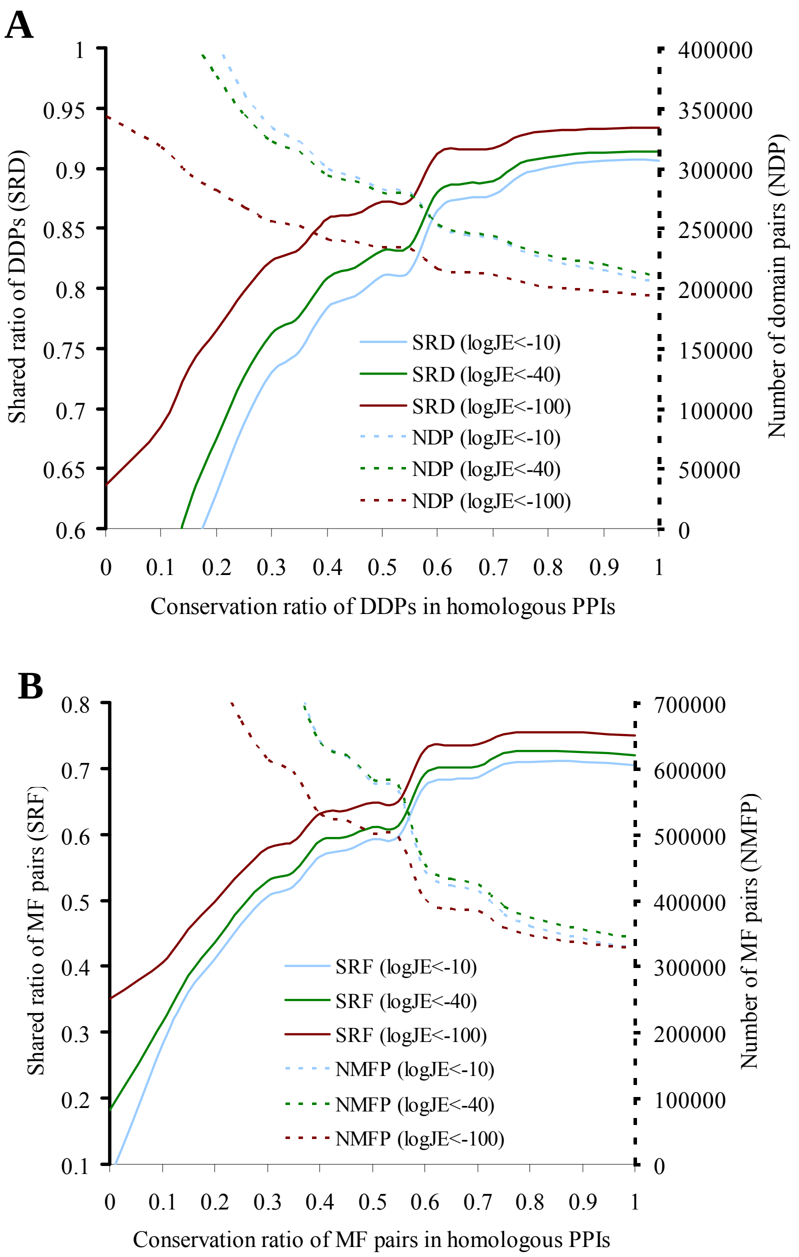

Figure 8. Evaluations of the PPI family. (A) The relationships between conservation ratios of DDPs with shared ratios of DDPs and with the number (dotted lines) of DDPs derived from 103,762 PPI families. The shared ratio of DDPs is 0.88 and the number of DDPs is 252,728 when the conservation ratio is 0.6 and joint *E*-value is 10^{-40} (green lines). (B) Relationships between conservation ratios of molecular function pairs (MFPs) with shared ratios of MFPs and with the number (dotted lines) of MFPs derived from 106,997 PPI families. The shared ratio of MFPs is 0.69 and the number of MFPs is 454,251 when the conservation ratio is 0.6 and joint *E*-value is 10^{-40} (green lines).

3.3 Example analysis of PPI family

Figures 3C and D show search results using σ1A-adaptin (UniProt accession number: P61967) and γ1-adaptin (P22892) of *Mus musculus* as the query. These two proteins are components of the heterotetrameric adaptor protein complex 1 (AP-1), which medicates clathrin-coated vesicle transport from the *trans*-Golgi network to endosome[33]. According to the crystal structure (PDB code: 1w63)[34], this protein pair is a physical interaction, but it is not recorded in the annotated PPI database. For this query, the PPISearch server identifies 14 homologous PPIs, a PPI family, from four species (human, mouse, fruit fly, and yeast). This PPI family has four DDPs (Figure 3E)—PF01217-PF01602 (*CRD* is 1.0), PF01217-PF02883 (0.93), PF1217-PF02296 (0.14), and PF01217-PF07718 (0.07). Two DDPs (PF01217-PF01602 and PF01217-PF02883) with highest *CRD* ratios are the domain compositions of the query and PF01217-PF01602 is the interacting domains[34].

This server allows users to choose the J_E threshold of homologous PPIs. For example, when J_E is set to 10^{-100} (default value is 10^{-40}), the number of homologous PPIs decreases from 14 to 10 by filtering out the last four PPIs (Figure 3D). These 10 homologous PPIs consistently include the two DDPs PF01217-PF01602 and PF01217-PF02883, each with a *CRD*=1.0. Furthermore, users can choose the best match or number of homologous PPIs in a species. In this manner, the PPISearch server is able to select the primary homologous PPIs of each species for specific applications, such as evolutionary analysis of essential proteins.

3.4 Limitations and biases of PPI family

There were some families which have numerous homologous PPIs from particular organisms but few homologous PPIs from some organisms, such as the family of transforming growth factor-beta3 (TGFB3) and TGF-beta type II receptor (TGFBR2). In these case studies, we found that the limitations and biases may be caused by the annotated database and the method of searching homologs by local alignment using BLASTP. In this section, we proposed some real cases to confirm these problems and the improvements for these limitations and biases.

3.4.1 Limitations of the annotated database for identifying PPI family

In this study, we identified the PPI family through sequence similarity from the annotated database. Recently, there are rapidly increasing number of PPIs which are identified and characterized by many high-throughput experimental approaches. Therefore, PPIs in the annotated database collected from five public databases (IntAct, BioGRID, DIP, MIPS, and MINT) have many PPIs derived from large-scale PPI identification. We compared the number of proteins in Integr8 database (6,352,363 protein sequences in 2,274 species), which integrated information about deciphered genomes and their corresponding proteomes, and the number of PPIs in the annotated database. Table 1 presents the number of PPIs and proteins in organisms that are commonly used in molecular researches. There are 56,006 proteins (24.19 % in 11 species) of *Homo sapiens* but only 45,963 PPIs (20.39 %). In contrast, 5,727 proteins (2.47 %) and 113,161 PPIs (50.42 %) of *Saccharomyces cerevisiae* are recorded in Integr8 database and the annotated database, respectively. These results indicated that the annotated database was biased by PPIs of *Saccharomyces cerevisiae*. Conversely, the number of PPIs from *Homo sapiens* in the annotated database is underestimated. Similarly, there are few PPIs in *Danio rerio*, *Bos Taurus*, and *Gallus gallus*. Based on these results, we found that the number of PPIs is underestimated in some organisms but overestimated in some organisms since we have merged the common used PPI databases. It meant the annotated PPI database is dominated by few species. On the other hand, there are more human PPIs than currently known from high-throughput and other experimental evidence[35].

For improving this limitation, we used the Integr8 database as the new protein sequence database instead of original protein sequence database derived from the annotated database. Integr8, which is composed of complete genomes, reflects the genomic distribution of proteins in the nature but is not biased by few species.

3.4.2 Biases of local alignment by BLASTP in searching homologs

We used the tool of BLASTP that searching similar proteins by local alignment to identify homologs through sequence similarity (*E*-value ≤ 10^{-10}). The PPI family of human transforming growth factor-beta3 (TGFB3) and TGF-beta type II receptor (TGFBR2), which is searching from the annotated database, consists of 23 homologous PPIs in human and only one homologous PPI in muse (Table 2). Otherwise, the SB-PPI family that using the template (PDB code: 1ktz[36]; TGFB3-TGFBR2) identified 4 and 2 homologous PPIs in human and mouse, respectively, among 28 SB-HomPPIs of six organisms from the Integr8 database (Table 3). More PPIs and organisms were included in this SB-PPI family than the PPI family. Additionally, Figure 9 shows the conservation of domain-domain pairs between PPI family and SB-PPI family in *homo sapiens* and *Mus musculus*, and all of domain annotations are assigned from Pfam database. The SB-PPI family emphasizes the interface of the template 1ktz has more conserved domain-domain pairs than the PPI family. According to the structural interface of 1ktz, TGFB3 binds with TGFBR2 through the TGF_beta domain and ecTbetaR2 domain. The TGF_beta domain and ecTbetaR2 domain are also considered as the interacting domain pair in the iPfam database. However, among these 24 homologous PPIs, there are only three homologous PPIs that also have the interacting domain pair in human (Table 2). The Activin_recp domain instead of the ecTbetaR2 domain in other homologous PPIs implied that these 21 homologous PPIs may not belong to this PPI family. We observed that sequence alignments of these homologous PPIs emphasize on the region of the PKinase domain which is the larger common domain of protein kinase (Figure 9).

Table 2. The PPI family search results [using the protein sequence pair of human TGFB3 (P10600) and TGFBR2 (P37173) as the query]

Protein 1				Protein 2				Species	Joint *E*-value	Rank
UniProt AC	Gene name	*E*-value	Domain	UniProt AC	Gene name	*E*-value	Domain			
P10600	TGFB3	0.0	TGFb_propeptide TGF_beta	P37173	TGFBR2	0.0	ecTbetaR2 Pkinase	*Homo sapiens*	0.0	1
P61812	TGFB2	1e-129	TGFb_propeptide TGF_beta	P37173	TGFBR2	0.0	ecTbetaR2 Pkinase	*Homo sapiens*	3.2e-155	2
P01137	TGFB1	1e-82	TGFb_propeptide TGF_beta	P37173	TGFBR2	0.0	ecTbetaR2 Pkinase	*Homo sapiens*	1.0e-131	3
P10600	TGFB3	0.0	TGFb_propeptide TGF_beta	P36897	TGFBR1	9e-63	Activin_recp TGF_beta_GS Pkinase	*Homo sapiens*	9.5e-122	4
P10600	TGFB3	0.0	TGFb_propeptide TGF_beta	P37023	ACVRL1	7e-52	Activin_recp TGF_beta_GS Pkinase	*Homo sapiens*	2.6e-116	5
P61812	TGFB2	1e-129	TGFb_propeptide TGF_beta	P36897	TGFBR1	9e-63	Activin_recp TGF_beta_GS Pkinase	*Homo sapiens*	3.0e-96	6
P01137	TGFB1	1e-82	TGFb_propeptide TGF_beta	P36897	TGFBR1	9e-63	Activin_recp TGF_beta_GS Pkinase	*Homo sapiens*	9.5e-73	7
P01137	TGFB1	1e-82	TGFb_propeptide TGF_beta	P37023	ACVRL1	7e-52	Activin_recp TGF_beta_GS Pkinase	*Homo sapiens*	2.6e-67	8
O14793	MSTN	2e-27	TGFb_propeptide TGF_beta	Q13705	ACVR2B	3e-76	Activin_recp Pkinase	*Homo sapiens*	7.7e-52	9
P18075	BMP7	2e-21	TGFb_propeptide TGF_beta	Q13705	ACVR2B	3e-76	Activin_recp Pkinase	*Homo sapiens*	7.7e-49	10
P18075	BMP7	2e-21	TGFb_propeptide TGF_beta	P27037	ACVR2A	3e-76	Activin_recp Pkinase	*Homo sapiens*	7.7e-49	11
O95390	GDF11	4e-21	TGFb_propeptide TGF_beta	Q13705	ACVR2B	3e-76	Activin_recp Pkinase	*Homo sapiens*	1.1e-48	12
P22004	BMP6	2e-20	TGFb_propeptide TGF_beta	Q13705	ACVR2B	3e-76	Activin_recp Pkinase	*Homo sapiens*	2.4e-48	13
P22004	BMP6	2e-20	TGFb_propeptide TGF_beta	P27037	ACVR2A	3e-76	Activin_recp Pkinase	*Homo sapiens*	2.4e-48	14
P08476	INHBA	2e-19	TGFb_propeptide TGF_beta	Q13705	ACVR2B	3e-76	Activin_recp Pkinase	*Homo sapiens*	7.7e-48	15
P08476	INHBA	2e-19	TGFb_propeptide TGF_beta	P27037	ACVR2A	3e-76	Activin_recp Pkinase	*Homo sapiens*	7.7e-48	16
P12643	BMP2	4e-18	TGFb_propeptide TGF_beta	P27037	ACVR2A	3e-76	Activin_recp Pkinase	*Homo sapiens*	3.5e-47	17

Table 2. The PPI family search results [using the protein sequence pair of human TGFB3 (P10600) and TGFBR2 (P37173) as the query] (Continued)

Protein 1				Protein 2				Species	Joint *E*-value	Rank
UniProt AC	Gene name	*E*-value	Domain [a]	UniProt AC	Gene name	*E*-value	Domain			
P09529	INHBB	5e-13	TGFb_propeptide TGF_beta	Q13705	ACVR2B	3e-76	Activin_recp Pkinase	*Homo sapiens*	1.2e-44	18
P09529	INHBB	5e-13	TGFb_propeptide TGF_beta	P27037	ACVR2A	3e-76	Activin_recp Pkinase	*Homo sapiens*	1.2e-44	19
P43026	GDF5	3e-12	TGFb_propeptide TGF_beta	Q13705	ACVR2B	3e-76	Activin_recp Pkinase	*Homo sapiens*	3.0e-44	20
P43026	GDF5	3e-12	TGFb_propeptide TGF_beta	P27037	ACVR2A	3e-76	Activin_recp Pkinase	*Homo sapiens*	3.0e-44	21
P18075	BMP7	2e-21	TGFb_propeptide TGF_beta	Q13873	BMPR2	3e-61	Activin_recp Pkinase	*Homo sapiens*	2.4e-41	22
P22004	BMP6	2e-20	TGFb_propeptide TGF_beta	Q13873	BMPR2	3e-61	Activin_recp Pkinase	*Homo sapiens*	7.7e-41	23
Q9Z1W4	Gdf11	4e-21	TGFb_propeptide TGF_beta	P27040	Acvr2b	2e-76	Activin_recp Pkinase	*Mus musculus*	8.9e-49	1

PPI family consists of 24 homologous PPIs of human transforming growth factor and its receptor, including transforming growth factor-beta3 (P10600 with interacting domain TGF_beta) and TGF-beta type II receptor (P37173 with interacting domain ecTbetaR2) searching on Integr8 database. The threshold of the Joint *E*-value is set to 10^{-40}.

[a] Domain annotations of protein are assigned from Pfam database.

Table 3. The SB-PPI family search results (using the structural template 1ktz of human TGFB3 and TGFBR2 heterodimer as the query)

Protein 1			Protein 2								
UniProt AC	Gene name	Domain[a]	UniProt AC	Gene name	Domain	Species	*Z*-score	P / N[b]	Rank	RSS of BP[c]	RSS of CC[d]
P17246	Tgfb1	TGFb_propeptide TGF_beta	P38438	Tgfbr2	ecTbetaR2 Pkinase	*Rattus norvegicus*	3.702	-	1	1.00	0.75
Q07258	Tgfb3	TGFb_propeptide TGF_beta	P38438	Tgfbr2	ecTbetaR2 Pkinase	*Rattus norvegicus*	3.702	-	2	0.86	0.75
P04202	Tgfb1	TGFb_propeptide TGF_beta	Q62312	Tgfbr2	ecTbetaR2 Pkinase	*Mus musculus*	3.124	-	1	0.90	0.00
Q3UFN7	Tgfb3	TGFb_propeptide TGF_beta	Q62312	Tgfbr2	ecTbetaR2 Pkinase	*Mus musculus*	3.124	-	2	0.84	0.66
P10600	TGFB3	TGFb_propeptide TGF_beta	P37173	TGFBR2	ecTbetaR2 Pkinase	*Homo sapiens*	3.723	P	1	0.89	0.46
P01137	TGFB1	TGFb_propeptide TGF_beta	P37173	TGFBR2	ecTbetaR2 Pkinase	*Homo sapiens*	3.723	P	2	0.89	0.48
P61812	TGFB2	TGFb_propeptide TGF_beta	P37173	TGFBR2	ecTbetaR2 Pkinase	*Homo sapiens*	3.137	P	3	0.92	0.55
Q9NR23	GDF3	TGF_beta	P37173	TGFBR2	ecTbetaR2 Pkinase	*Homo sapiens*	3.097	-	4	-[e]	0.00
Q66I23	tgfb3	TGFb_propeptide TGF_beta	B0S5M2	ORF=DKEY-101K6.5-001	ecTbetaR2 Pkinase	*Danio rerio*	4.034	-	1	0.37	-
Q7ZZU7	tgfb1	TGFb_propeptide TGF_beta	B0S5M2	ORF=DKEY-101K6.5-001	ecTbetaR2 Pkinase	*Danio rerio*	4.016	-	2	0.37	-
Q66I23	tgfb3	TGFb_propeptide TGF_beta	Q58EQ1	tgfbr2	ecTbetaR2 Pkinase	*Danio rerio*	3.753	-	3	0.37	-
Q7ZZU7	tgfb1	TGFb_propeptide TGF_beta	Q58EQ1	tgfbr2	ecTbetaR2 Pkinase	*Danio rerio*	3.735	-	4	0.37	-
Q9W6I7	lft2	TGFb_propeptide TGF_beta	Q58EQ1	tgfbr2	ecTbetaR2 Pkinase	*Danio rerio*	3.614	-	5	0.37	-
Q7SZV4	tgfb2	TGFb_propeptide TGF_beta	B0S5M2	ORF=DKEY-101K6.5-001	ecTbetaR2 Pkinase	*Danio rerio*	3.338	-	6	0.37	-
Q9W6I7	lft2	TGFb_propeptide TGF_beta	B0S5M2	ORF=DKEY-101K6.5-001	ecTbetaR2 Pkinase	*Danio rerio*	3.205	-	7	0.37	-
Q7SZV4	tgfb2	TGFb_propeptide TGF_beta	Q58EQ1	tgfbr2	ecTbetaR2 Pkinase	*Danio rerio*	3.112	-	8	0.37	-
P09531	TGFB1	TGFb_propeptide TGF_beta	Q90999	TGFBR2	ecTbetaR2 Pkinase	*Gallus gallus*	3.639	-	1	0.53	0.00
P30371	TGFB2	TGFb_propeptide TGF_beta	Q90999	TGFBR2	ecTbetaR2 Pkinase	*Gallus gallus*	3.039	-	2	0.53	0.55

Table 3. The SB-PPI family search results (using the structural template 1ktz of human TGFB3 and TGFBR2 heterodimer as the query) (Continued).

Protein 1			Protein 2								
Uniprot AC	Gene name	Domain[a]	Uniprot AC	Gene name	Domain	Species	Z score	P / N[b]	Rank	RSS of BP[c]	RSS of CC[d]
Q4RLV8	ORF=GSTENG00032323001	TGFb_propeptide TGF_beta	Q4SUT7	ORF=GSTENG00012314001	ecTbetaR2 Pkinase	*Tetraodon nigroviridis*	3.758	-	1	0.37	-
Q4RFT1	ORF=GSTENG00035186001	TGFb_propeptide TGF_beta	Q4SUT7	ORF=GSTENG00012314001	ecTbetaR2 Pkinase	*Tetraodon nigroviridis*	3.739	-	2	0.37	-
Q4RPB2	ORF=GSTENG00031190001	TGFb_propeptide TGF_beta	Q4SUT7	ORF=GSTENG00012314001	ecTbetaR2 Pkinase	*Tetraodon nigroviridis*	3.735	-	3	0.37	-
Q4RR49	ORF=GSTENG00030322001	TGFb_propeptide TGF_beta	Q4SUT7	ORF=GSTENG00012314001	ecTbetaR2 Pkinase	*Tetraodon nigroviridis*	3.567	-	4	0.37	-
Q4RLV8	ORF=GSTENG00032323001	TGFb_propeptide TGF_beta	Q4S0V9	ORF=GSTENG00025846001	ecTbetaR2 Pkinase	*Tetraodon nigroviridis*	3.496	-	5	0.37	-
Q4RFT1	ORF=GSTENG00035186001	TGFb_propeptide TGF_beta	Q4S0V9	ORF=GSTENG00025846001	ecTbetaR2 Pkinase	*Tetraodon nigroviridis*	3.477	-	6	0.37	-
Q4RR49	ORF=GSTENG00030322001	TGFb_propeptide TGF_beta	Q4S0V9	ORF=GSTENG00025846001	ecTbetaR2 Pkinase	*Tetraodon nigroviridis*	3.456	-	7	0.37	-
Q4RPB2	ORF=GSTENG00031190001	TGFb_propeptide TGF_beta	Q4S0V9	ORF=GSTENG00025846001	ecTbetaR2 Pkinase	*Tetraodon nigroviridis*	3.363	-	8	0.37	-
Q4SBM1	ORF=GSTENG00020909001	TGF_beta	Q4SUT7	ORF=GSTENG00012314001	ecTbetaR2 Pkinase	*Tetraodon nigroviridis*	3.265	-	9	-	0.00
Q4SVQ4	ORF=GSTENG00011872001	TGFb_propeptide TGF_beta	Q4SUT7	ORF=GSTENG00012314001	ecTbetaR2 Pkinase	*Tetraodon nigroviridis*	3.263	-	10	0.37	-

SB-PPI family consists of 4 positive and 0 negative PPIs of the structural template (PDB code 1ktz), including transforming growth factor-beta3 (chain A with interacting domain TGF_beta) and TGF-beta type II receptor (chain B with interacting domain ecTbetaR2) searching on Integr8 database. The threshold of the Z-score is set to 3.0.

[a] Domain annotations of protein are assigned from Pfam database.

[b] PPI is a positive (P, recorded in database) or negative case (N, RSS scores of BP and CC are less than 0.4).

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

[e] The protein pair don't have Gene Ontology annotation in BP or CC.

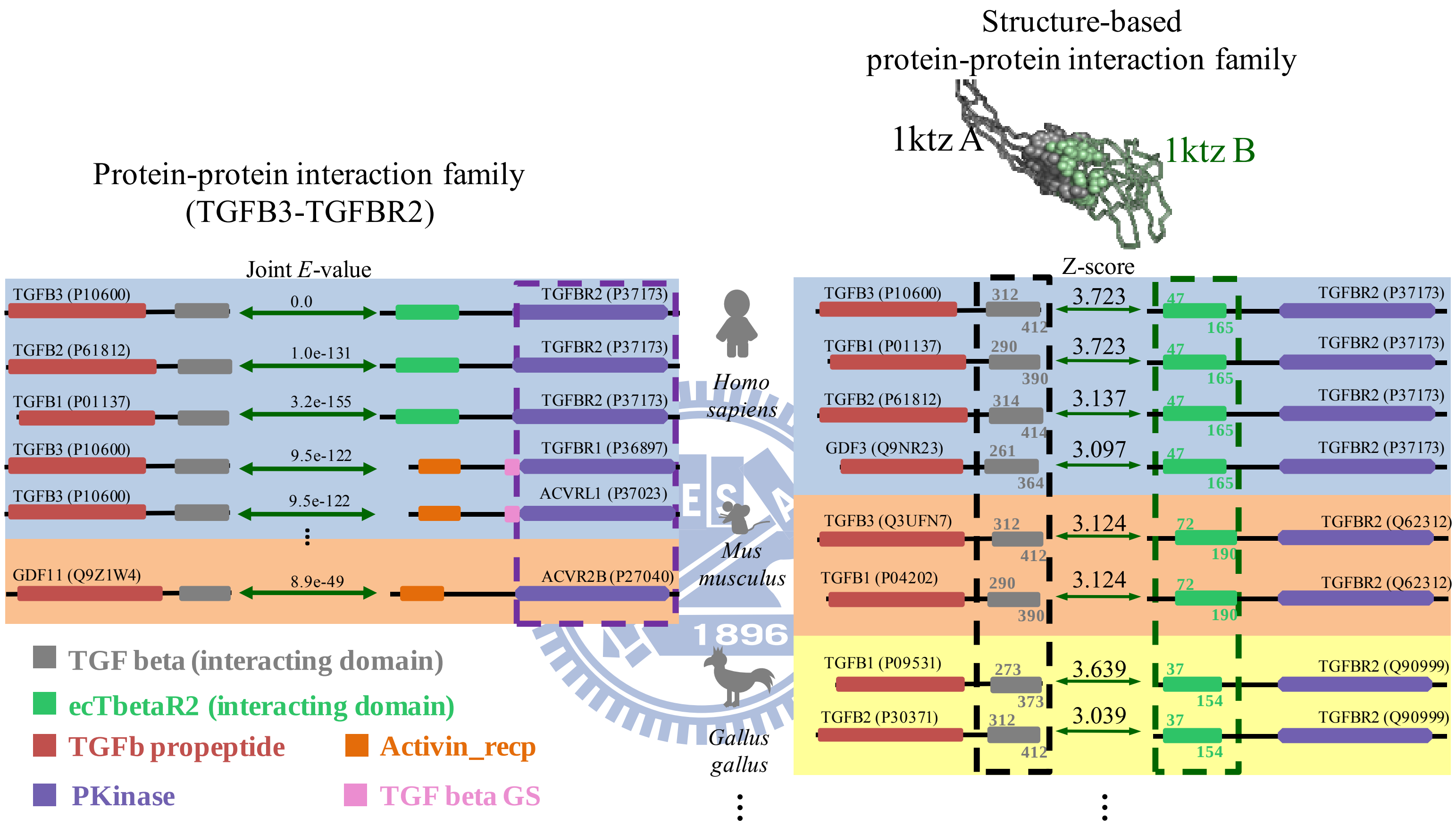

Figure 9. Comparing the PPI family with the SB-PPI family using the PPI of human transforming growth factor-beta3 (TGFB3; P10600) and TGF-beta type II receptor (TGFBR2; P37173) or their 3D heterodimer template (PDB code: 1ktz). The detail information of members in PPI family and SB-PPI family are showed on the Table 1 and 2. The interacting domains (black and green dotted box) are more conserved in SB-PPI family than PPI family. The sequence alignment regions between TGFBR2 and its homologs (purple dotted box) in PPI family are not involved in protein-protein interacting.

This problem often occurs in the PPI family, which have the highly similar and larger domain among multiple domains because of local alignment of BLASTP.

In addition, we also found another case to support our assumption. Growth-regulatory signals strictly regulate the transition from the G1 to the S phases of the cell cycle which is involved in marking an irreversible commitment to DNA synthesis and proliferation[37], such as Rb-E2F pathway. In Rb-E2F pathway, the retinoblastoma (Rb) protein binds to the E2F transcription factors for negatively regulating the G1-S transition. Figure 10 presents the PPI family of the transcription factor E2F1 (E2F1) and retinoblastoma-associated protein (Rb1) which consists of five homologous PPIs in *homo sapiens* and one homologous PPI in *Rattus argentiventer*. We also identified the SB-PPI family, which includes 40 SB-HomPPIs in seven species, using the structural template (2ast) of E2F1 (chain B) and RB1 (chain C). Unfortunately, there are two homologous PPIs in the PPI family but not in the SB-PPI family (*e.g.*, RBL1-E2F1 in *Homo sapiens* and Rb1-E2f1 in *Rattus argentiventer*). Based on the interface between chain B and C of 2ast, the Rb1 binds with E2F1 through the RB_C domain and a region that is not a domain in Pfam database. In this PPI family, the sequence similarity (1e-45) between the homolog (RBL1) in the homologous PPI (RBL1-E2F1) and the query protein (RB1) is due to the DUF3452 domain and part of the RB_A domain but not the RB_C domain. Among 6 conserved contact residues, we found that the sequence alignment of the RB_C domain (contact residue identity is 40%) only conserves one contact residue that is involved in forming hydrogen bond. It meant that RBL1 may bind to E2F1 weakly or even not bind. Moreover, the E2f1 protein (O09139) lacks the region of the interface in another homologous PPI (Rb1-E2f1). This fact implied this PPI may not belong to this family or even be a false PPI because of losing interacting domains of RB1-E2F1. These results could help us to find some potential problems for identifying PPI family.

For addressing these biases, we applied the method "3D-domain interolog mapping" to the SB-PPI family for emphasizing the interface region. Therefore, all SB-HomPPIs of the SB-PPI family using template 1ktz share the same interacting domain pair (TGF_beta-ecTbetaR2). On the other hand, the SB-PPI family could identify SB-HomPPIs from some organisms that were not in the PPI family. The new approach of SB-HomPPI, which is combination of the concept of PPI family and 3D-domain interologs, is presented on the next section.

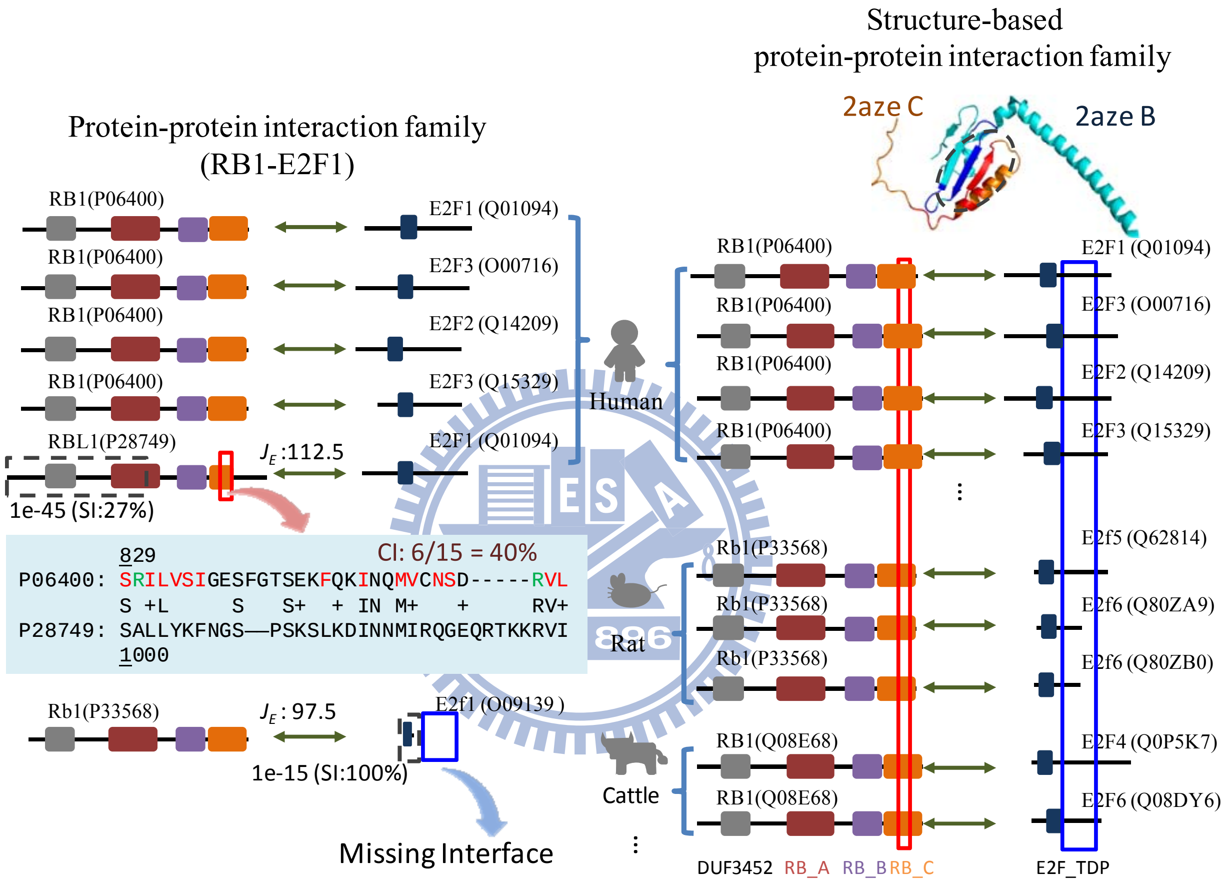

Figure 10. Comparison of searching results between PPI family and SB-PPI family of transcription factor E2F1 (E2F1) and retinoblastoma-associated protein (Rb1). The interface of the structural template (2aze) consists of the RB_C domain of Rb1 (red box) and the region of E2F1 that is not a domain in Pfam database (blue box). In PPI family, the homolog (RBL1) in RBL1-E2F1 interaction is determined through sequence similarity (*E*-value is 10^{-45}) due to the sequence alignment (gray dotted box) by BLASTP, but the region of sequence alignment does not located on the interface. The protein E2f1 of the PPI in rat lacks the region of interface

3.5 The criteria for identifying SB-PPI family

The approach of SB-HomPPI identifies SB-PPI family that utilizes the method of 3D-domain interologs to determine SB-HomPPIs across multiple species. In this study, the criteria and verifications of SB-HomPPIs followed the previous studies in our lab. Precision, recall and F-measure were utilized to decide the threshold (*Z*-score) of the SB-HomPPI for searching homologous complexes. The F-measure is given as (2 × precision × recall) / (precision + recall) where the precision and recall using the standard positive and negative sets. Figure 11 shows the relationships between *Z*-score and recall and precision using NR-563 on the Integr8 database. The recall significantly decreases when *Z*-score ≥ 3; conversely, the precision increases slightly when joint *Z*-score is between 3 and 4. The recall and precision are 0.79 and 0.63, respectively, and the SB-HomPPI yields the highest F-measure value (0.70) if the threshold of *Z*-score is set to 3. We identify homologous families of a 3D-dimer template *T* with significant sequence similarity (PSI-BLASTP *E*-values ≤ 10^{-10}). A SB-PPI family of template *T* is composed of SB-HomPPIs from homologous families when these SB-HomPPIs have interface sequence identity ≥ 25%, contact residue identity ≥ 25%, *Z*-score ≥ 3, and they ranked in the Top 25 in one species. Among NR-1895 set, we identified 1,638 SB-PPI families including of 224,713 SB-HomPPIs in 1,715 species.

1896

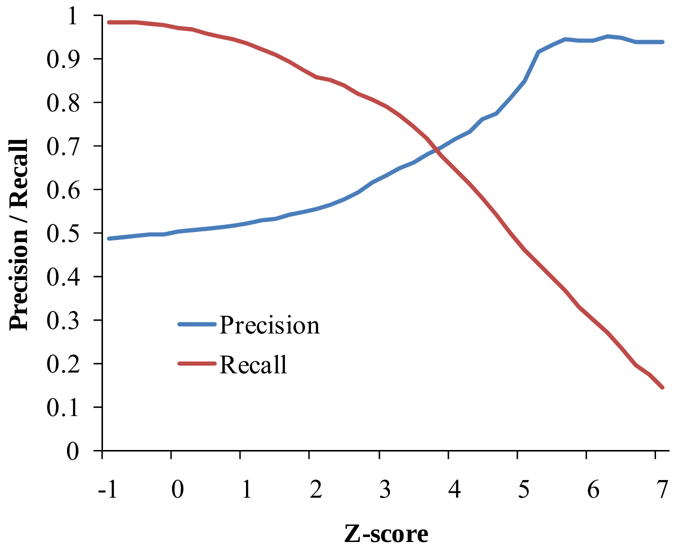

Figure 11. The precision and recall of SB-HomPPI with Z-scores on the data set NR-563 and the Integr8 database. The precision is increasing and the recall is decreasing if the Z-score is increasing. In this study, the threshold of Z-score is set to 3.0 for considering a protein-protein interaction candidate because of the highest F-score (0.70).

3.6 Conservation of interacting domain pairs and GO term pairs in SB-PPI family

Based on 359 heterodimer structures with IDPs annotated in iPfam database, we statistically evaluated IDP conservation of SB-PPI families and the consensus (transferability) of IDPs between these heterodimer structures and their SB-HomPPIs. SB-PPI family is highly conserved on interacting domain pairs due to high average consensus ratio (more than 0.7) in 339 families (94.4%) (Figure 12, black lines). It implied that SB-PPI family could transfer the consensus interacting domain pair(s) to the member for understanding the newly determined PPI, if the member lacked the annotation of interacting domain pair(s).

Additionally, we utilized the MFP transferability (the shared ratio), which is statistically derived from 281 heterodimer structures with MF annotation in GO database, to estimate MFPs shared by the 3D-dimer structures and its homologous PPIs. Figure 13A shows the shared ratio of MFP is more than 0.97 when the conservation ratio ≥ 0.6. Similarly, we also estimated transferability of BPPs and CCPs between 3D-dimer structures and their SB-HomPPIs, and there are 292 and 245 heterodimer structures with BP and CC annotation, respectively (Figure 13B and C). While the conservation ratio are more than 0.6, we found that the shared ratios of BPP and CCP are higher than 0.94. These results indicated SB-PPI family achieves high agreements on GO term pairs.

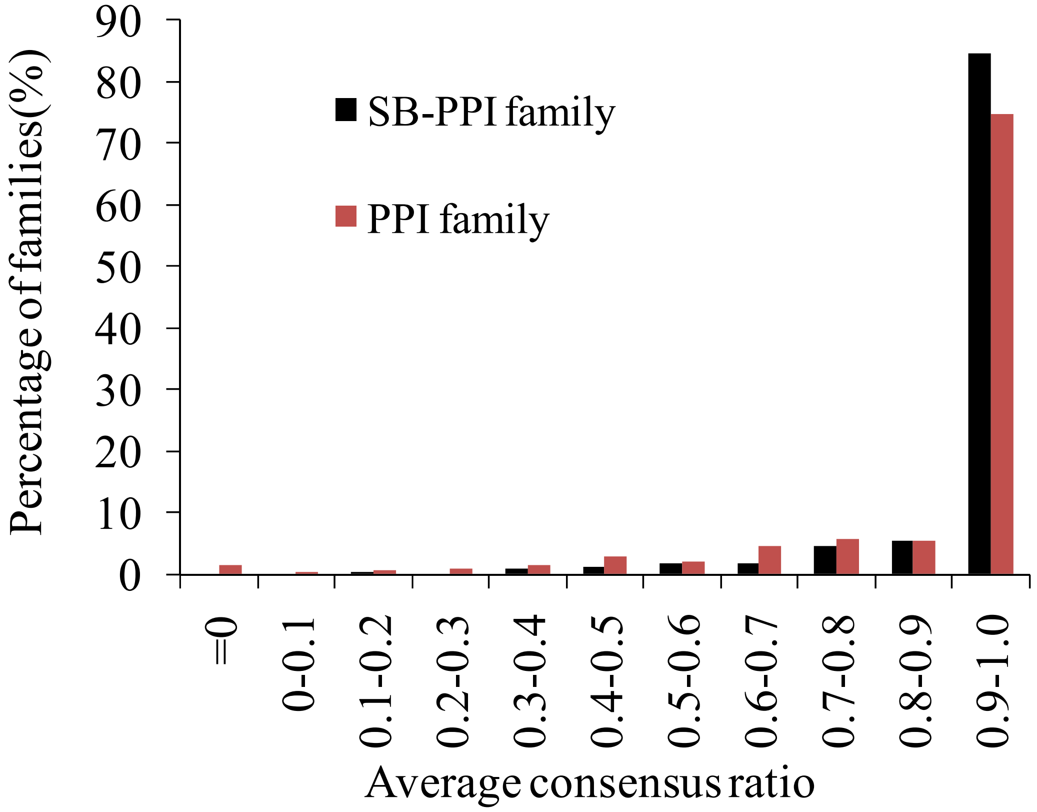

Figure 12. The distributions of the average consensus ratio of interacting domain pairs between 12,053 PPI families and 359 SB-PPI families. The consensus ratios of 304 (84.7%) and 339 (94.4%) families are above 0.9 and 0.7 in SB-PPI families, receptively. Among 12,053 PPI families, there are 9,004 (74.7%) and 10,312 (85.6%) PPI families while their consensus ratios are above 0.9 and 0.7, respectively.

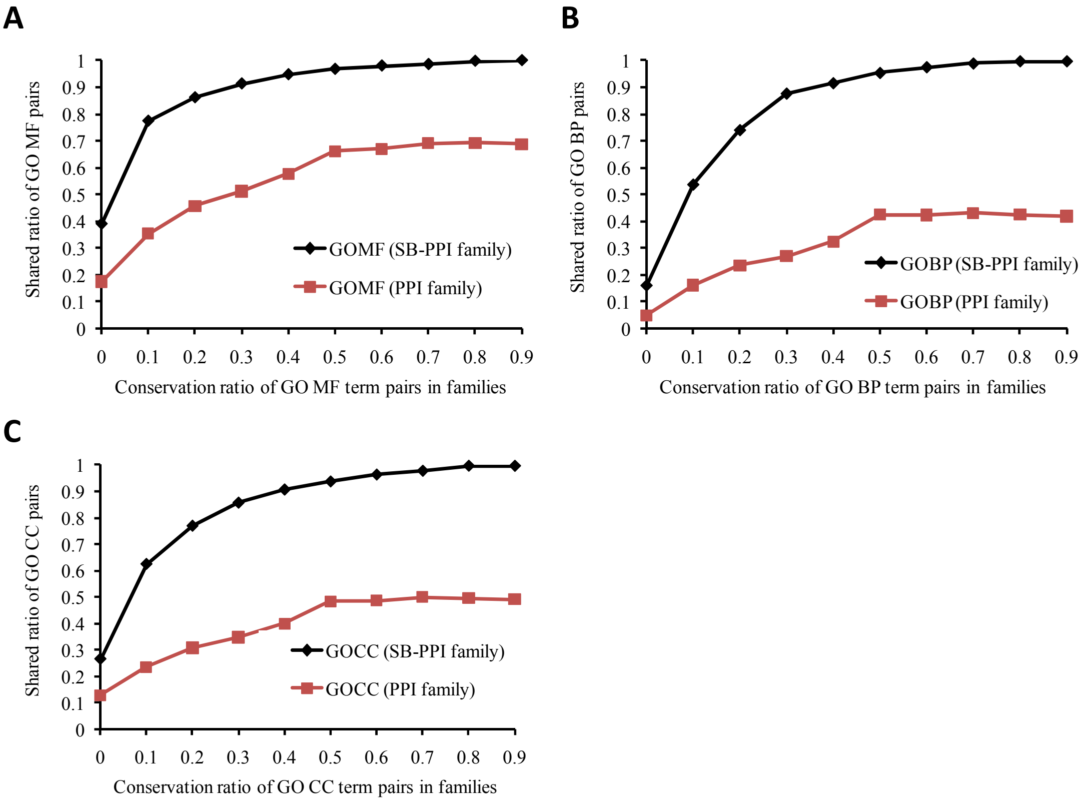

Figure 13. The transferability (shared ratios) and conservation ratios of GO term pairs between SB-PPI families and PPI families. (A) While the conservation ratios are above 0.6, the shared ratios of molecular function pairs (MFPs) are above 0.97 and 0.66 in 281 SB-PPI families and 78,970 PPI families, respectively. (B) The shared ratios of biological process pairs (BPPs) of 292 SB-PPI families and 72,607 PPI families are above 0.95 and 0.42 when their conservation ratios ≥ 0.6. (C) 245 SB-PPI families with the conservation ratios ≥ 0.6 have the shared ratio of CCPs above 0.94, and 67,317 PPI families with the conservation ratios ≥ 0.6 have the shared ratio of cellular component pairs (CCPs) above 0.48.

3.7 Example analysis of SB-PPI family

Figure 5 shows the SB-PPI family searching results of using structure template (PDB entry 1ktz-A and 1ktz-B) sequences as the query, which are human TGFB3 (UniProt accession number P10600) and human TGFBR2 (P37173), respectively. The transforming growth factor β (TGFβ) pathway controls the differentiation, growth, and final fate of metazoan cells[38]. TGFB3 and TGFBR2 are ligand and receptor in this pathway. There are 153 homologs (*E*-value ≤ 10^{-10}), such as TGFB1, TGFB2, and TGFB3, of human TGFB3 found in 14 organisms by using PSI-BLAST scanning on Integr8 database. Likewise, nine homologs (*e.g.*, TGFBR2) of Human TGFBR2 are yielded. We found 28 SB-HomPPIs with significant interface similarity (sequence identity ≥ 25%, contact residue identity ≥ 25%, and Z-value ≥ 3) in six organisms, including *Homo sapiens*, *Mus musculus*, and *Gallus gallus* (Figure 5 and Table 2). There are 3 interactions recorded in an annotated PPIs database, TGFB3, TGFB2, and TGFB1 with TGFB2 in human. Nine PPIs have highly relative gene names in different organisms. Moreover, GDF3 in *homo sapiens* and lft2 in *Danio rerio* are belong to TGFβ signaling molecules[39-40]. Among these 28 interactions, all of them use TGF_beta domain and ecTbetaR2 domain as interacting domain pair that is recorded in iPfam database.

A compact heterodimers and an interacting surface, which is lined with hydrophobic and hydrophilic residues, formed by the protein TGFB2 interacts with the protein TGFBR2 (Figure 5B and D). In TGFBR2, two residues Glu119 and Asp32 are involved in the formation of hydrogen-bonded ion pairs (red dotted line) with positively charged amino acids, Arg25 and Arg 94, respectively[36]. Figure 5D show the interacting evolution analysis built by 8 interactions reveals that three residues (Leu27, Thr51 and Ile53) in TGFBR2 and two residues (Trp32 and Tyr90) in TGFB3 are conserved in these species. These residues are involved in forming the convex hydrophobic ridge and the convex hydrophobic ridge[36].

3.8 Application of verify the large-scale PPIs

In **Section 3.4.1**, the annotated PPI database, including many PPIs that are identified and characterized by high-throughput experimental approaches, is dominated by few species, especially yeast. Following above results, we found that the PPI family and SB-PPI family are conserved in interacting domain pairs and molecular function pairs. In this section, we applied the conservation of the interacting domain pairs and GO term pairs in SB-PPI family and PPI family to verify the large-scale PPIs. Moreover, there are few PPIs in some species, such as *Danio rerio*, *Bos Taurus*, and *Gallus gallus*. For recruiting the PPIs in these species, we assessed the ability for predicting PPIs across species by using SB-PPI family and PPI family.

3.8.1 Conservation of interacting domain pair

The PPI family has highly conservation of domain-domain pairs (see **Section 3.2**). It also implied the consensus of interacting domain pairs. Domains are assumed to mediate more stable interactions and assemblies of proteins into complexes[41]. It is well known that domains often exhibit evolutionary conservation in sequence and three-dimensional structure.

Therefore, we might be expected that the same domain pairs mediate PPIs in different organisms[42]. For verifying the large-scale PPIs through the conservation of interacting domain pair between PPI family and SB-PPI family, we annotated the domain and interacting domain pairs by Pfam and iPfam databases, respectively. There are 359 SB-PPI families and 12,053 PPI families, which have the annotations of interacting domain pair in the iPfam database. Figure 12 presents the percentage of the number of families in each interval of average consensus ratio. The consensus ratios of 304 (84.7%) and 339 (94.4%) SB-PPI families are more than 0.9 and 0.7, receptively. These ratios are higher than that of PPI family (74.7% and 85.6%). It meant that the members of SB-PPI family through the binding model prefer to share the same interacting domain pairs with the structural template. There are some reasons may cause this difference, such as bias of local alignment by BLASTP in searching protein homologs (see **Section 3.4.2**). However, the SB-PPI family using the structural interface to build the binding model could emphasize on the interacting domain pairs to obtain a 10% increase in family numbers when the consensus ratios are more than 0.9. It meant that SB-PPI family achieves higher agreement on conservation of interacting domain pairs than PPI family. Our preliminary result demonstrated that interacting domain pairs are the important component in PPI family and SB-PPI family. Therefore, we could evaluate the reliability of the large-scale PPIs through the conservation of interacting domain pairs.

3.8.2 Conservation of GO term pairs

In our previous analysis of PPI family, the members of a PPI family have similar molecular functions, and *SRF* ratios are highly correlated with conservation ratios (*CRF*) of MFPs (see **Section 3.2**). We compared the conservation of not only the GO molecular function pairs (MFP) but also biological process pairs (BPP) and cellular component pairs (CCP) in that homologous PPIs may be involved in the same pathway and locate in the same cellular component. Figure 13 shows the comparison of transferability for GO term pairs between PPI family and SB-PPI family. In above results, PPI family achieved high agreement (about 0.69) on the transferability of MFP while the conservation ratios are more than 0.6. The result of SB-PPI family is better than that of PPI family (Figure 13A). The shared ratio achieves 0.97 when the conservation ratios ≥ 0.6. The possibility of the template in a family having a MFP is about 97% while more than 60 % homologous PPIs of the family includes this MFP. Furthermore, higher differences of shared ratios about 0.53 and 0.45 appear in the conservations of BPP and CCP (the conservation ratios ≥ 0.6), respectively, between PPI family and SB-PPI family (Figure 13B and C). The SB-PPI family is considered as more conserved than PPI family in GO term pairs. The preliminary results implied that GO term pairs, especially molecular function, are important components of family. Therefore, we could evaluate the conservation of GO term pairs to verify the reliability of large-scale PPIs.

3.8.3 The performance of predicting PPIs using approaches of identifying PPI family and SB-PPI family

In this section, we applied two methods of PPI family and SB-PPI family to predict protein-protein interactions across multiple species. Among NR-563 set, there are 437 heterodimers that two protein sequences are recorded in UniProt database. For the purpose of predicting PPIs across multiple species, the Integr8 database (6,352,363 protein sequences in 2,274 species) is considered as the searching database and the positive and negative set were described in **Section 2.8**. Here, we evaluated the precision between PPI family and SB-PPI family at each number of true positives (Figure 14). Under the condition of finding the same number of true positives, we found that all precisions of SB-PPI family between Z-score ≥ -100 and Z-score ≤ 100 are higher than precisions of PPI family between $J_E \geq 10^{-10}$ and $J_E \leq 10^{-180}$. The largest difference of precisions between 0.61 in PPI family and 0.93 in SB-PPI family was occurred while two methods found about 690 true positives. It implied two methods could be used to predicting PPIs and the method of SB-PPI family is better than that of PPI family.

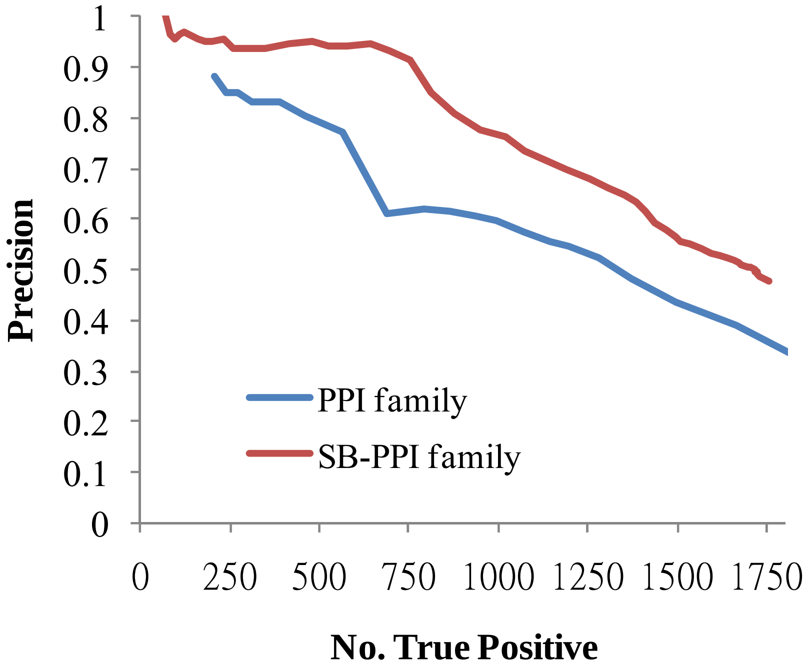

Figure 14. The comparison of precisions between two approaches of PPI family and SB-PPI family. The method of SB-PPI family has higher precision than PPI family in any number of true positives.

1896

3.9 Application of identifying protein complex family

Protein complexes are considered fundamental units of macromolecular organization and their composition is also known to vary due to cellular requirements[32]. Previous studies used template-based methods (*i.e.*, comparative modeling[19] and fold recognition[20]) to model a large set of yeast complexes by searching a 3D-complex library[43-44]. They often modeled one organism to another but not across multiple organisms. In the above section, we have verified PPI family and SB-PPI family and obtained high agreements on interacting domain pairs and GO term pairs. Here, we applied the concept of PPI family and the approach of SB-HomPPI to identify the protein complex family, especially muti-chain complex and membrane protein complex that are difficult to crystalize. Figure 15 shows the protein complex family of succinate ubiquinone oxidoreductase (PDB code: 1yq3)[45]. Mitochondrial respiratory Complex II (succinate ubiquinone oxidoreductase, E.C. 1.3.5.1) is involved in Krebs tricarboxylic acid cycle (oxidizing succinate to fumarate) and plays a role in an entry point for electrons into the respiratory chain at the level of ubiquinol[45]. It includes a large flavoprotein subunit containing covalently bound FAD (1yq3 chain A), an iron-sulfur protein (IP) (chain B), and two small membrane anchor subunits (chains C and D) ligating a single low spin heme of type B. This template consists of four interfaces, Chain A-B, Chain B-D, Chain B-C, and Chain C-D. These interfaces could build four SB-PPI families. According to these four SB-PPI families, we could identify a protein complex family through combing four SB-PPI families. We found five homologous complexes in *Homo sapiens*, *Mus musculus*, *Bos taurus*, and *Ustilago mayds 521*. Interestingly, two homologous complexes lacking gene names in UniProt database could be annotated gene names through other homologous complexes. Moreover, the protein of chains C in *Gallus gallus* was not recorded in UniProt and NCBI databases, but we could transfer the gene name and other descriptions through the protein complex family. It implied that the protein complex may have conservation of evolution as the PPI family and SB-PPI family. I am involving in this study with Yu-Shu Lo in our lab now.

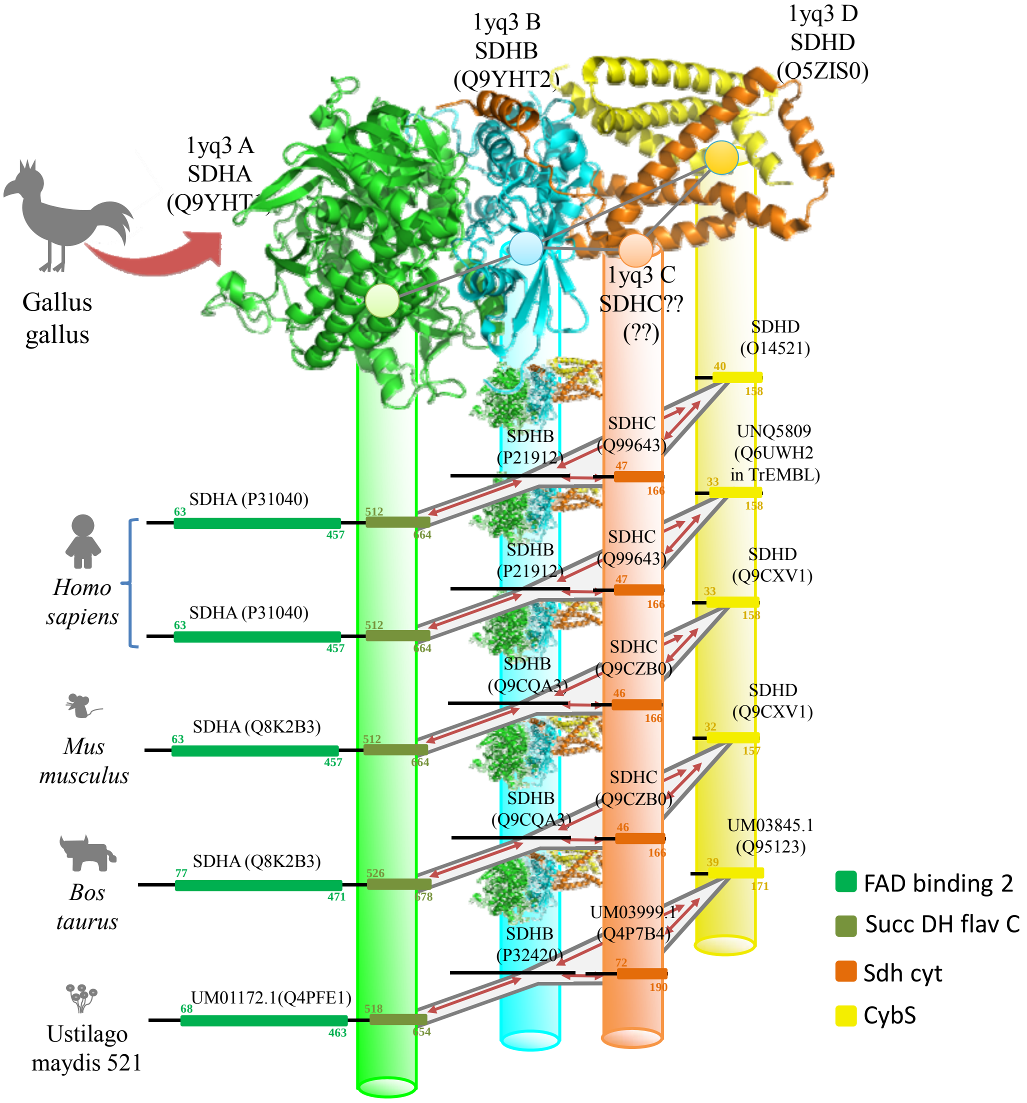

Figure 15. The protein complex family of succinate ubiquinone oxidoreductase (PDB code: 1yq3). There are five homologous protein complexes in *Homo sapiens*, *Mus musculus*, *Bos taurus*, and *Ustilago mayds 521*.

Chapter 4 Conclusion

4.1 Summary

We have developed a new concept of "PPI family" to identify homologous protein-protein interactions as a family in sequence level. Moreover, we also combine the concept of "PPI family" and the method of "3D-domain interolog mapping" to a new approach of SB-HomPPI which finding SB-HomPPIs (a SB-PPI family). We compared conservations of interacting domain pairs and GO term pairs between two approaches to evaluate the transferability. These results implied that PPISearch can broadly build the family by only using the protein sequence. However, the approach of SB-HomPPI can identify SB-HomPPIs to construct higher reliable SB-PPI families than PPISearch. Two approaches may be complementary for classifying PPIs to families. In this study, we get some critical conclusions as follows:

1. PPISearch is the first public server identifying PPI family across multiple species while users input a pair of protein sequences. We identified 90,715 PPI families excluding two proteins of the PPI from different species, and these families consist of 93,406 PPIs in 174 species.
2. PPI family achieves high agreements on interacting domain pairs and molecular function pairs between query protein pairs and their corresponding homologous PPIs.
3. The annotated database, which is dominated by few species, has limitations and biases for identifying PPI family. The method of searching homologs by local alignment using BLASTP has a bias in favor of the large domain when the domain is highly similar with the query but not involving in protein-protein interacting.
4. By integrating the method of "3D-domain interolog mapping" with "PPI family", SB-HomPPI identifies SB-PPI family across multiple complete genomes through using the interface of heterodimer structure as the template. Among NR-1895 query set, we identified 1,638 SB-PPI families including of 224,713 SB-HomPPIs in 1,715 species.
5. In SB-PPI family, which uses Integr8 as searching sequence database, the SB-HomPPI approach emphasizing the interface could avoid the limitation of the annotated database and the bias of searching homologs by local alignment using BLASTP.
6. Comparing the conservations of interacting domain pairs and GO term pairs, SB-PPI family is more conserved than PPI family.
7. SB-PPI family has much higher predicting precision than PPI family which only considers

the protein sequence.

8. The concept of PPI family could apply to verify and recruit the PPI database and construct the protein complex family.

4.2 Major contributions and future works

Recently, there are rapidly increasing number of reliable PPIs which are recorded in many databases, such as IntAct, DIP, MIPS, MINT, and BioGRID. We identified the PPI family which consists of a group of homologous PPIs to provide a description of the functional and evolutionary relationships of PPIs. The new PPI family, which combines the PPI families and SB-PPI families, may be more and more complete and reliable while more reliable PPIs and 3D-structure are identified. Additionally, The PPI family could help biologists to study the comparison of biochemical networks across multiple species. The cross-species network comparison can be used to identify the corresponding pathways from one organism to another. For systems biology, the comparison of networks provides clues for understanding some important issue, such as evolution of networks.

Some important issues will be discussed in the future. The PPI family could combine with the SB-PPI family for verifying the reliability of the members in a family in older to building the reliable PPI family. Moreover, we also want to know whether this reliable PPI family may have hierarchical relationship as the protein family (*e.g.*, SCOP). Figure 16 shows the hypothesis of the hierarchical relationship in PPI families. A PPI superfamily may consist of more than one PPI family and some PPI families further include many core families. This relationship implies that the evolution may be involved in a protein interacting with another protein. Our major assumption, which considered that a group of homologous PPIs may be from a common ancestry, could be further proved when co-evolution occurs in homologous PPIs of the reliable PPI family.

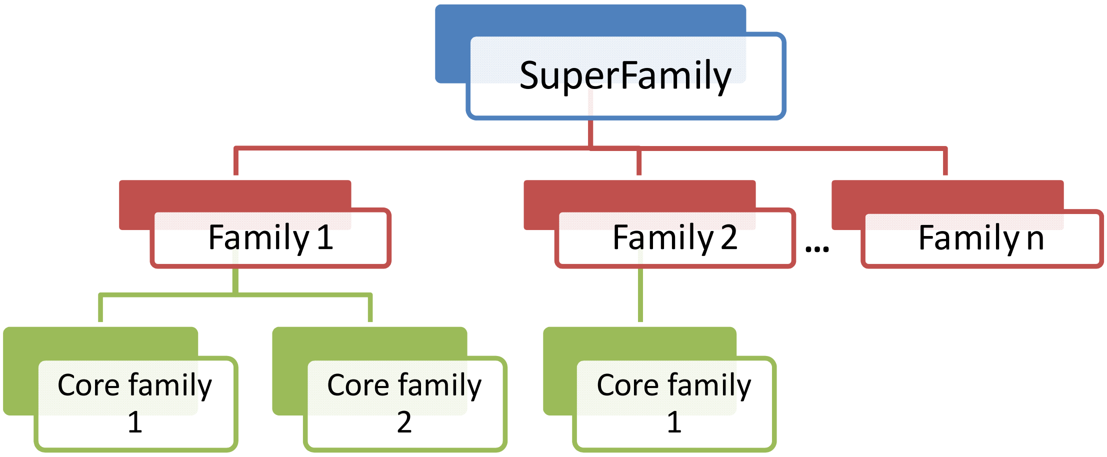

Figure 16. The hypothesis of the hierarchical relationship between PPI families.

Based on the methods of PPI family and SB-PPI family, we also could verify the PPI in a family, especially these PPIs from large-scale experiments. Each PPI of a family may be scored through a new scoring function in the future. This scoring function may be composed of several scores. For example, the reliability score may consider the existence of crystal structure and the PPI identified by small-scale experiments or many different large-scale experiments. The interface (or interacting domain pair) is highly conserved and important for protein-protein interacting, and the similarity of interface could provide a clue to evaluate a PPI. In addition, the conserved biological properties, such as GO terms, also could be considered biological meaning score according to their highly conservation in our studies. It is useful to construct the reliable PPI family and offer biologists to realize evolutions of homologous PPIs and PPI families.

References

1 Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Research* **38**, D211-222 (2010).

2 Wu, C. H. *et al.* PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Research* **32**, D112-114 (2004).

3 Lo Conte, L. *et al.* SCOP: a structural classification of proteins database. *Nucleic Acids Research* **28**, 257-259 (2000).

4 Greene, L. H. *et al.* The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Research* **35**, D291-297 (2007).

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

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

7 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 Sciences of the United States of America* **96**, 4285-4288 (1999).

8 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 Research* **35**, W561-567 (2007).

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

10 Shoemaker, B. A. & Panchenko, A. R. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Computational Biology* **3**, e42 (2007).

11 Kerrien, S. *et al.* IntAct--open source resource for molecular interaction data. *Nucleic Acids Research* **35**, D561-565 (2007).

12 Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Research* **34**, D535-539 (2006).

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

14 Mewes, H. W. *et al.* MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Research* **36**, D196-201 (2008).

15 Ceol, A. *et al.* MINT, the molecular interaction database: 2009 update. *Nucleic Acids Research* **38**, D532-539 (2010).

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

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

18 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 Research* **37**, W369-375 (2009).

19 Aloy, P. & Russell, R. B. Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 5896-5901 (2002).

20 Lu, L., Lu, H. & Skolnick, J. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins* **49**, 350-364 (2002).

21 Stein, A., Panjkovich, A. & Aloy, P. 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Research* **37**, D300-304 (2009).

22 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).

23 Chen, Y. C., Chen, H. C. & Yang, J. M. DAPID: a 3D-domain annotated protein-protein interaction database. *Genome Informatics* **17**, 206-215 (2006).

24 Deshpande, N. *et al.* The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Research* **33**, D233-237 (2005).

25 Kersey, P. *et al.* Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Research* **33**, D297-302 (2005).

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

27 Bork, P. *et al.* Predicting function: from genes to genomes and back. *Journal of Molecular Biology* **283**, 707-725 (1998).

28 Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. A

combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-86 (1999).

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

30 Kriventseva, E. V., Fleischmann, W., Zdobnov, E. M. & Apweiler, R. CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. *Nucleic Acids Research* **29**, 33-36 (2001).

31 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 Research* **34**, 2137-2150 (2006).

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

33 Bonifacino, J. S. & Traub, L. M. Signals for sorting of transmembrane proteins to endosomes and lysosomes. *Annual Review of Biochemistry* **72**, 395-447 (2003).

34 Heldwein, E. E. *et al.* Crystal structure of the clathrin adaptor protein 1 core. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 14108-14113 (2004).

35 Tillier, E. R. & Charlebois, R. L. The human protein coevolution network. *Genome Research* **19**, 1861-1871 (2009).

36 Hart, P. J. *et al.* Crystal structure of the human TbetaR2 ectodomain--TGF-beta3 complex. *Nature Structural Biology* **9**, 203-208 (2002).

37 Rubin, S. M., Gall, A. L., Zheng, N. & Pavletich, N. P. Structure of the Rb C-terminal domain bound to E2F1-DP1: a mechanism for phosphorylation-induced E2F release. *Cell* **123**, 1093-1106 (2005).

38 Massague, J., Blain, S. W. & Lo, R. S. TGFbeta signaling in growth control, cancer, and heritable disorders. *Cell* **103**, 295-309 (2000).

39 Massague, J. TGF-beta signal transduction. *Annual Review of Biochemistry* **67**, 753-791 (1998).

40 Bisgrove, B. W., Essner, J. J. & Yost, H. J. Regulation of midline development by antagonism of lefty and nodal signaling. *Development* **126**, 3253-3262 (1999).

41 Pawson, T. & Nash, P. Assembly of cell regulatory systems through protein interaction domains. *Science* **300**, 445-452 (2003).

42 Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S. K., Teichmann, S. A. & Weiner, J.,

3rd. The evolution of domain arrangements in proteins and interaction networks. *Cellular and Molecular Life Sciences* **62**, 435-445 (2005).

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

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

45 Huang, L. S. *et al.* 3-nitropropionic acid is a suicide inhibitor of mitochondrial respiration that, upon oxidation by complex II, forms a covalent adduct with a catalytic base arginine in the active site of the enzyme. *Journal of Biological Chemistry* **281**, 5965-5972 (2006).

Appendix A

List of publications

Journal papers

1. Chen, C.-C., **Lin, C.-Y.**, Lo, Y.-S. and Yang, J.-M. PPISearch: a web server for searching homologous protein-protein interactions across multiple species. *Nucleic Acids Research* **37**:W376-W383 (2009). (Impact factor: 6.878)

2. Lo, Y.-S., **Lin, C.-Y.** and Yang, J.-M. (2010) PCFamily: a web server for searching homologous protein complexes. *Nucleic Acids Research* (Accepted)