國立交通大學

生物科技學系

博士論文

同源蛋白質-蛋白質交互作用之研究

A Study of Homologous Protein-protein Interactions

研究生：陳俊辰

Chen Chun-Chen

指導教授：楊進木博士

Advisor: Yang Jinn-Moon, Ph. D.

中華民國九十八年七月

誌謝

此論文能夠完成，俊辰得到了太多的幫助和愛護——師長、父母、朋友——謹藉著這小小的篇幅，致上我最深切的謝意。

首先最應當深深感謝的，是恩師楊教授進木。是老師將俊辰引領進科學的領域，在我遇上困難時，您竭心盡力給予支持和指導；您嚴謹的治學態度和寬廣的視野，更令俊辰獲益良多。這樣的成長與變化，不但是在治學上，也更是變化了俊辰的人生態度。您與臺大流病所金教授傳春給予的這項寶貴禮物，俊辰將會受益許久，回思此情，感激實難言喻。

我的父母親陳天華先生、黃鳳娟女士一直給予俊辰溫暖的包容、無聲的支援。完成這本論文的過程是順遂和低潮相互倚伏的組合，歡喜的時候，我能與您分享，鬱滯的時刻，您們鼓勵我，是我最大的心靈支柱。感謝妹妹怡安，聽我訴苦，分擔我的情緒，在忙碌的工作中抽空關懷問候，謝謝你們，我卻愧無回報，彌增慚怍而已。

感謝實驗室的好朋友們：宇書、峻宇，如果沒有你們，這本論文無法完成。許多個一起熬夜的夜晚，拉著你們一起討論的 meeting，我深深感激。彥甫、其樺、章維、怡馨、凱程、志達、彥修、敬立、彥超、力仁、韋帆，謝謝你們給這論文的許多建議、協助、還有關心，共同走過來的這一段日子，真的是非常美好的回憶。

俊辰何其幸運，竟然能得到這樣多的幫助、愛護。最後請容我再一次說：謝謝，感恩您們。

Abstract

The discovery of sequence and structural homologs to a known protein often provides clues, such as biochemical function and domain, for understanding a newly determined protein. In this thesis, we have proposed a new concept of "homologous protein-protein interaction (PPI)", which is helpful for understanding a newly determined PPI. We proposed evidence of the existence of homologous PPIs, developed a methodology to rapidly identify homologous PPIs, and infer the transferability of interacting domains and functions of a query protein pair. We found that homologous PPIs can be distinguished when homolog pairs were in the annotated database and have significant joint sequence similarity (*E*-value $\leq 10^{-40}$) with the query protein pair. As an increasing number of reliable PPIs become available and high-throughput experimental methods provide systematic identification of PPIs, there is a growing need for fast and accurate methodologies for discovering homologous PPIs of a PPI. For this demand, we combined the methodology with an annotated PPI database (290,137 PPIs in 576 species) to construct a PPISearch server (http://gemdock.life.nctu.edu.tw/ppisearch/) to supply service for global researchers. In addition, we have utilized the concept of homologous PPIs to cross-species PPI prediction and cross-species network comparison. Currently, large-scale interaction networks are available for only several model organisms. Our preliminary results suggested that the concept of homologous PPI are useful for a systematic transfer of PPI networks between multiple species.

中文摘要

對新發現的蛋白質而言，其同源序列(sequence homolog)及同源結構(structural homolog)通常能提供研究者關於該蛋白質生物功能(function)、功能區塊(domain)的若干線索。更進一步，我們提出一個新的概念「同源蛋白質-蛋白質交互作用(homologous protein-protein interaction, 簡稱homologous PPI)」，此概念對於了解新被發現的蛋白質交互作用是很有幫助的。在此研究中，我們提出一套合理定義、鑑別homologous PPIs的方法論，並提供了自然界中homologous PPIs確實存在的證據。而在現今PPI資料迅速增加的情況下，研究者迫切地需要能夠快速搜尋並提供合理homologous PPIs的方法，以對未知功能的PPI進行深入的研究。為達成此目的，我們將本研究中發展出的方法論與一個大型PPI資料庫(包含來自 576 個物種的 290,137 筆PPIs)結合，建立了一個網路服務工具PPISearrh (http://gemdock.life.nctu.edu.tw/ppisearch/)，使用者可輸入任一提問蛋白質對(query protein pair)，它具有快速搜尋該提問蛋白質對之homologous PPIs (亦稱PPI family)的功能，並已開放給全球研究者使用。

本研究亦提供統計基礎，將homologous PPIs所具有的生物功能及功能區塊用來合理地註解未知性質的蛋白質間交互作用，其觀念與用同源序列(或結構)來了解某個新發現蛋白質相類。本研究中我們提出，homologous PPIs須具備三個條件：是提問蛋白質對中兩個蛋白質的同源序列之配對；已知紀錄在資料庫中；以及足夠顯著的(*E*-value $\leq 10^{-40}$)合成序列相似度(joint sequence similarity)。

在本研究中，我們也初步將 homologous PPI 的新概念運用於兩個研究課題，第一是跨物種的蛋白質間交互作用之預測(cross-species PPI prediction)，另一個是跨物種間蛋白質交互作用網絡之比對(cross-species network comparison)。目前只有數種模式生物(model organisms)已累積大量的 PPI 資料，其蛋白質交互作用網絡可被較完整地了解，我們的初步成果顯示，homologous PPI 之概念可以對研究上述兩個課題有所助益。

Contents

Chapter 1.

Introduction

1.1 Motivation

Many sequence and structure alignment methods have been developed to identify homologs of newly determined sequences and structures[1, 2]. The discovery of sequence and structural homologs to a known protein often provides clues, such as function and domain, for understanding a newly determined protein. Likewise, we proposed a new concept "homologous protein-protein interaction (PPI)". As an increasing number of reliable PPIs become available, there is a growing need for fast and accurate methodolgies for discovering homologous PPIs to understand a newly determined PPI.

In this thesis, we provided evidence to supply the existence of homologous PPIs, developed a methodology to search homologous PPIs across multiple species, and annotate the query protein pair. Public PPI databases, such as IntAct, BioGRID and MINT, offer a good basis for researchers to investigate this issue. In addition, we supplied the new concept of homologous PPIs, and combined our method identifying homologous PPIs with an annotated PPI data set (was consisted of five public PPI database) to construct a web server, PPISearch, to supply service for global researchers.

1.2 Background

The interactions between proteins are critical for most of biological functions. For example, signals from the exterior of a cell are mediated to the inside of that cell by protein-protein

interactions of the signaling molecules. This process, called signal transduction, plays a fundamental role in many cellular processes and in multiple diseases (e.g. cancers). Those cellular behaviors and phenotypes, for example, growth, cell division, cellular differentiation, and apoptosis, mainly are mediated by protein-protein interactions.

The interactions between proteins are critical to most of cellular processes. To identify and characterize protein-protein interactions 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[3], known 3D complexes[4] and interologs[5]] have been proposed[6]. Some PPI databases, such as IntAct[7], MIPS[8], DIP[9], MINT[10], and BioGRID[11] 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.

Sequence homologs of a known protein often provide clues to understand 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[12, 13] applied homologous interactions to assess the reliability of PPIs. Otherwise, several computational approaches, such as Yu *et al.* (2004)[5], Scott *et al.* (2007)[14], and InteroPORC[15], combined known PPIs from one or more species and orthology or homology relationships between the source and targets species. These data sets of predicted PPIs, which are constructed by these methods, are available online. However, only a few PPIs in these data sets have been identified by experiments and these sets have not supplied service of searching PPIs across species. Currently, there is no server supplies users to input a pair of interacting proteins

and provides their homologous interactions from known PPIs for investigating the query protein pair.

According to our knowledge, this study first showed the existence of homologous PPIs, querying a limited number of databases. Additionally, PPISearch is also the first server that identifies homologous PPIs from annotated PPI databases and infers the 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. We supplied a threshold of conservation ratio of biological characteristics (e.g. domains and function). These characteristics have higher ratio than the threshold would be reliably assigned to the query protein pair. Our results demonstrated that this server achieves high agreements on interacting domain-domain pairs and function pairs between query protein pairs and their respective homologous PPIs.

1.3 Thesis overview

The thesis is organized as follows. In **Chapter 2**, we proposed the PPI concept and the methodology to find the homologous PPIs for a given query PPI. Then, we combined the methodology with a non-redundant PPI data set to construct a web server, PPISearch. The PPI data set consists of 290,137 PPIs derived from 576 species.

In **Chapter 3**, we proposed evidence supplying the existence of PPI families, the results of homologous PPIs search, and discussed the observations of PPI families. We examined the new concept of homologous PPI based on four insights. Moreover, we used case studies to describe the insights into the concept of homologous PPI, and the statistical analyses of PPI families. Our results demonstrated the utility and feasibility of the PPISearch server in

identifying homologous PPIs and inferring conserved domains and functions from PPI families. By allowing users to input a pair of protein sequences, PPISearch is the first server that can identify homologous PPIs from annotated PPI databases and infer transferability of interacting domains and functions between homologous PPIs and a query. We showed that PPISearch is a fast homologous PPIs search server and is able to provide valuable annotations for a newly determined PPI.

In **Chapter 4**, we applied the concept of homologous PPI (i.e. PPI family) to cross-species prediction of PPIs and cross-species network comparisons. In recent years, for complementing experimental techniques (e.g. yeast two-hybrid system and mass spectroscopy), a number of computational methods, such as PathBLAST[16, 17] and interologs[5, 18], have been developed to predict PPIs[19]. The concept of interologs has been extended to be a "generalized interolog mapping" method[5]. Our results showed that our discovery can be used to advance the generalized interolog mapping method. In addition, we used case studies to present that the concept of homologous PPI are useful for a systematic transfer of PPI networks between multiple species.

Chapter 2.

Methods and Materials for Finding Homologous Protein-protein Interactions

In this chapter, we presented the concept of PPI family and the method to find the homologous PPIs for a given query PPI. Figure 1 illustrates the concept of searching homologous PPIs. For this purpose, we constructed a non-redundant PPI data set for searching homologous PPIs. The data set consists of PPIs derived from five public databases, IntAct, MIPS, DIP, MINT and BioGRID. Total number of PPIs in this data set is 290,137 in 576 species.

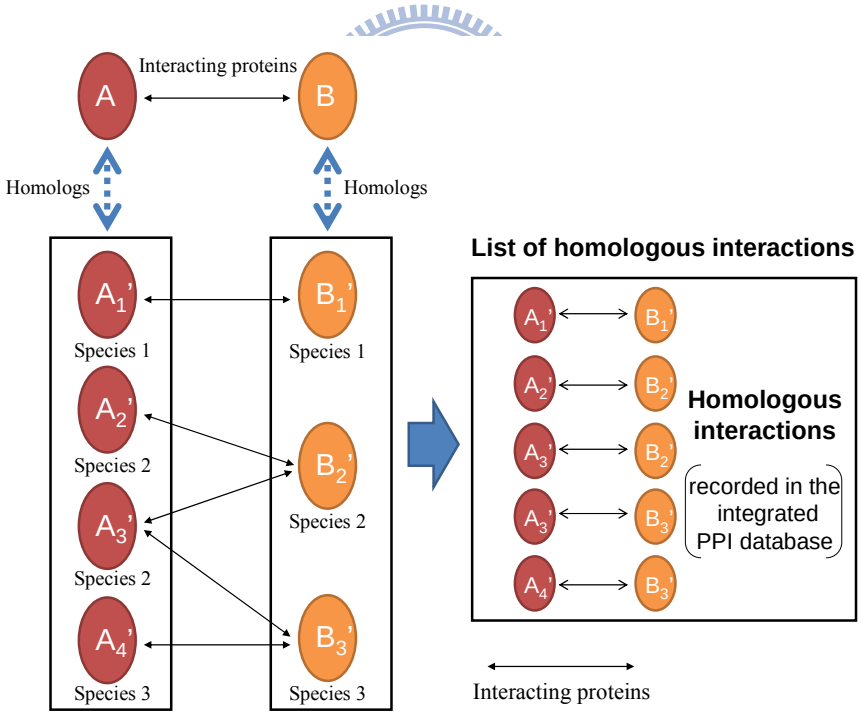


Figure 1. Illustration of searching homologous interactions. Interacting proteins A and B is the query protein pair given by users. A_1'-A_4' and B_1'-B_3' are the homologs (defined by BLASTP *E*-value) of proteins A and B, respectively. The homolog pairs recorded in the integrated PPI data set (290,137 PPIs) are considered as potential homologous interactions.

Additionally, we presented how we evaluate the reliability of homologous PPIs, which are defined by sequence similarity (BLASTP *E*-values) and joint sequence similarity (joint *E*-value) between the query protein pair and these homologous PPIs.

2.1 Overview of homologous protein-protein interaction search

In this study, we developed a methodology for searching homologous PPIs and used it to constructing a web server, PPISearch. Figure 2 shows the details of the PPISearch server to search homologous PPIs of a query protein pair (A and B) by the following steps (Figure 2A). 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 (Figure 2B 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 $\leq 10^{-40}$) between candidates and the query (Figure 2D). Then, we measure the conservation ratios of domain-domain pairs (DDPs; Pfam[20] and InterPro[21] domains) and protein functions (Gene Ontology annotations[22]) derived from these homologous PPIs of the query (Figure 2E).

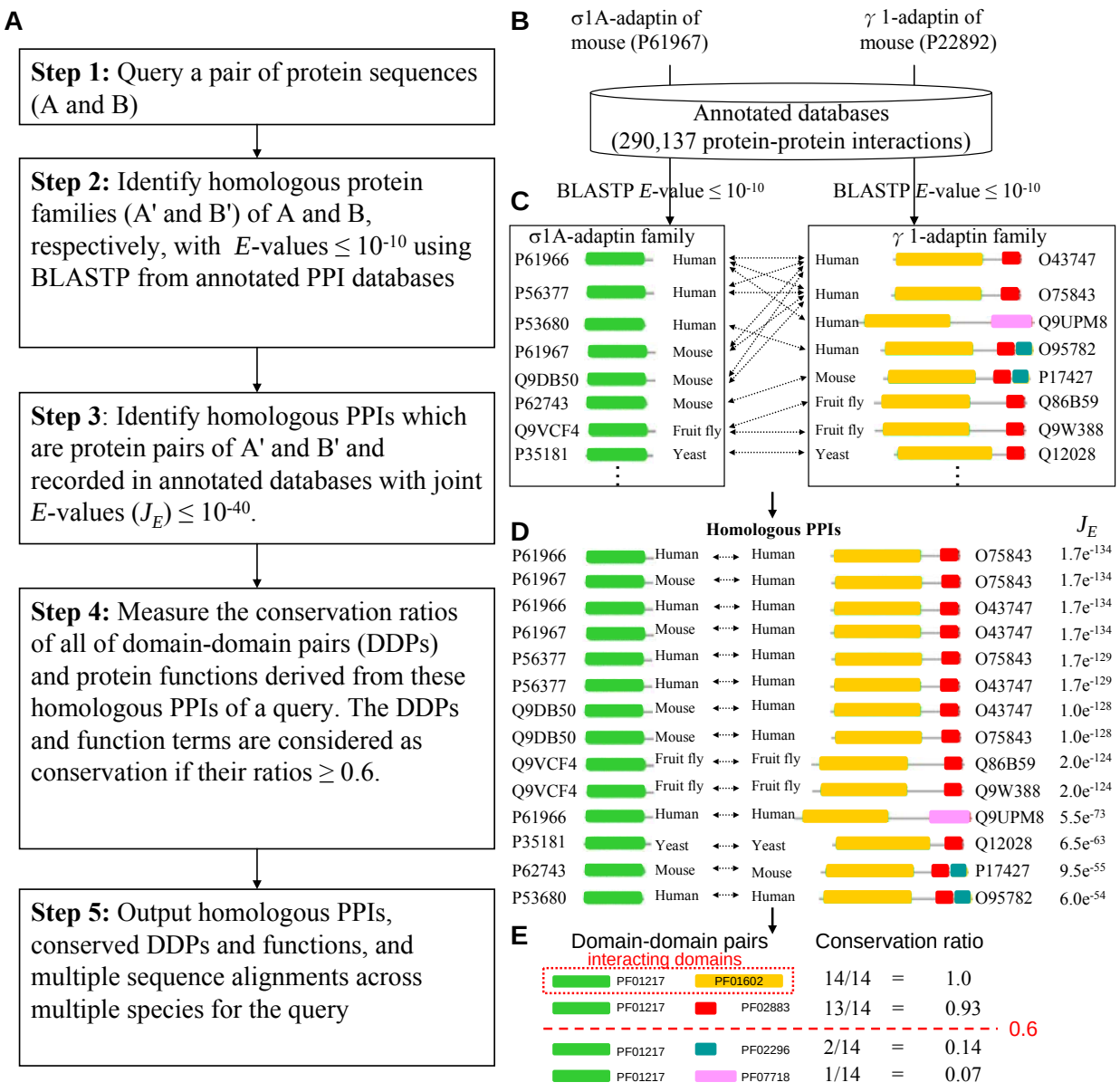


Figure 2. Overview of the PPISearch server for homologous protein-protein interaction search and conservation analysis using proteinsσ1A-adaptin and γ1-adaptin as the query. (**A**) The main procedure. (**B**) Identify homologs ofσ1A-adaptin and γ1-adaptin using BLASTP to scan the annotated PPI databases. (**C**) The homologous families of σ1A-adaptin and γ1-adaptin with *E*-values ≤10^{-10}. (**D**) Homologous PPIs of the query. (**E**) Conservation ratios of domain-domain pairs derived from homologous PPIs.

2.2 Homologous protein-protein interaction

The concept of homologous PPI is the core of the this study to identify the PPI family and measure DDPs and functional conservations of a query protein pair (A and B). We defined a homologous PPI as follows: (1) homologs of A and B are proteins with significant sequence similarity BLASTP *E*-values $\leq 10^{-10}$;[5, 18] (2) significant joint sequence similarity (joint *E*-value $J_E \leq 10^{-40}$) 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. This work followed previous studies[5, 18] to define joint sequence similarity as

$$J_E = \sqrt{E_A \times E_B} \quad (1)$$

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[23] (see **Chapter 3**).

2.3 Non-redundant data set for searching homologous PPIs

Table 1 lists the dates and numbers of PPIs of the five public databases.

Table 1. Five source data sets of PPIs

Database	Number of PPIs	Date
IntAct	147,634	Dec. 14, 2008
MIPS	18,529	Oct. 1, 2008
DIP	52,445	Oct. 14, 2008
MINT	77,846	Oct. 28, 2008
BioGRID	150,827	Dec. 17, 2008
Total	447,281	
Non-redundant	290,137	

After removing redundant PPIs, the annotated data set used in this study has 290,137 PPIs. These PPIs were identified experimentally from 576 species.

We describe briefly these public databases as follows: (1) **IntAct**: All interactions are derived from literature curation or direct user submissions and are freely available[7]. IntAct is a freely available and open source database system of protein interaction data; (2) **MIPS**: The Munich Information Center for Protein Sequences (MIPS) combines automatic processing of large amounts of sequences with manual annotation of selected model genomes. PPIs of MIPS are annotated by the compilation of manually curated databases for protein interactions based on literature to serve as an accepted set of reliable annotated interaction data[8]; (3) **DIP:** The DIP database catalogs experimentally determined PPIs. It combines information from a variety of sources to create a single, consistent set of PPIs. The data stored within the DIP database were curated manually by both expert curators and automatically using computational approaches that utilize the knowledge about the PPI networks extracted from the most reliable, core subset of the DIP data[9]; (4) **MINT**: The Molecular INTeraction database focuses on experimentally verified protein-protein interactions mined from the scientific literature by expert curators[10]; (5) **BioGRID**: The BioGRID (Biological General Repository for Interaction Datasets) database was developed to house and distributes collections of protein and genetic interactions from major model organism species. BioGRID currently contains ~150,000 interactions from six different species, as derived from both high-throughput studies and conventional focused studies[11].

The tabular data files of PPIs from the five public databases were downloaded. We merged all of these PPIs and removed duplications by using UniProt accession numbers. A total of 290,137 PPIs in 576 species were included in our investigation.

2.4 Annotations of homologous protein-protein interactions

A query protein pair and its homologous PPIs, significant both in sequence and joint sequence similarity, can be considered a PPI family. The concept of PPI families is similar to that of protein sequence family[20, 21] and protein structure family[24]. We believe that PPI families can be applied widely in biological investigations. Here, we assume that the members of a PPI family are conserved on specific functions and in interacting domain(s). Using these conservations of a PPI family, our server can be used to annotate the protein functions and DDPs of a query protein pair.

2.4.1 Transferability of molecular function

These members of a PPI family often have similar molecular functions. We used the molecular function (MF) terms of Gene Ontology[22] to annotate the functions of a query protein pair. The conservation ratio (CRF_m) of an MF term pair (MFP) m in homologous PPIs of a query i is utilized to measure the agreement and is defined as

$$CRF_m = \frac{\text{Number of homologous PPIs with a GO MF term pair } m}{\text{Number of homologous PPIs of query } i} \tag{2}$$

Additionally, the shared ratio of MFPs (SRF), which is statistically derived from 106,997 annotated queries, is utilized to estimate the transferability of conserved function pairs shared by the query and its homologous PPIs. The SRF against different ratio k is defined as

$$SRF = \frac{\sum_{i \in Q} f_i(CRF_m \geq k)}{\sum_{i \in Q} F_i(CRF_m \geq k)} \tag{3}$$

where Q is a set of annotated PPIs in databases; i is a query protein pair; $f_i(\mathrm{CRF}_m \geq k)$ is the number of MFPs with CRF_m values exceeding k and these MFPs are shared by the query i and its homologous PPIs; and $F_i(\mathrm{CRF}_m \geq k)$ is the total number of MFPs with $\mathrm{CRF}_m \geq k$, where MFPs are derived from homologous PPIs of the query i. Here, k is set to 0.6.

2.4.2 Transferability of domain-domain pairs

A query protein pair and its homologous PPIs often show conserve interacting DDPs. To measure the occurrence 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

$$CRD_p = \frac{\text{Number of homologous PPIs with a domain pair } p}{\text{Number of homologous PPIs of query } i} \quad (4)$$

Figure 2D and E show an example to calculate the CRD values of four DDPs. In addition, to evaluate the transferability of DDPs between a query and its homologous PPIs statistically, this study defines the shared ratio (SRD) of DDPs using CRD_p and 103,762 annotated PPIs as query protein pairs. The SRD of DDPs against different ratio c is given as

$$SRD = \frac{\sum_{i \in Q} d_i(CRD_p \geq c)}{\sum_{i \in Q} D_i(CRD_p \geq c)} \quad (5)$$

where Q is a set of annotated PPIs in databases (here, the total number of PPIs in Q is 103 762); i is a query protein pair; $d_i(\mathrm{CRD}_p \geq c)$ is the number of DDPs with CRD_p values exceeding c; and these DDPs are shared by the query i and its homologous PPIs. $D_i(\mathrm{CRD}_p \geq c)$ is the total number of the DDPs with $\mathrm{CRD}_p \geq c$, 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.

2.5 Data sets for evaluating the approach of searching homologous protein-protein interactions

For evaluating the usefulness of the PPISearch server for the discovery of homologous PPIs and for the annotations of a query protein pair, we constructed two query protein sets, termed HOM and ORT. For searching homologous PPIs, HOM and ORT data sets are used to assess performance of PPISearch and to determine the threshold of joint *E*-value J_E [Equation (1)] (Figure 3A).

2.5.1 HOM data set

The HOM set includes all of 290,137 PPIs. The HOM set was applied to infer the relations between conservation ratios [CRF and CRD defined in Equations (2) and (4)] and the transferability of DDPs and MFPs, respectively, between a query and its homologous PPIs.

2.5.2 ORT data set

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[23] and CluSTr[25] databases. These clusters contain all sequenced organisms (1125 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.

Chapter 3.

Evidence Supplying the Existence of Homologous Protein-protein Interactions

In this chapter, we presented the evidence of existence of homologous PPIs (**Section 3.1**), the results of homologous PPIs search (i.e. PPISearch) (**Sections 3.2-3.3**) and discussed the observations of PPI family. We used case studies to describe the insights used to examine the concept of homologous PPI and statistically analyze PPI families (**Section 3.4**).

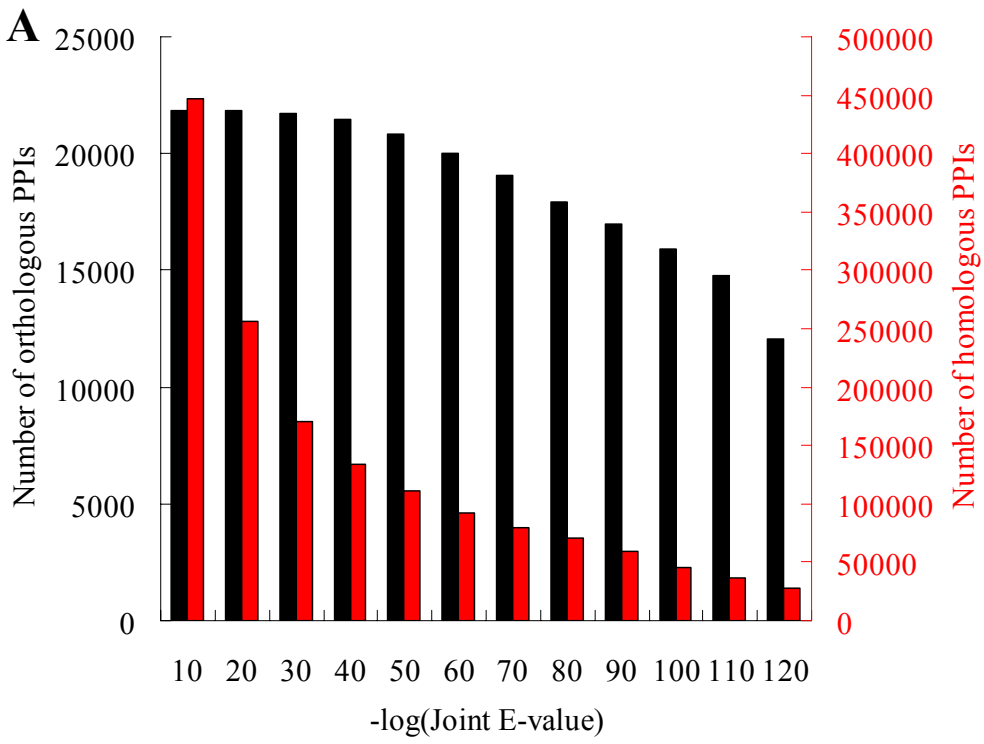
3.1 Evidence of the existence of homologous PPIs

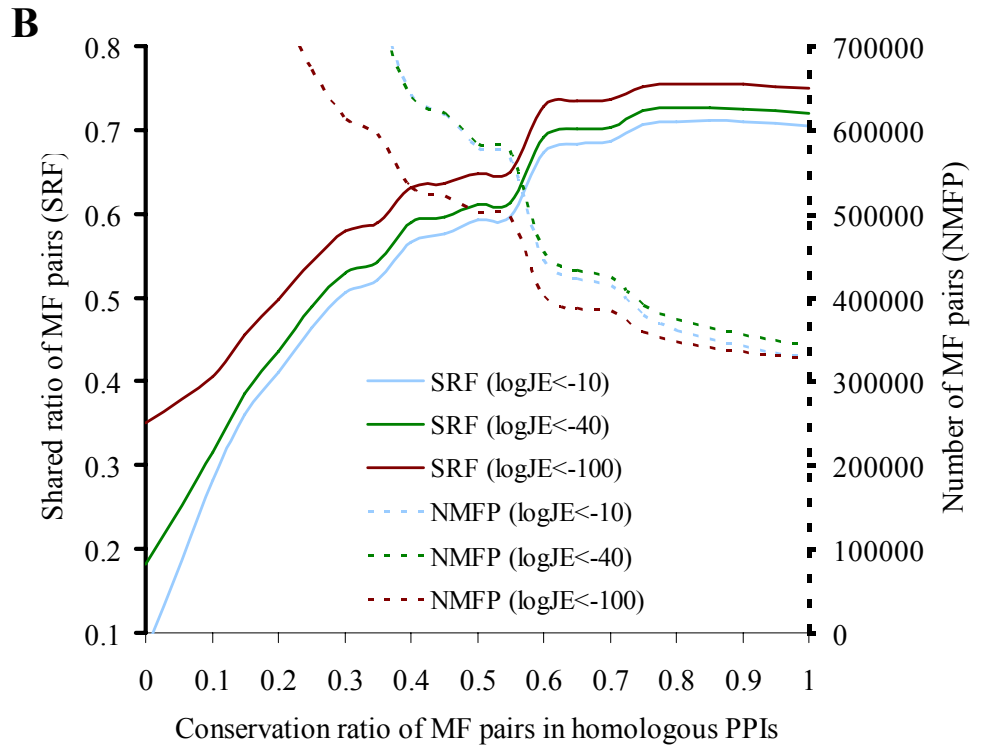
We analyzed the results of homologous PPIs by four views. Firstly, we observed the conservation of biological function in PPI families. Secondly, we observed the conservation of domain pairs in PPI families. Thirdly, because of the sharing of conserved domain pairs in PPI families, we observed the conservation of interacting domains in PPI families (based on protein 3D structures). Finally, because of the sharing of conserved interacting domains in PPI families, we observed the conservation of binding interface between two proteins of each PPI in PPI families. These evidences from the four views showed the existence of homologous PPIs (i.e. PPI family).

3.1.1 Conservation of molecular function in PPI families

To verify the discovery of homologous PPIs, we selected two query protein sets, termed HOM and ORT. To search homologous PPIs, HOM and ORT are used to assess PPI families and to evaluate the threshold of joint *E*-value J_E (Figure 3A). In addition, the HOM set was applied to infer the relations between conservation ratios [CRF defined in **Chapter 2**] and the transferability of MFPs, respectively, between a query and its homologous PPIs (Figure 3B). The HOM set includes all 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.

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 3A shows the relationships between joint *E*-value J_E 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 shows that the proposed method is able to identify 98.2% orthologous PPIs with a reasonable number of homologous PPIs when $J_E \leq 10^{-40}$.





(Continued on next page)

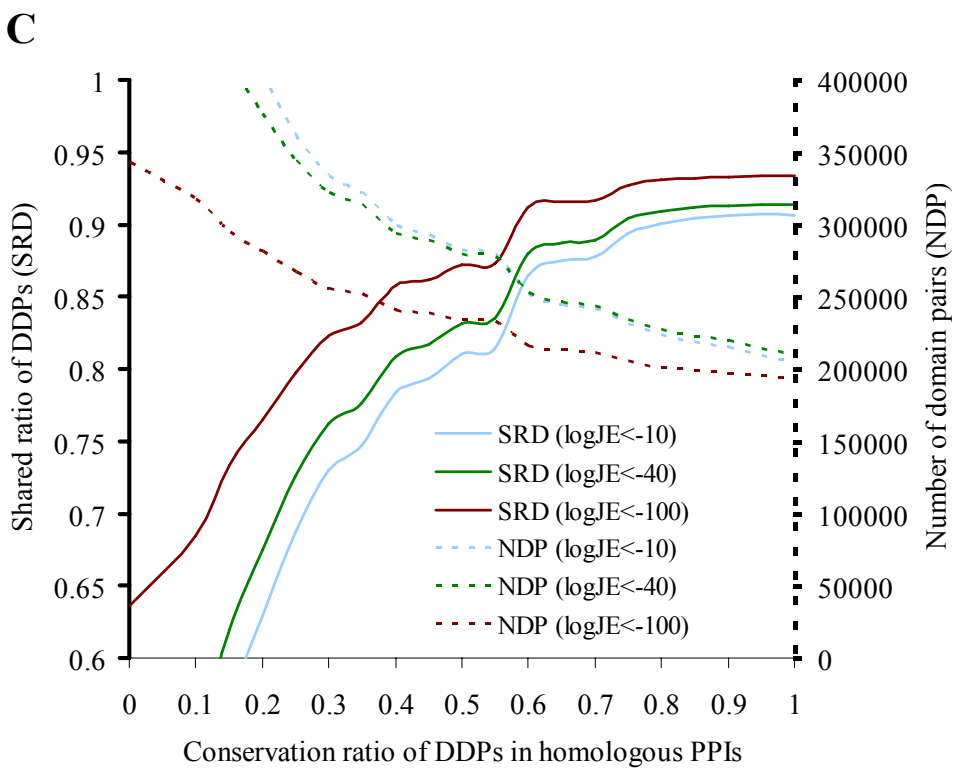


Figure 3. Conservations of biological functions and domain pairs in PPI families. (**A**) The relationships between joint *E*-value J_E and the numbers of orthologous PPIs (black) and homologous PPIs (red) derived from 290,137 annotated PPIs. (**B**) The relationships between the conservation ratios of molecular function pairs (MFPs) with the 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 if the conservation ratio is 0.6 and the joint *E*-value is 10^{-40} (green lines). (**C**) 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).

To evaluate the transferability of MFPs between a query and its homologous PPIs, we used the SRF [Equation (3)]. The HOM set is also used to evaluate the utility of the PPISearch server in annotating the query protein pair. By excluding proteins without molecular function annotations of GO from the query set, 106,997 PPIs are used to evaluate the transferability (SRF) of conserved MFPs between these query PPIs and their respective homologous PPIs (Figure 3B). 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 SRD is 0.69 and the number of MFPs is 454,251.

3.1.2 Conservation of domain pairs in PPI families

In addition, the HOM set was applied to infer the relations between conservation ratios [CRD defined in **Chapter 2**] and the transferability of DDPs, respectively, between a query and its homologous PPIs (Figure 3C). To evaluate the transferability of DDPs between a query and its homologous PPIs, we used the SRD [Equation (5)]. 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 respective homologous PPIs (Figure 3C).

Figure 3C 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 $\leq$ 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 and the number of DDPs is 252,728. This result demonstrates that members of a PPI family reliably share DDPs (or interacting domains). Additionally, similar results were obtained for the transferability of conserved functions between homologous PPIs and the query (Figure 3B).

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

3.1.3 Conservation of interacting domains in PPI families

The two above evidence were acquired by sequence-based searching. As an increasing number of structural data was available (e.g. protein complexes in PDB), we used structure-based views to examine the concept of homologous PPIs. In this section 3.1.3, we observed the conservation of interacting domains in PPI families because of the assumption of "the members of a PPI family have similar interacting domains".

Firstly, we collected a data set of protein complexes. Each complex was composed of two protein chains (i.e. heterodimer or homodimer) and was (1) recorded in the annotated PPI database (290,137 PPIs) and (2) recorded in iPfam database. iPfam is a database []. After selecting protein complexes from PDB, a data set of 1,014 complexes (in other words, 1,014 PPIs) was constructed.

Figure 4 shows the method we calculated the conservation of interacting domain in PPI families. Proteins A-B is an interacting protein pair, in which there are two physical domain-domain interactions (DDIs). If a protein pair between homologs (*E*-value $\leq 10^{-10}$) A' and B' kept $\geq$ 1 DDI, we considered the protein pair has similar interacting domains to the query pair A-B. We compared the two partitions of "PPI families" and "Non PPI families". The "PPI families" consisted of PPIs with $J_E \geq 10^{-40}$, for example, the PPIs circled by blue. Conversely, the "Non PPI families" consisted of PPIs having $J_E < 10^{-40}$.

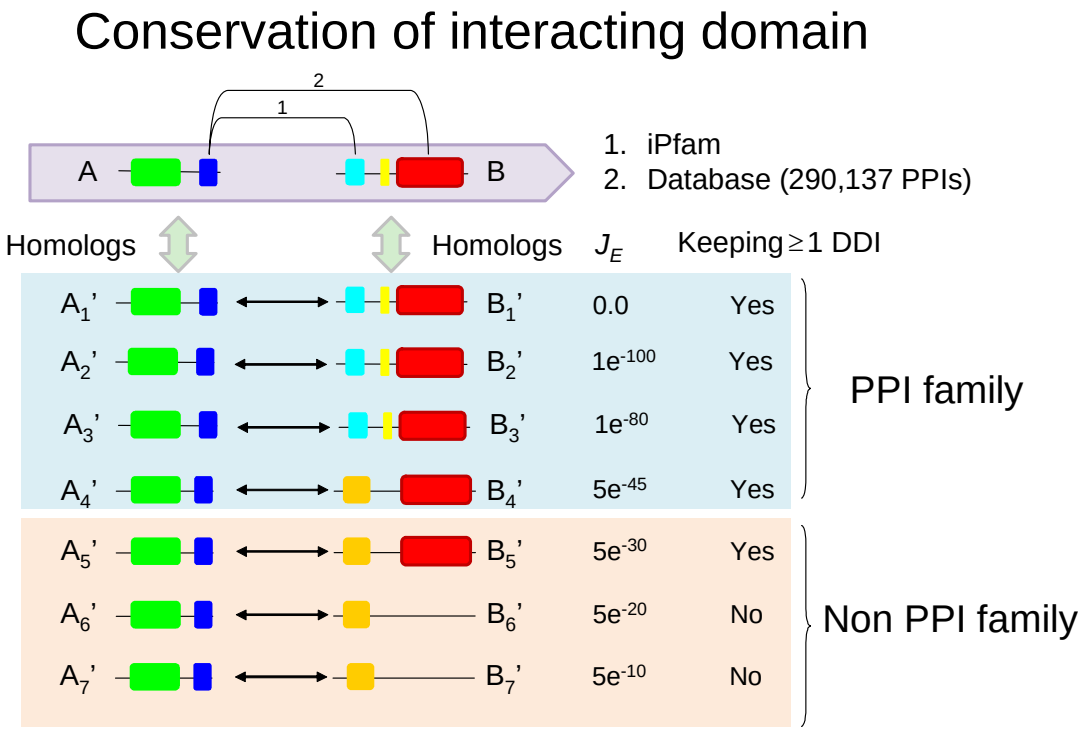


Figure 4. Illustration of the method we calculated the conservation of interacting domain in PPI families. The rectangles colored by green, blue, light blue, yellow, and red mean domains. The query interacting protein pair A-B has two DDIs. All of the pairs of A' and B' homologs are marked "Yes" (keeping ≥ 1 DDI with the query PPI A-B) or "No" (keeping no DDI with the query PPI A-B).

Figure 5 indicates the results of observing the conservation of interacting domains in PPI families. We found that the number of PPIs keeping ≥ 1 DDI (11,060 PPIs) were 2.35-fold more than that of PPIs not keeping DDI (4,699 PPIs) in the set "PPI families". In comparison, the number of PPIs not keeping (27,264 PPIs) were 1.74-fold that of PPIs keeping DDI ≥ 1 DDI (15,653 PPIs) in the set "Non PPI families".

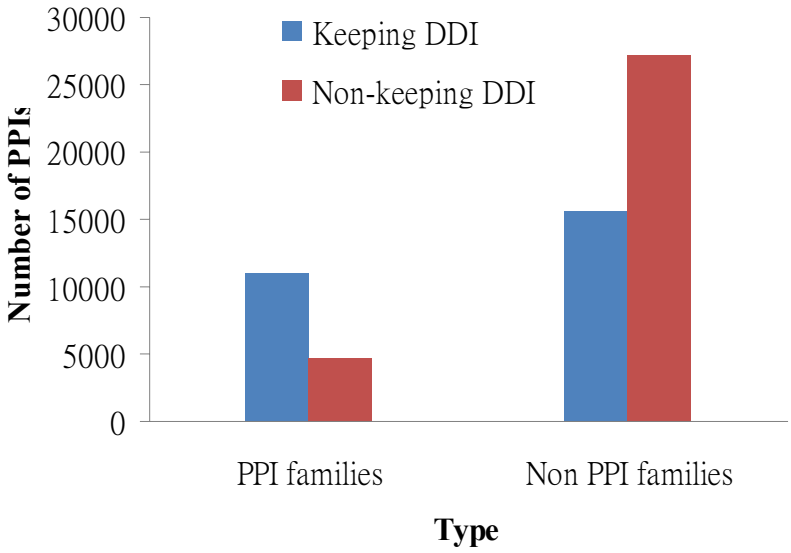


Figure 5. Conservation of interacting domains in PPI families. The number of PPIs keeping ≥ 1 DDI is 11,060 PPIs (blue) and that of PPIs not keeping DDI is 4,699 PPIs (red) in the set "PPI families". In comparison, the number of PPIs not keeping is 27,264 PPIs (red) and that of PPIs keeping DDI ≥ 1 DDI is 15,653 PPIs (blue) in the set "Non PPI families".

These results indicated that there was higher conservation of interacting domains in PPI families than in non-PPI families. In the **Section 3.4 Discussion**, we supplied a possible reason of why there were 4,699 PPIs which not keeping DDI in the set "PPI families".

3.1.4 Conservation of binding model in PPI families

After we acquired the evidence of conservation of interacting domain, we were interesting to observed the similarity of structural binding interfaces within PPI families. This idea was derived from the assumption that the PPI members of a PPI family have similar structural binding interfaces between two protein partners of each PPI. We have developed 3D-partner,

which is a web tool to predict interacting partners and binding models of a query protein sequence through structure complexes and a new scoring function.

For above purpose, we collected a data set of protein complexes from PDB, which was composed of 517 heterodimers because 3D-partner was developed based on protein structures of heterodimers. Similar to the description in **Section 3.1.3**, we compared the two subsets of "PPI families" (4,998 PPIs) and "Non PPI families" (9,102 PPIs). The results of comparison between the two data subsets are showed in Figure 6. We used a threshold *Z*-score to measure the similarity of binding model and identify interacting partners with the query. The *Z*-score reveals that the proportion of true positives rises when a higher *Z*-score is utilized. The *P* value of *T* test between the *Z*-scores of the two subsets "PPI families" and "Non PPI families" is less than 10^{-30}. The results indicated that there was significant difference between the two subsets.

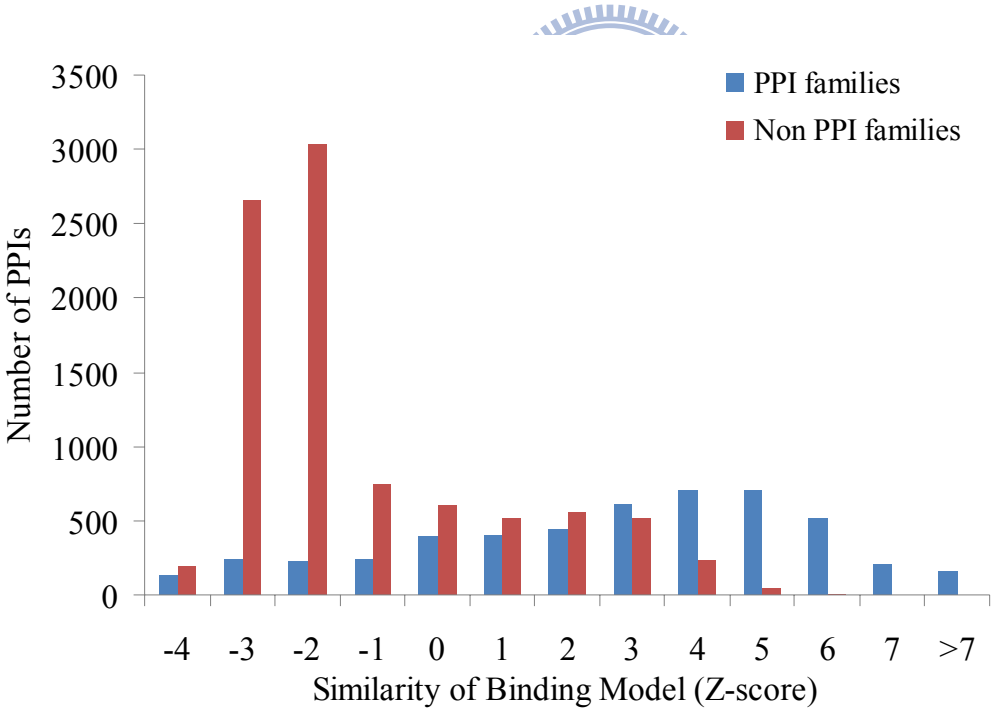


Figure 6. Distribution of *Z*-score (i.e. similarity of binding models) in two subsets of "PPI families" (blue) and "Non PPI families" (red). There are 4,998 and 9,102 PPIs in the two subsets, respectively.

In addition, we were interesting that why many PPIs in “PPI families” subset have low similarity with the query PPIs. For this purpose, we used a descriptor, aligned contact residue identity (CI), to observe the similarity of binding interface in a PPI family. Figure 7 shows an illustration of how we calculated CI values.

We selected the PPIs in "PPI families" subset for observation (Figure 8) and example analysis. Figure 8 indicates the distributions of CI values in PPIs with *Z*-score ≥ 1.96 (i.e. 95% confidence interval) and that with *Z*-score < 1.96. We found that 93.5% of PPIs with *Z*-score < 1.96 have CI = 0. In other words, these results suggested that we would get PPIs with different binding model from the query PPIs through the search method we currently used. In the **Section 3.4 Discussion**, we supplied a possible reason of causing this observation.

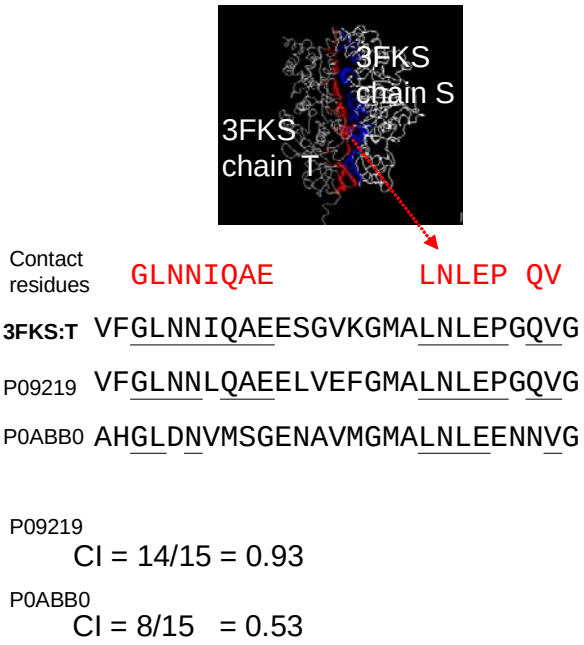


Figure 7. An example of how to calculate CI values. The 15 residues colored by red are part of contact residues (on 3FKS chain T) in the interacting interface between 3FKS chains T and chain S. The underlined residues in the aligned sequences P09219 and P0ABB0 are the residues which are identical to the contact residues on 3FKS chain T. In this case, CI of P09219 is 14/15 = 0.93 and that of P0ABB0 is 8/15 = 0.53, respectively.

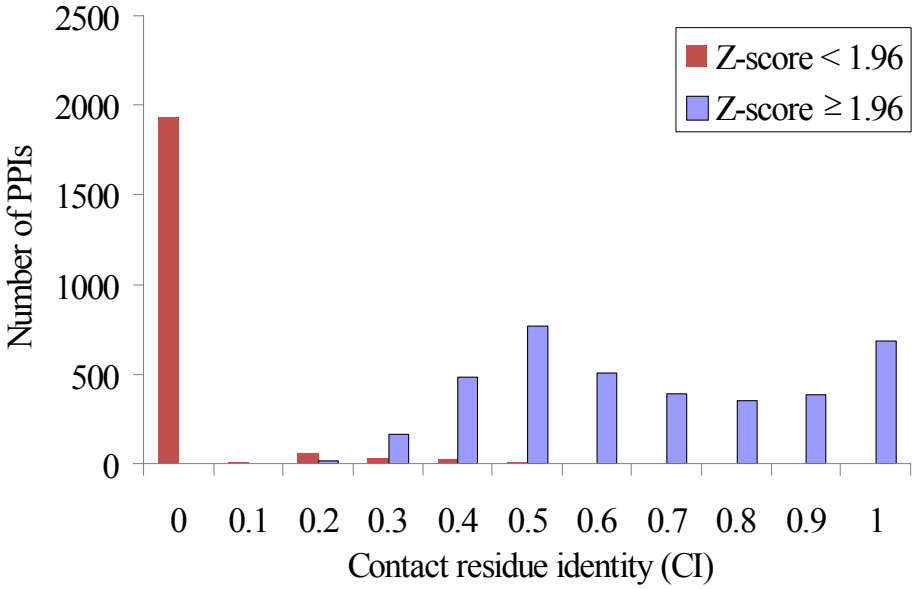


Figure 8. Distribution of CI values in PPIs with *Z*-score ≥ 1.96 (blue) and *Z*-score < 1.96 (red).

3.2 Input, output, and options of the PPISearch server

The PPISearch is an easy-to-use web server (Figure 9). Users input a pair of protein sequences in FASTA format or UniProt ID, and choose *E*-value thresholds for homologs and for homologous PPIs (Figure 9A). In addition, users can assign the CRD and CRF thresholds, specific species and the number of homologous PPIs in a species.

To evaluate the usefulness of the PPISearch server for the discovery of homologous PPIs and for the annotations of a query protein pair, we selected two query protein sets, termed HOM and ORT. To search homologous PPIs, HOM and ORT are used to assess PPISearch performance and to determine the threshold of joint *E*-value J_E (Figure 3A). In addition, the HOM set was applied to infer the relations between conservation ratios [CRD and CRF defined in **Chapter 2**] and the transferability of DDPs and MFPs, respectively, between a query and its homologous PPIs (Figure 3B and C). The HOM set includes all 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.

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 3A shows the relationships between joint *E*-value J_E 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 shows that the proposed method is able to identify 98.2% orthologous PPIs with a reasonable number of homologous PPIs when $J_E \leq 10^{-40}$.

To evaluate the transferability of DDPs and MFPs between a query and its homologous PPIs, we used the SRD [Equation (3)] and SRF [Equation (5)]. 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 respective homologous PPIs (Figure 3B). 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 3C).

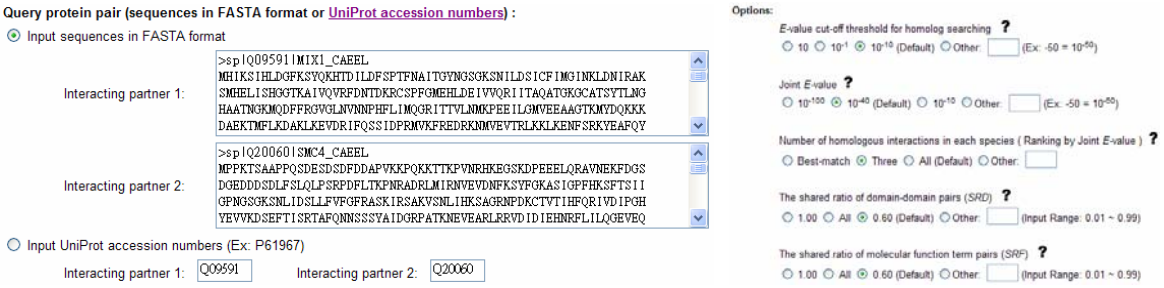
Figure 3B shows the relationship between conservation ratios (CRD) of DDPs and the SRD ratios. The SRD ratio increases significantly (solid lines) when the CRD increases and CRD $\leq$ 0.6. Conversely, the number of DDPs derived from 103,762 PPI families decreases (dotted lines) as CRD increases. If the CRD is set to 0.6 and the joint *E*-value is set to 10^{-40} (green lines), the SRD is 0.88 and the number of DDPs is 252,728. This result demonstrates that members of a PPI family derived by PPISearch reliably share DDPs (or interacting domains). Additionally, similar results were obtained for the transferability of conserved

functions between homologous PPIs and the query (Figure 3C). 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 and the number of MFPs is 454,251.

These results reveal that the PPISearch server 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 demonstrate that this server achieves high agreement on DDPs and MFPs between the query and their respective homologous PPIs.

Typically, the PPISearch server yields homologous PPIs within 20 seconds when sequence length is $\leq$ 350 (Figure 9B). This server identifies homologous PPIs in multiple species; conservations and GO annotations of protein functions; conservations and annotations of DDPs; and the best-matched protein pairs of the query (Figure 9C). Additionally, the PPISearch server provides multiple sequence alignments of homologous PPIs and indicates the conserved residues based on amino acid types. For each homologous PPI, this server shows the alignments and experimental annotations (e.g. interaction types, experimental methods, gene names and GO terms).

A



B

Homologous Interactions of the Query Protein Pair

Query pair	Query 1 :**Q09591** Query 2 :**Q20060**
Number of Homologous Interactions	**7** Multiple sequence alignment
Number of Species	**4**
Number of GO annotation pairs	**6** (>=0.6), 136 (Total)
Number of Pfam domain pairs	**4** (>=0.6), 4 (Total)
Number of InterPro annotation pairs	**4** (>=0.6), 3 (Total)

Query 1: Q09591			Query 2: Q20060			Interaction Similarity		Databases					
TaxID	Protein	*E*-value	TaxID	Protein	*E*-value	Joint *E*-value ?	Rank in species	IntAct	DIP	MIPS	MINT	BioGRID	Detail
6239	Q09591	0.0	6239	Q20060	0.0	0.0	1	✓					View
6239	Q09591	0.0	6239	P48996	1e-165	3.2e-173	2	✓					View
4932	P38989	1e-131	4932	Q12267	1e-144	3.2e-138	1				✓	✓	View
9606	O95347	1e-163	9606	Q9NTJ3	2e-98	4.5e-131	1	✓			✓	✓	View
4896	P41003	1e-129	4896	P41004	4e-77	2.0e-103	1	✓				✓	View
4932	P47037	3e-40	4932	Q12267	1e-144	1.7e-92	2	✓	✓	✓	✓	✓	View
4932	P38989	1e-131	4932	P32908	2e-21	1.4e-76	3	✓	✓	✓	✓	✓	View

C

Couple of GO annotaions			Couple of Pfam domains			Couple of Interpro annotations		
Q09591	Q20060	Ratio	Q09591	Q20060	Ratio	Q09591	Q20060	Ratio
GO:0005524 F:ATP binding	GO:0005524 F:ATP binding	**1.00**	PF02463 SMC_N x1	PF02463 SMC_N x1	**1.00**	IPR010935 SMC hinge	IPR010935 SMC hinge	**1.00**
GO:0005515 F:protein binding	GO:0005515 F:protein binding	**0.86**	PF02463 SMC_N x1	PF06470 SMC_hinge x1	**1.00**	IPR010935 SMC hinge	IPR003395 RecF/RecN/SMC N	**1.00**
GO:0005515 F:protein binding	GO:0005524 F:ATP binding	**0.86**	PF06470 SMC_hinge x1	PF02463 SMC_N x1	**1.00**	IPR003395 RecF/RecN/SMC N	IPR010935 SMC hinge	**1.00**
GO:0005524 F:ATP binding	GO:0005515 F:protein binding	**0.86**	PF06470 SMC_hinge x1	PF06470 SMC_hinge x1	**1.00**	IPR003395 RecF/RecN/SMC N	IPR003395 RecF/RecN/SMC N	**1.00**
GO:0051301 P:cell division	GO:0007076 P:mitotic chromosome condensation	**0.71**						
GO:0007076 P:mitotic chromosome condensation	GO:0007076 P:mitotic chromosome condensation	**0.71**						

Figure 9. The PPISearch server search results using proteins MIX-1 and SMC-4 of *Caenorhabditis elegans* as the query. (**A**) The user interface for assignments of query protein sequences and *E*-value thresholds of homologs and homologous PPIs. (**B**) Homologous PPIs of MIX-1–SMC-4 in multiple species and public databases. (**C**) Conserved protein functions (GO terms) and domain-domain pairs (Pfam and InterPro) of homologous PPIs with a conservation ratio ≥ 0.6.

3.3 Example analysis of homologous PPI search

3.3.1 σ1A-adaptin and γ1-adaptin

Figure 2C 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[26]. According to the crystal structure (PDB code 1W63)[27], 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 2E) — 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[27].

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 2D). 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.

3.3.2 MIX-1 and SMC-4

Mitotic chromosome and X-chromosome-associated protein (MIX-1, Q09591) and structural maintenance of chromosomes protein 4 (SMC-4, Q20060) of *Caenorhabditis elegans* are members of SMC protein family, and are required for mitotic chromosome segregation[28]. Both MIX-1 and SMC-4 are essential components in forming the condensin complex for interphase chromatin to convert into mitotic-like condense chromosomes[28, 29]. Using *C. elegans* MIX-1 and SMC-4 as the query protein pair and J_E is set to 10^{-40}, the PPISearch server found seven homologous interactions from annotated PPI databases (Figure 9B). These seven homologous PPIs are consistently SMC–SMC protein interactions, including SMC-2–SMC-4, SMC-3–SMC-4 and SMC-2–SMC-1, in four species. Among these homologous PPIs, two PPIs, Q95347-Q9NTJ3 (*Homo sapiens*) and P38989-Q12267 (*Saccharomyces cerevisiae*), are orthologous interactions of the query MIX-1–SMC-4[23].

These seven homologous PPIs of MIX-1 and SMC-4 include 136 GO term pairs. Among these GO terms, the CRF ratios of four GO MF term pairs and two GO BP term pairs exceed 0.6 (Figure 9C). These six GO term pairs are consistent with the term-pair combinations of MIX-1 and SMC-4. For example, MIX-1 and SMC-4 have the same two GO MF annotations, protein binding (GO:0005515) and ATP-binding (GO:0005524). Additionally, these seven homologous PPIs contain four DDPs with CRD ratios of 1.0. These four DDPs, PF02463-PF02463, PF06470-PF02463, PF02463-PF06470 and PF06470-PF06470, are recorded in *i*Pfam[20] and are consistent with the query pair. The hinge-hinge interaction (PF02463-PF02463) is experimentally proved, and is conserved in the eukaryotic SMC-2–SMC-4 heterodimer[30]. These analytical results reveal that the PPISearch server is able to identify homologous PPIs that share conserved DDPs and MFPs with the query.

3.4 Discussion

3.4.1 Example analysis for giving more insights into PPI family

In above content, we brought up the concept of homologous PPIs, and give statistic evidence and biological examples to support it. At next step, we will provide more evidence to verify the homologous PPIs identified by our methodolgy. For this purpose, we will verify this issue based on four views: (1) domain composition of PPIs, (2) biological functions of PPIs, (3) the locations of PPIs in pathways, and (4) PPIs in manually curated complexes. In other words, we assume that if a PPI is "homologous" to another PPI, they have the same specific function, interacting domains, and they are experimentally identified in the same pathway and/or in the same protein complex.

The first two views have been used to evaluate the concept of homologous PPIs through Pfam annotations and GO terms. Currently, we are starting to gain insights into homology of PPIs by the last two views. Preliminarily, we use components of the transforming growth factor β (TGF-B) system as an example to test our assumption.

Nearly 30 members of the TGF-B family have been described in human, and many orthologs are known in mouse and other vertebrates[31]. The family is divided into two general branches, the BMP/GDF and TGF-B/Activin/Nodal branches, whose members have diverse, while often complementary, effects[31]. TGF-Bs are potent fibrotic factors responsible for the synthesis of extracellular matrix. TGF-Bs act through the TGF-B type I and type II receptors (TGFBR-1 and TGFBR-2) to activate intracellular mediators, such as Smad proteins, the p38 mitogen-activated protein kinase (MAPK), and the extracellular signal-regulated kinase pathway[32]. Figure 10A shows two branches of the Smad pathway mediate signaling by the two main groups of TGF-B family agonists. The BMPs and related GDFs, as well as AMH/MIS,

trigger receptors that signal through Smads 1, 5, and 8. The TGF-Bs, Activins, and Nodals (and the Nodal-related Xnr factors from *Xenopus*) trigger receptors that phosphorylate Smads 2 and 3. Orthologs from fruit fly are presented in red color (Figure 10B). Alternative type I receptor names are: ALK3 (BMPR-IA), ALK4 (ActR-IB), ALK5 (TbR-I) and ALK6 (BMPR-IB). Activins and BMPs share some of their type II receptors, as indicated[31].

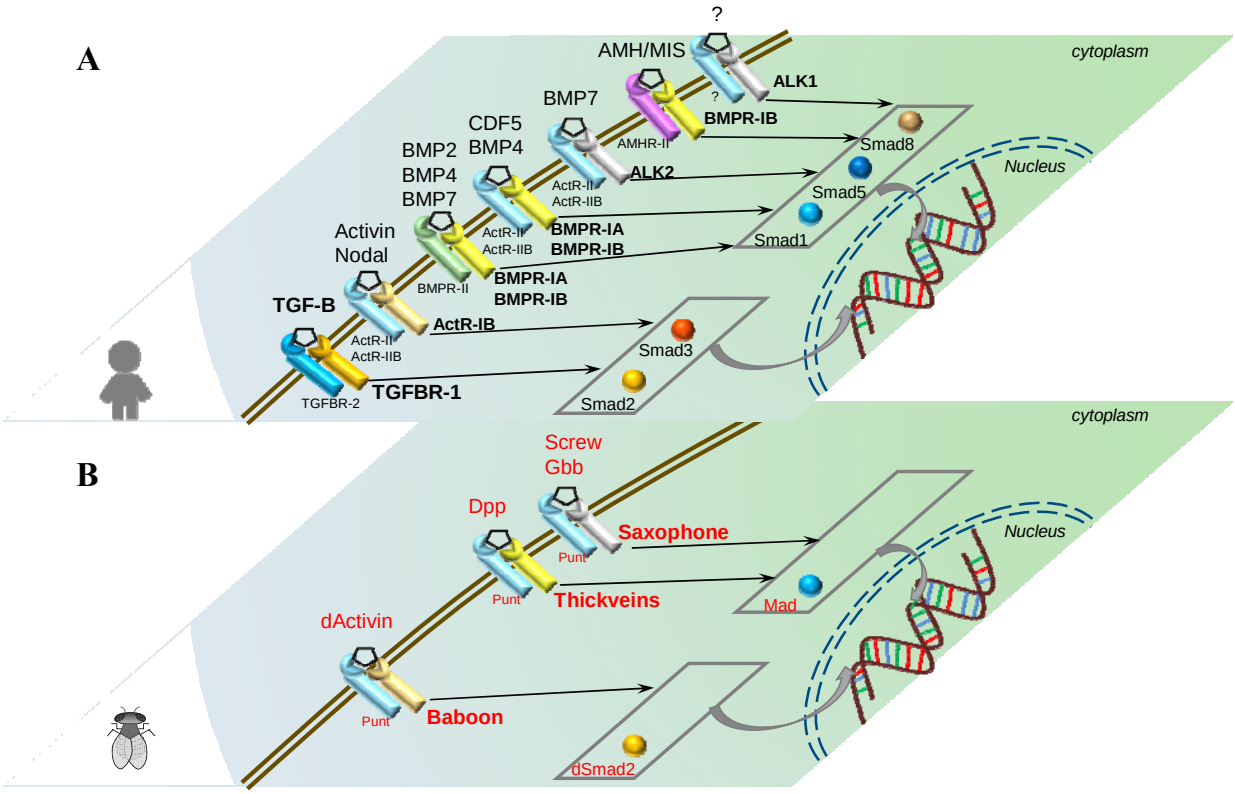


Figure 10. Ligand, receptor, and Smad relationships in the TGF system. (**A**) Two branches of the Smad pathway mediate signaling by the two main groups of TGF-B family agonists. The TGF-Bs, Activins, and Nodals (and the Nodal-related Xnr factors from *Xenopus*) trigger receptors that phosphorylate Smads 2 and 3. The BMPs and related GDFs, as well as AMH/MIS, engage receptors that signal through Smads 1, 5, and 8. (**B**) Orthologs in fruit fly are presented in red color.

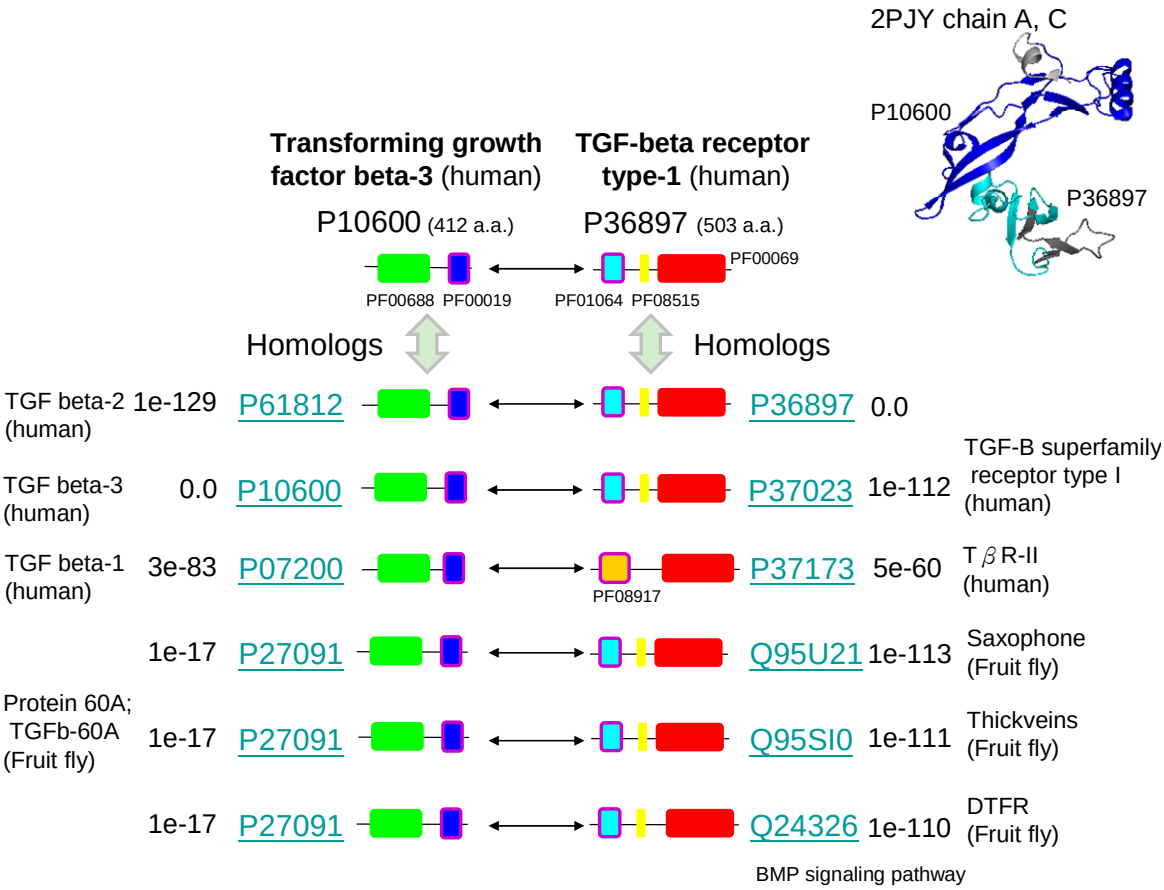


Figure 11. Illustration of cross-species PPI family of TGF-B3—TGFBR-1 interaction. Here we present six of 42 homologous PPIs ($J_E \leq 10^{-40}$) across four species, human, fruit fly, mouse, and chicken. All of the 42 PPIs are protein pairs between TGF-B family and TGFBR family. Interacting domain pairs are circled by purple lines.

The search result of homologous PPIs is showed in Figure 11. PPIsearch found 42 PPIs from the annotated PPI database. All of the 42 PPIs are protein pairs between members of TGF-B family and TGFBR family. Protein structure of TGF-B3—TGFBR-1 complex (PDB code: 2PJY)[33] gives us the information of interacting domains. TGF-B3 engages with TGFBR-1 with domain PF00019 (TGF-B like domain) and PF01064 (Activin types I and II receptor domain).

In this example, we observed homology of PPIs through the two views: pathway and complex insights. The result shows that the PPIs searched by PPISearch are in the TGF signaling pathway, which is identified by experiments, and in protein complexes, which are evolutionarily homologous to each other. The preliminary observation suggests that the two views may be useful to help us to identify homologous PPIs. In the future, we will collect data sets of pathways (e.g. from KEGG[34]) and complexes (e.g. EcoCyc[35]) to verify our concept on large scale.

3.4.2 A question caused by local sequence alignment

In this study, we supplied the concept of homologous PPIs and inferred the transferability of domain and function pairs. Moreover, we applied the concept to construct a web server, PPISearch. In the process that we evaluated the results of searching homologous PPIs, a question was found. We utilized BLASTP as the fast sequence alignment tool to search potential homologs, however, these search results might be biased by local sequence alignments.

We presented an example to describe this question in Figure 12. Q8INB9 is a RAC serine/threonine-protein kinase of fruit fly with three Pfam domains. In this kinase, a protein kinase domain (red) has 258 amino acids, which covers 42% of the whole sequence length, 611 amino acids. The potential homologs with BLASTP *E*-value $\leq 10^{-10}$ have the protein kinase domain, however, they lose the interacting domain PF00169 (deep blue, pleckstrin homology domain). This example indicated the question of searching potential homologs. The question will be reformed in the future works.

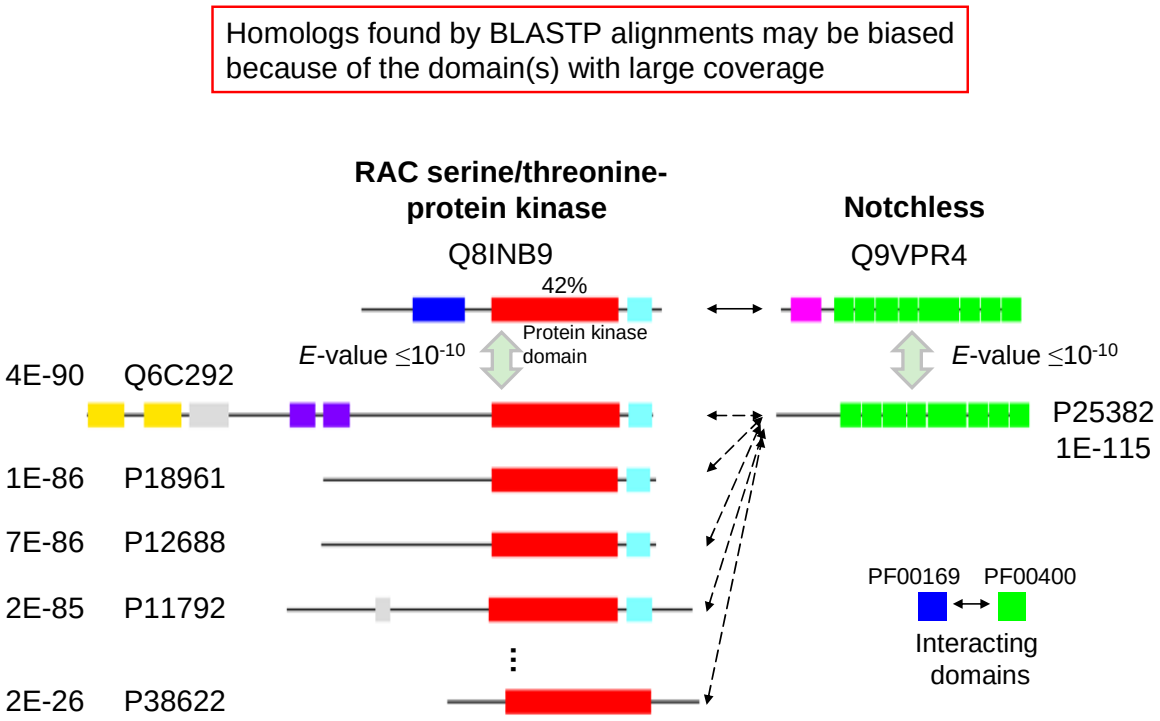


Figure 12. The search result of an interacting protein pair Q8INB9-Q9VPR4. The potential homologs of Q8INB9 keep the protein kinase domain (colore red) but have no PF00169 domain (colored deep blue) to interact with the potential homolog of Q9VPR4.

Additionally, as described in **Section 3.1.3** and **3.1.4**, we discussed possible reasons of why members of a PPI family have different domains and binding models with the query protein. Figure 13 shows an example of our observation. Transcription factor IIIB (TFIIIB), consisting of the TATA-binding protein (TBP), TFIIB-related factor (BRF-1) and BDP-1, is a central component in basal and regulated transcription by RNA polymerase III[36]. In this case, we found that when we searched homologs of yeast BRF-1 by using BLASTP, there were protein sequences with *E*-values $\leq 10^{-10}$, which had no interacting domains (colored by blue), in searching results. This observation suggested that local alignment methods, such as BLASTP,

may get unreliable homologs because of locally similar regions on sequences. The question will be reformed in the future works, too.

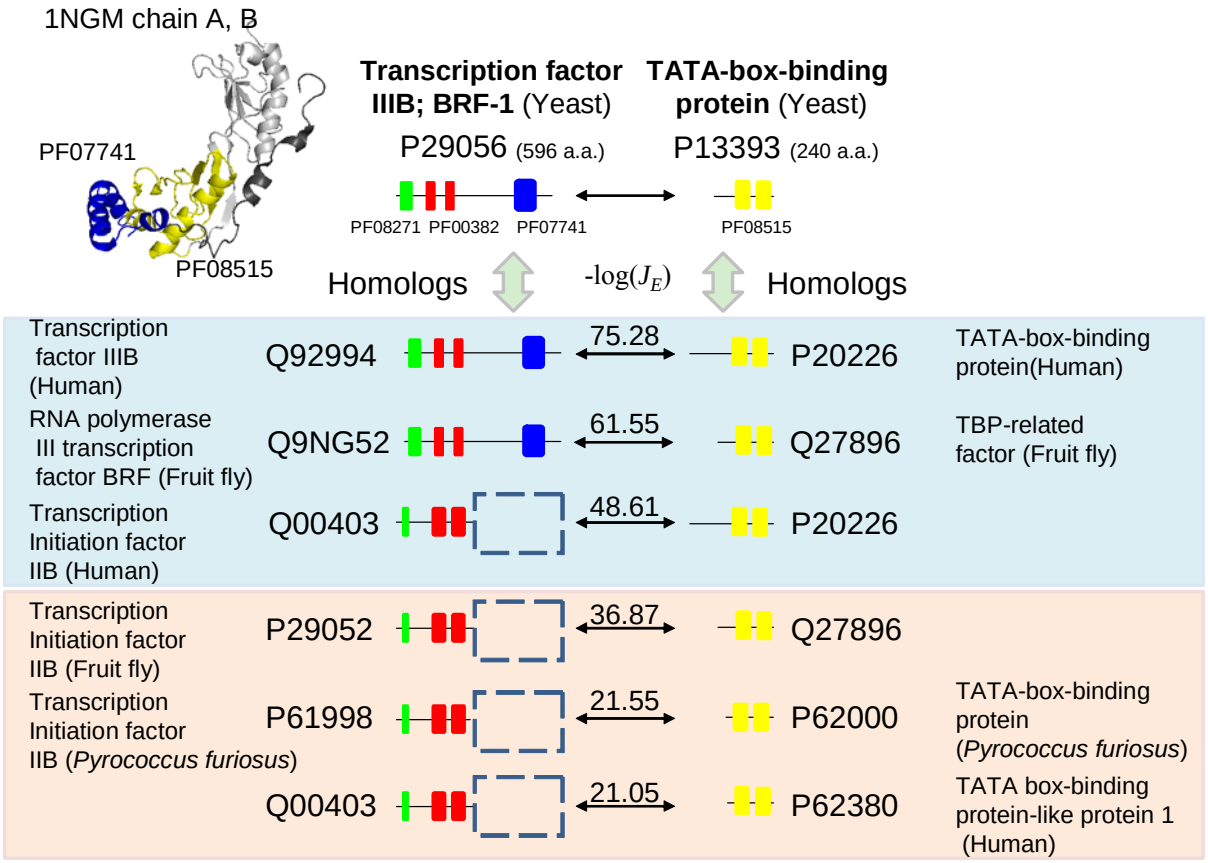


Figure 13. An example of our method selecting proteins without interacting domain (*E*-values $\leq 10^{-10}$) as homologs. The interacting domains of the complex 1NGM chains A and B are colored by blue (PF07741) and yellow (PF08515), respectively. Q00403, the transcription initiation factor IIB of human, does not has the domain PF07741 but has *E*-value $\leq 10^{-10}$.

Chapter 4.

Applications of Homologous Protein-protein Interactions

In this chapter, we applied homologous PPI to cross-species prediction of PPIs, and cross-species network comparisons. Many experimental approaches, for example, yeast two-hybrid system, mass spectroscopy, and tandem affinity purification, have been used to decipher PPI networks. To complement these experimental techniques, a number of computational methods for predicting PPIs, such as PathBLAST[16, 17] and interologs[5, 37] (i.e. conservation of interactions across species), have been developed[19].

The concept of interolog (originally introduced by Walhout *et al.*[37]) combines known PPIs from one or more source species and orthology relationships between the source and target species to predict PPIs in the target species. Yu *et al.* (2004)[5] extended and assessed the concept of interologs to provide a "generalized interolog mapping" method (see below). We considered that our discovery (described in **Section 3.4 Discussion**) can be used to advance the generalized interolog mapping method.

In addition, cross-species network comparison provides insights into the relationships between the proteins of an organism thereby contributing to a better understanding of cellular processes. However, large-scale interaction networks are available for only several model organisms. We considered that the concept of homologous PPI are useful for a systematic transfer of PPI networks between multiple species.

4.1 Cross-species prediction of protein-protein interactions

4.1.1 Background

Protein-protein interactions play an essential role in cellular functions. For rapidly increasing of sequenced genomes, it has been of significant value to provide the approaches of predicting PPIs from one organism (with abundant known interactions) to another organism (with less interaction data). In other words, to reliably transfer PPI annotation from one organism to another[5].

The concept of "interologs" means: If interacting proteins A and B in one organism (source) have interacting orthologs A' and B' in another organism (target), the pair of A-B and A'-B' are called interologs. Operationally, the ortholog of a protein is defined as its best-matching homolog in another organism. Matthews *et al.* (2001)[18] proposed a "best-match mapping" method to predict worm (*C. elegans*) interactions from yeast (*S. cerevisiae*) interactome. This method considered all pairs of best-matching homologs (homologs are defined by BLASTP *E*-value $\leq 10^{-10}$) of interacting yeast proteins as potential interologs.

Additionally, Yu *et al.* (2004)[5] extended and assessed the concept of interologs to provide a "generalized interolog mapping" method. The mapping method regards all pairs of homologs, which have joint similarities (see **Section 4.1.5**) larger than a certain cutoff, as possible interologs. Their results showed that interaction annotation could be reliably transferred between two organisms if a pair of proteins has a joint *E*-value (J_E) $\leq 10^{-70}$.

There are interesting questions in best-match and generalized interolog mapping methods. Firstly, best-match mapping method suffers from low coverage of the total interactome[5], because of using only best matches. For this question, Yu *et al.* (2004) proposed the method of generalized interolog mapping. Secondly, in generalized interolog mapping method, the

homologs of a query protein selected at a certain *E*-value would sometimes be different in subcellular compartment, biological process, or function from the query protein. For example, YLL034C in yeast has a low *E*-value (< 10^{-120}) with protein Q01853 in mouse, but YLL034C has no CDC48 domain (for protein degradation)[38]. The protein pairs having these sequences may be not reliable candidates of interologs.

Orthology of different organisms are usually used in predicting interactions[39]. The third question is that, orthologous protein interactions between two species have various J_E. For example, two protein pairs P47857-P12382 and Q8R317-P40142 in mouse have orthologous interactions YMR205C-YGR240C and YMR276W-YBR117C with J_E =10^{-171} and 10^{-27} in yeast, respectively. In other words, a certain cutoff would usually lose part of orthologous interactions.

To improve these three questions, we preliminarily propose a new "ranked-based interolog mapping" method for predicting protein-protein interactions between species. This method uses only part, not all, of homologs of interacting proteins to gather possible interologs.

4.1.2 Results and discussion

4.1.2.1 Accuracy of rank-based interolog mapping

For practicability of approach to predict interactions, we wish to develop a method which has reliable predicting accuracy and acceptable coverage. In this preliminary study, we propose a new "rank-based interolog mapping" method. This method looses the best-match mapping to get a higher coverage of the total interactome. On the other hand, this method selects part, not all, of homologs in the target organism to amend the two questions of generalized interolog mapping.

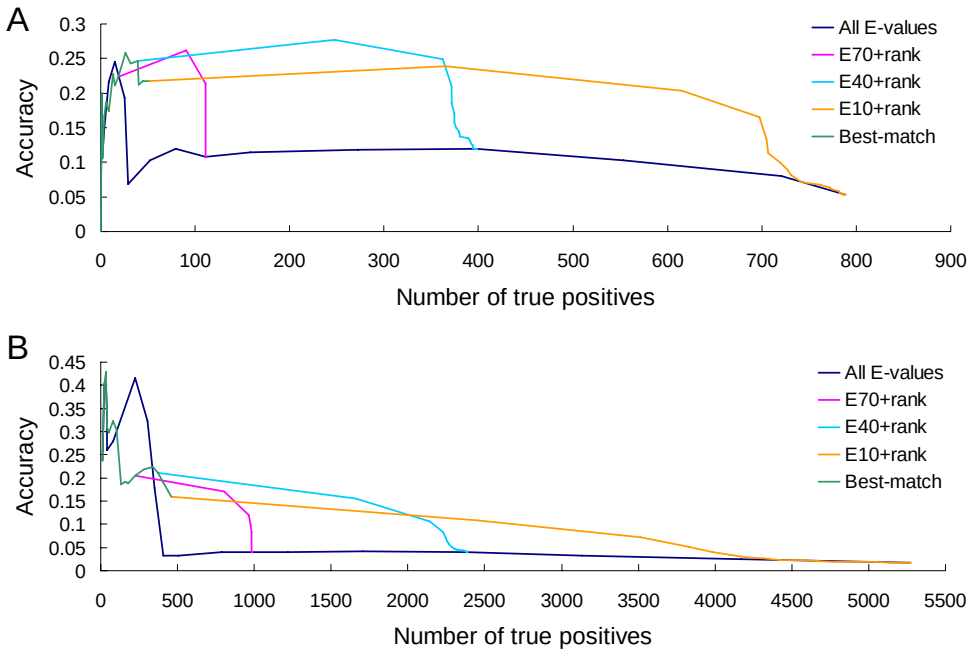


Figure 14. The comparison of accuracy of rank-based interolog (yellow, blue, and pink lines), best-match (green line), and generalized interolog mapping (deep blue line) methods in (**A**) worm-yeast mapping and (**B**) the four mappings. "E10+rank", "E40+rank", and "E70+rank" mean $Acc(10^{-10},R)$, $Acc(10^{-40},R)$, and $Acc(10^{-70},R)$, $R \in$ [1,5,10,15,20,25,30,35,40,45,50,55, 60,65,70,75,80,85,90,95,100, 'All'], respectively.

First, we map only worm interactions onto the yeast genome. We assess the predicting accuracy of our method, best-match and generalized interolog mapping against sets of gold standard positives ***P*** and negatives ***N*** (see **Section 4.1.5**). Figure 14A shows the relationship between accuracy and coverage in the worm-yeast mapping. The blue line indicates the accuracy of generalized interolog mapping from $J_E \leq 10^{-190}$ to 10^{-10}. The green line indicates the accuracy of best-match mapping at $J_E \leq 10^{-10}$. There are three clear observations:

1. While selecting only pairs of top ***R*** homologs (see **Section 4.1.5**) as candidate inter-actions (i.e. rank-based interologs), the accuracy would be usually better than the accuracy of generalized interolog mapping. For example, the purple line consists of plots $Acc(10^{-70},\boldsymbol{R})$,

$\boldsymbol{R} \in$ [1,5,10,15,20,25,30,35,40, 45,50,55,60,65,70,75,80,85,90,95,100, 'All'], $J_E \leq 10^{-70}$ was used as a good threshold of predicting interactions in Yu *et al.* (2004). The accuracy of $\boldsymbol{R}$ = 1, 5, and 10 are 0.22, 0.26, and 0.21, respectively, are better than $Acc(10^{-70}$, 'All') = 0.11. 'All' means $\boldsymbol{R}$ has no limit.

2. If $J_E \leq 10^{-110}$, $Acc(J_E$, 'All') will raise sharply but the number of true positives are ≤ 25. In other words, a very low coverage of yeast interactions.

3. The max number of true positives in best-match mapping is 50 (at $J_E \leq 10^{-10}$). Similarly, there is a low coverage of yeast interactions.

To gather better statistics, we map inter-actions in worm, fruit fly, mouse, and human onto the yeast genome, assessing them against our gold standards. We perform a similar analysis in Figure 14B. The number of true positives dotted in Figure 14B is sum of the true positives in worm-yeast, fly-yeast, mouse-yeast, and human-yeast mappings. The accuracy is calculated by sum of the true and false positives in the four mapping processes.

In Figure 14B, the comparison among rank-based interolog, best-match, and generalized interolog mapping is similar to that in Figure 14A. The accuracy of $\boldsymbol{R}$ = 1, 5, and 10 are 0.21, 0.17, and 0.12, respectively, are better than $Acc(10^{-70}$, 'All') = 0.04.

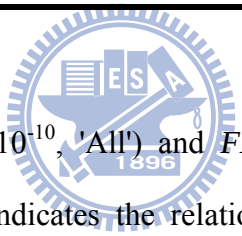
4.1.2.2 Functional similarity between homologous protein pairs

For quantitatively assessing the unreliable homologous pairs in rank-based interologs, we construct sets of ***P'*** and ***N'*** (see **Section 4.1.5**). Genome Ontology (GO) consortium provides a standardized vocabulary, in which three structured ontologies have been proposed, which allow the description of molecular function (MF), biological process (BP) and cellular component (CC)[40]. This annotation particularly allows for assessing the functional similarity of genes or

their products. Based on Wu *et al.* (2006)[41], we calculate the functional similarities between query (in the four organisms) and target (in yeast) interactions by using GO annotations. However, not all of protein sequences have GO annotations. Table 2 shows the percentage of *TP*'(10^{-10}, 'All') and *FP*'(10^{-10}, 'All') with the terms of CC and BP ontologies in each organism. Here *TP*'(10^{-10}, 'All') = *TP*(10^{-10}, 'All') ∩ ***P'*** and *FP*'(10^{-10}, 'All') = *FP*(10^{-10}, 'All') ∩ ***N'***.

Table 2. Comparison of true and false positives selected by J_E and that evaluated by CC and BP annotations

Species	***TP*(10^{-10},** 'All')	***FP*(10^{-10},** 'All')	***TP'*(10^{-10},** 'All')	***FP'*(10^{-10},** 'All')
Worm	788	13971	412 (52.3%)	2778 (19.9%)
Fly	780	73235	362 (46.4%)	23148 (31.6%)
Mouse	912	37636	685 (75.1%)	27770 (73.8%)
Human	2790	187149	1661 (59.5%)	128752 (68.8%)



The statistics of recall of *TP*'(10^{-10}, 'All') and *FP*'(10^{-10}, 'All') in four mappings are showed in Figure 15. Figure 15A indicates the relationship between recall and rank. For example, in worm-yeast mapping, the recall of *TP*'(10^{-10}, 'All') at ***R*** = 1, 5, 10 are 5.1% (21/412), 59.0% (243/412), and 98.3% (405/412), respectively. At ***R*** = 1, 5, 10, the recall of *FP*'(10^{-10}, 'All') is 0.5% (15/2778), 6.7% (185/2778), and 14.8% (411/2778). There are similar trends in the four mappings from the source organisms to yeast.

Otherwise, there is no given J_E could satisfy the demands together: High recall of true positives and low recall of false positives. For example, the recall of *FP*'(10^{-10}, 'All') is 16.3% at $J_E \leq 10^{-40}$, near that at ***R*** = 10, but the recall of *TP*'(10^{-10}, 'All') is only 58.3%. At $J_E \leq 10^{-40}$, the recall of true and false positives are 10.4% and 2.6%, respectively. This result suggests that rank-based mapping method could predict more reliable interactions under a given percentage of false positives than best-match and generalized interolog mapping methods.

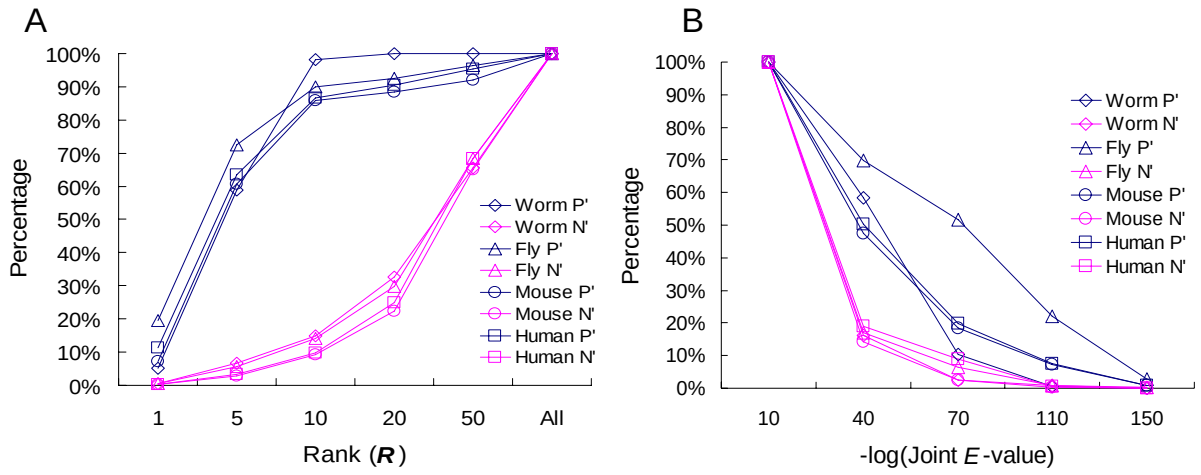


Figure 15. The relationship between recall of *TP*'(10^{-10}, 'All') and *FP*'(10^{-10}, 'All') against (**A**) rank and (**B**) J_E. *TP*'(10^{-10}, 'All') and *FP*'(10^{-10}, 'All') of each mapping are represented by blue and pink solid lines, respectively.

4.1.2.3 Orthologous interactions

Figure 16A and 16B shows the distribution of orthologous interactions against rank and J_E. The total number of orthologous interactions of four mappings is 1,626. Obviously, the orthologous interactions between four organisms and yeast concentrate in top-1—top-5 (totally 99.4%, 1616/ 1626; If top-1—top-10, 99.8%, 1622/1626) but spread at various J_E (e.g. 6.9% in $10^{-40} < J_E \leq 10^{-50}$ and 10.5% in $10^{-70} < J_E \leq 10^{-80}$). This result supplies two suggestions: First, best-match mapping may be not good because it will lose ~44% of orthologous interactions. Second, generalized interolog mapping with any given J_E would lose part of orthologous interactions. Although loosing J_E could raise the coverage of orthologous interactions, the false positives would increase sharply. Our rank-based interolog mapping method could supply higher coverage of orthologous interactions and acceptable quantity of false positives.

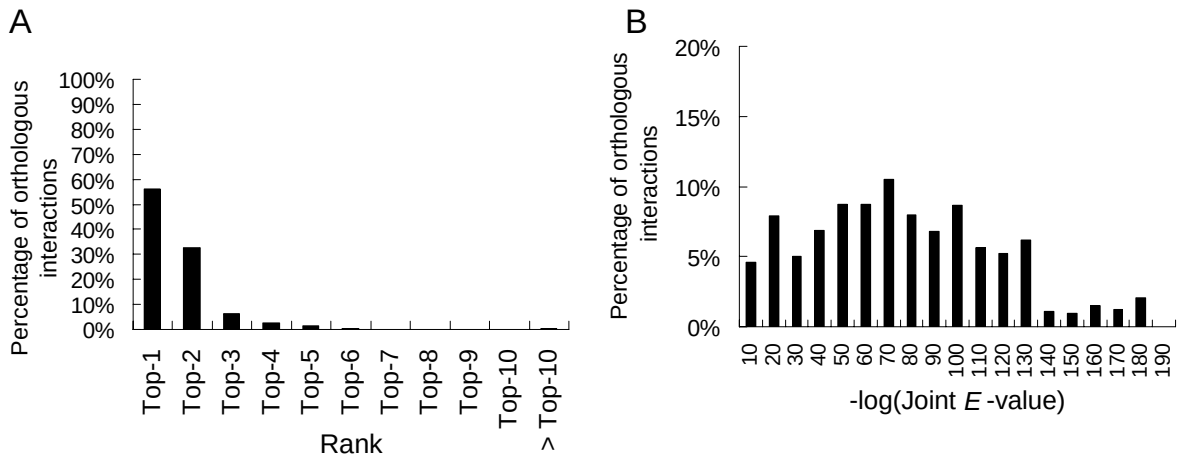


Figure 16. Distribution of total orthologous inter-actions in the four mappings against (**A**) rank and (**B**) J_E.

4.1.3 Discussion

4.1.3.1 Case analysis

We use two cases to explain why rank-based interolog mapping method could work. In first case, the query interaction P43686-P62195 in human has 231 pairs of possible homologs (*E*-value $\leq 10^{-10}$) in yeast. P43686 and P62195 are 26S protease regulatory subunit 6B and 8 of proteasome, respectively[42]. Figure 17A shows that all of 15 true positives ($0.8 < RSS^{GO}_{A-B,A'-B'} \leq 1.0$) are $\leq$ top 10. 97% (29/30) of true and false positives with $RSS^{GO}_{A-B,A'-B'} \leq 0.8$ are out of top 10. The rank of each pair is calculated in the same way described in Figure 16. Comparing to Figure 17B, reliable true positives spread in $10^{-170} \leq J_E \leq 10^{-40}$, this suggests that any given J_E would lose part of true positives.

Similarly, Figure 17C and 17D represent another case, interaction O16368-Q9GZH5 in worm. O16368 is 26S protease regulatory subunit 4. Q9GZH5 is non-ATPase protein 1 of proteasome regulatory particle [43]. All of 49 true positives ($0.8 < RSS^{GO}_{A-B,A'-B'} \leq 1.0$) are $\leq$ top 10.

These reliable true positives spread in $10^{-130} \leq J_E \leq 10^{-30}$. In this case, there are no true and false positives with $RSS^{GO}_{A-B,A'-B'} \leq 0.8$.

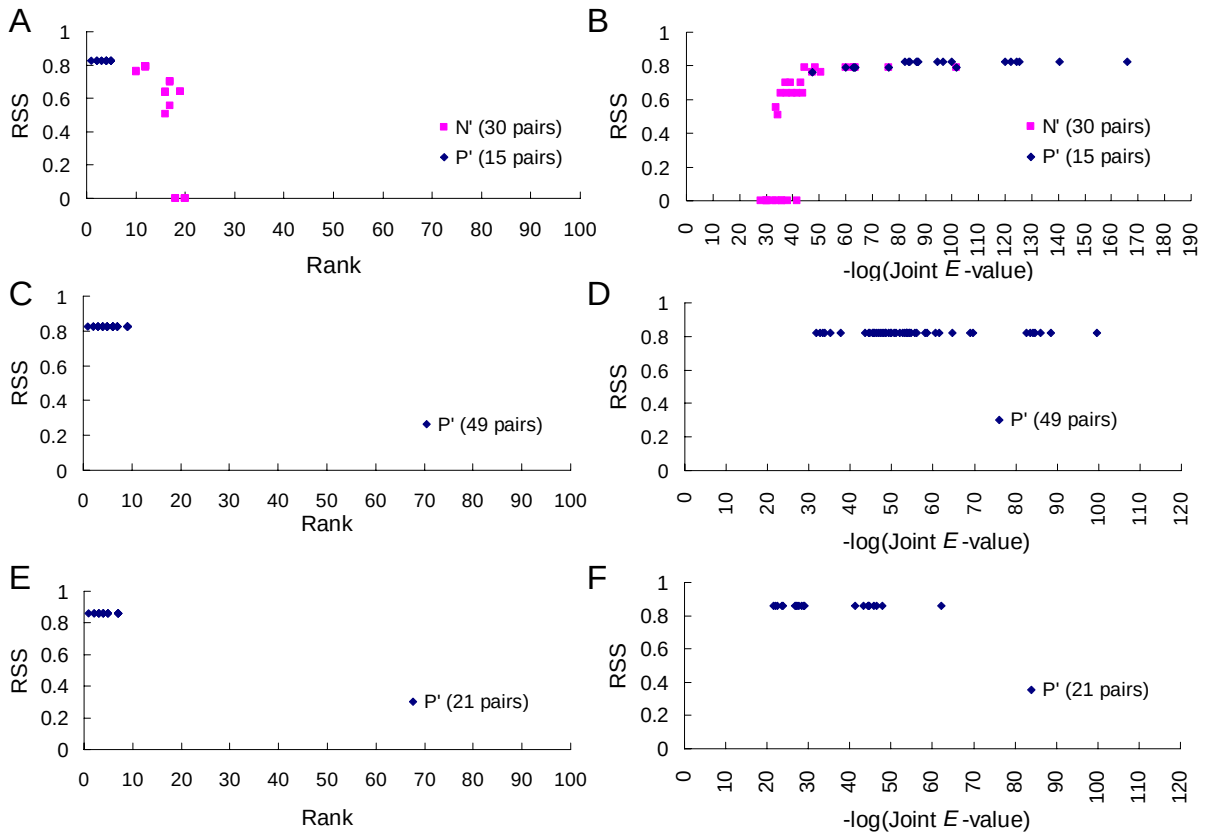


Figure 17. Three cases of rank-based interolog mapping. (A) and (B) show the *TP'*(10^{-10}, 'All') (colored blue) and *FP'*(10^{-10}, 'All') (colored pink) of the query interaction P43686-P62195 in human. Similarly, (C) and (D) show the *TP'*(10^{-10}, 'All') and *FP'*(10^{-10}, 'All') of the query interaction O17071-Q09583 in worm. (E) and (F) show the *TP'*(10^{-10}, 'All') and *FP'*(10^{-10}, 'All') of the query interaction O44156-Q27488 in worm. RSS is $RSS^{GO}_{A-B,A'-B'}$.

4.1.3.2 Three types of good predictions

We classify our predictions between organisms into three types. As $J_E \leq 10^{-70}$ was considered as a good threshold for predicting interactions, we represent the advantages of rank-based interologs in detail at $J_E \leq 10^{-70}$.

First, the true and false positives of a query interaction have various J_E. This suggests that any given J_E, such as 10^{-70}, may be not a good cutoff. For example, O16368-Q9GZH5 is a known interaction in worm. They have 21 and 2 possible homologs (E-value $\leq 10^{-10}$) in yeast. In the 21 × 2 = 42 homolog pairs, the total number of true positives (i.e. $|TP(10^{-10}, \text{'All'})|$) and false positives (i.e. $|FP(10^{-10}, \text{'All'})|$) are 14 and 14, respectively. In this case, the generalized interologs with $J_E \leq 10^{-70}$ have $|TP(10^{-70}, \text{'All'})| = 8$ and $|FP(10^{-70}, \text{'All'})| = 7$, the predicting accuracy is 0.53. For the rank-based interologs with $\boldsymbol{R} = 5$ and 10, $|TP(10^{-10}, 5)| = 10$ and $|FP(10^{-10}, 5)| = 0$, $|TP(10^{-10}, 10)| = 14$ and $FP(10^{-10}, 10) = 4$. The $Acc(10^{-10}, 5)$ and $Acc(10^{-10}, 10)$ are 1.0 and 0.78, both better than $Acc(10^{-70}, \text{'All'})$.

In cases of the second type, true positives of a query interaction have J_E from higher to lower than 10^{-70} and there are no or few false positives. Most of true positives are ranked in top $\boldsymbol{R}$. For this type, the query interaction O17071-Q09583 in worm is used as an example. O17071 has 21 homologs (E-value $\leq 10^{-10}$) and Q09583 has 7 homologs (E-value $\leq 10^{-10}$) in yeast, respectively. $|TP(10^{-10}, \text{'All'})|$ is 49 in the total 147 homolog pairs. In this case, $|TP(10^{-70}, \text{'All'})| = 7$ and $|FP(10^{-70}, \text{'All'})| = 0$, the predicting accuracy of using threshold $J_E \leq 10^{-70}$ is 0.14. Otherwise, $|TP(10^{-40}, \text{'All'})| = 43$ and $|FP(10^{-40}, \text{'All'})| = 0$, $|TP(10^{-10}, 5)| = 25$ and $|FP(10^{-10}, 5)| = 0$, $|TP(10^{-10}, 10)| = 49$ and $|FP(10^{-10}, 10)| = 0$. The accuracy of $\boldsymbol{R} = 5$ and 10 are 0.51 and 1.0, respectively.

Third, all pairs of $TP(10^{-10}, \text{'All'})$ of a query interaction have J_E higher than 10^{-70}. For this type, we get interaction O44156-Q27488 as an example. Both O44156 and Q27488 have 7

possible homologs (E-value $\leq 10^{-10}$) in yeast. The minimum J_E of all pairs of homologs is 7.7×10^{-63}. In other words, there is no true positives has $J_E \leq 10^{-70}$, $|TP(10^{-70}, \text{'All'})| = 0$. Otherwise, $|TP(10^{-10}, 5)| = 15$ and $|FP(10^{-10}, 5)| = 0$, $|TP(10^{-10}, 10)| = 21$ and $|FP(10^{-10}, 10)| = 0$, the accuracy of rank-based interolog mapping method at $\boldsymbol{R} = 5$ and 10 are 0.71 and 1.0, respectively.

4.1.4 Summary

In this preliminary study, we propose a rank-based interolog mapping method for predicting interactions across species. This method looses best-match mapping method to get a higher coverage of the total interactome. On the other hand, this method selects part, not all, of homologs in the target organism to amend generalized interolog mapping method.

Four mappings of worm-yeast, fly-yeast, mouse-yeast, and human-yeast are included in our preliminary study. In general, rank-based mapping method could predict more reliable interactions (including positives annotated by CC and BP ontologies and orthologous interactions) under a given percentage of false positives than best-match and generalized interolog mapping methods.

4.1.5 Methods

4.1.5.1 Rank-based interolog mapping

Interolog mapping is a process that maps interactions in the source organism onto the target organism to predict possible interactions. To address the three questions of best-match and generalized interolog mapping described above, we introduce a new “rank-based interolog mapping” method using part of possible homologs of interacting proteins. Operationally,

homologs could be defined as the proteins having an *E*-value $\leq 10^{-10}$ from BLASTP[18, 44]. An overview of the rank-based interolog mapping is depicted in Figure 18.

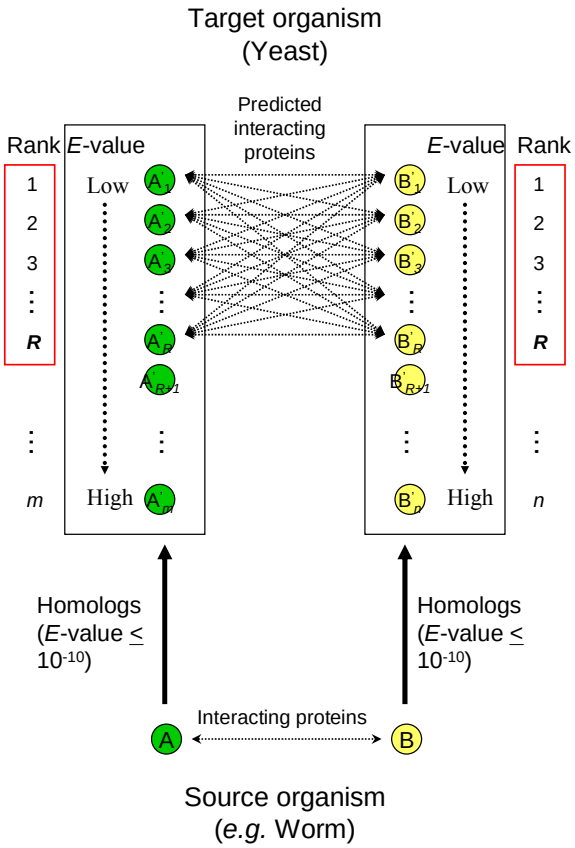


Figure 18. Schematic illustration of rank-based interolog mapping method. Proteins A_1', A'_2,…, A'_m and B_1', B'_2,…, B'_n are possible homologs (*E*-value $\leq 10^{-10}$) of proteins A and B in the source organism, respectively. All possible pairs between homologs A'_1,…, A'_R and B'_1,…, B'_R are called ranked-based interologs.

The steps are described as following:

1. For any given protein in the source organism (e.g. worm), we collect all of its homologs by BLASTP *E*-value $\leq 10^{-10}$.

2. These possible homologs of proteins A and B in the target organism (yeast) are ranked by

their *E*-values from low to high (i.e. from 0 to 10^{-10}), respectively.

3. These homologs ranked in top ***R*** are selected to pair with each other. These possible protein pairs between the homologs $A'_1,\ldots, A'_R$ and $B'_1 \ldots B'_R$ are called ranked-based interologs. Otherwise, the all protein pairs between the homologs $A'_1,\ldots, A'_m$ and $B'_1,\ldots, B'_n$ are generalized interologs.

The best-match mapping method considers pairs between the best-matching homologs as the candidates of interaction[18]. The generalized interolog mapping method uses all pairs of homologs, which have joint similarities larger than a certain cutoff, to find possible interactions in the target organism[5]. In this preliminary study, we consider the protein pairs between the top ***R*** possible homologs as the candidates of interaction in the target organism.

4.1.5.2 Source data sets

To assess the rank-based interolog mapping method, we need source organisms with known interaction data. In this preliminary study, worm, fruit fly, mouse, and human are used as source organisms. We collect the interactions of these four organisms recorded in IntAct database[45] (Table 3). We then map these interactions onto the yeast genome. The protein sequences of these four source organisms and the target organism yeast are from SWISS-PROT and SGD database[46], respectively.

4.1.5.3 Gold standard target data sets

Set of gold standard positives (P)

To assess the performance of interolog mapping, we need a collection of known interactions as positives in the target organism. Previously, a data set derived from the MIPS complex catalog, which contains 8,250 unique interacting protein pairs, has been used as a standard reference for known interactions[5, 47, 48]. We also consider the MIPS interactions as gold standard positives in this preliminary study.

Table 2. Source data sets derived from IntAct

Species	Worm	Human	Fly	Mouse	Total
Number of PPIs	4,653	18,943	19,774	2,728	46,098

Set of gold standard negatives (N)

A set of negatives (i.e. non-interacting proteins) in yeast is necessary for evaluating our method. Jansen *et al.* (2003)[49] considered pairs of proteins in different subcellular compartments as good estimates for non-interacting proteins. This set has 2,708,746 such protein pairs. Therefore, we find that 3,689 interactions in this set are also recorded in the core database of DIP[50]. We exclude these interactions and take 2,705,057 protein pairs as the set of gold standard negatives in this preliminary study.

4.1.5.4 Accuracy of interolog mapping

We assess the predicting accuracy of our method, best-match and generalized interolog mapping against ***P*** and ***N*** in yeast. The accuracy (*Acc*) is calculated as following:

$$Acc(J,R) = \frac{TP(J,R)}{TP(J,R) + FP(J,R)}$$

In this equation, $TP(J,R) = H(J,R) \cap \boldsymbol{P}$, $FP(J,R) = H(J,R) \cap \boldsymbol{N}$. $H(J,R)$ means the sets of rank-based interologs, best-matching homologous pairs, or generalized interologs in yeast at a certain cutoff. For example, in rank-based interolog mapping, J is a given joint E-value (see below) and R is the number of homologs selected by ranking (i.e. top ***R***). Otherwise, in generalized interolog mapping, J is a certain joint E-value and ***R*** has no limits. $|TP(J,R)|$ and $|FP(J,R)|$ are the number of true and false positives at a given J and ***R***.

4.1.5.5 Joint E-value (J_E)

J_E is the geometric means of *E*-values for the two pairs of interacting proteins. For example, if the *E*-values of A-A' and B-B' are $E_{A\text{-}A'}$ and $E_{B\text{-}B'}$, J_E between pairs A-B and A'-B' is $J_E = \sqrt{E_{A-A'} \times E_{B-B'}}$.

4.1.5.6 GO similarity measure

We assume that if a pair A'-B' in yeast is a reliable homologous pairs of A-B, A-A' and B-B' would be in similar subcellular compartment, biological process and have similar molecular function. Wu *et al.* (2006) proposed a method to measure semantic similarity between two proteins by using CC and BP annotations. Based on the relative specificity similarities (*RSS*) defined by their method, we calculate the similarity in cellular component and biological process between a protein pair A-B and its rank-based interologs A'-B' ($RSS^{GO}_{A-B,A'-B'}$). *RSS* values for the CC (RSS^{CC}) and BP (RSS^{BP}) ontologies mean the similarity of CC and BP terms of a given protein pair, respectively. The values are between 0 and 1.0. The equations we used are as follows.

$$RSS^{CC}_{A-B,A'-B'} = \sqrt{RSS^{CC}_{A-A'} \times RSS^{CC}_{B-B'}}$$

$$RSS^{BP}_{A-B,A'-B'} = \sqrt{RSS^{BP}_{A-A'} \times RSS^{BP}_{B-B'}}$$

$$RSS^{GO}_{A-B,A'-B'} = \sqrt{RSS^{CC}_{A-B,A'-B'} \times RSS^{BP}_{A-B,A'-B'}}$$

Wu *et al.* (2006) supplied both three confidence levels of yeast protein pairs annotated in the CC and BP ontologies. Their results showed that 78% interactions of their positive dataset fall into the high-confidence segment of $0.8 < RSS^{CC} \leq 1.0$ and $0.8 < RSS^{BP} \leq 1.0$. They suggested that the highest-confidence segment may contain most yeast protein-protein interactions. We use the thresholds to construct two datasets, ***P'*** and ***N'***. ***P'*** is the interaction

dataset including these protein-protein interactions with $0.8 < RSS^{GO}_{A-B,A'-B'} \leq 1.0$ in ***P***. ***N'*** is the dataset consisted of ***N*** and these interactions of ***P*** but having $RSS^{GO}_{A-B,A'-B'} \leq 0.8$.

4.1.5.7 Orthologous interactions between source and target organisms

We identified the orthologous proteins between the source organisms and yeast by ENSEMBL database (Mar, 2008)[51]. For comparing the coverage of orthologous interactions of our method, best-match and generalized interolog mapping, we identify and count the interacting pairs of orthologs in all pairs of possible homologs (*E*-value $\leq 10^{-10}$). For example, interacting proteins P47857-P12382 in mouse have orthologous interactions YMR205C-YGR240C. In Figure 16A, the label "Top-*n*" is the maximum of ranks of $E_{A\text{-}A}$' and $E_{B\text{-}B}$'. For example, YMR205C and YGR240C are the top 1 and top 2 in the ranking of homologs by *E*-values, respectively. The label of pair YMR205C-YGR240C is $n = \max(1, 2) = 2$.

4.2 Cross-species network comparison by using homologous PPIs

The comparison of biochemical networks across multiple species can be preliminarily studied by the concept of homologous PPIs. 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. Moreover, homologous PPIs may share similar functions and domains. We consider the concept "homologous PPIs", which to our knowledge is firstly proposed, as a starting point to compare and map protein-protein interaction networks across multiple species. 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[12, 13] applied homologous interactions to assess the

reliability of PPIs.

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. For drug discovery, the pathway-level comparison can indicate candidates of key targets (protein or protein-protein interactions) to inhibit a specific mechanism of a disease. As described above, the concept of homologous PPIs is useful to find PPIs which share similar functions and domains, we have applied this concept to search and, even re-build, the corresponding pathways from one organism to another.

We used the pathway of non-small cell lung cancer as an example (Figure 19A). The PPI of Raf (A-Raf proto-oncogene serine/threonine-protein kinase) and MEK (mitogen-activated protein kinase kinase) is key component of the Ras-dependent signaling pathway from receptors to the nucleus[52, 53]. Clinically, inhibition of MEKs suppressed Raf-mediated cellular growth. The three Raf family members, Raf-1, B-Raf, and A-Raf, are highly homologous and originally isolated as a oncogene contributing to cellular transformation and are one of the best characterized Ras effectors to activate the mitogen-activated protein kinase (MAPK) signaling pathway. Raf directly phosphorylates and activates MEK via two conserved serine residues in the kinase activation loop of MEK. The two member of MEK family, MEK-1 and MEK-2, in turn, are dual-specificity threonine and tyrosine kinases that phosphorylate and activate ERKs (extracellular signal-regulated kinases).

Figure 19B shows the homologous PPIs of the query PPI, A-Raf (UniProt Accession number: P10398) and MEK-1 (Q02750). For this query, the PPISearch server identifies a PPI family including five homologous PPIs in human and two homologous PPIs in mouse that belong to Raf family proteins (Raf-1, A-Raf, and B-Raf), which phosphorylate and activate MEK-1 and MEK-2. Moreover, these homologous PPIs share the same Pfam domain

annotation with the query.

Based on this search, we will be able to provide reliable corresponding Ras-dependent signaling pathway in mouse. The Raf family proteins, P10398, P15056, P04049 (human) and P28028 (mouse), can be used to network comparison. And so on, the MEK proteins, Q02750, P36507 (human) and P31938, Q63932 (mouse) can be used to build network comparison, too.

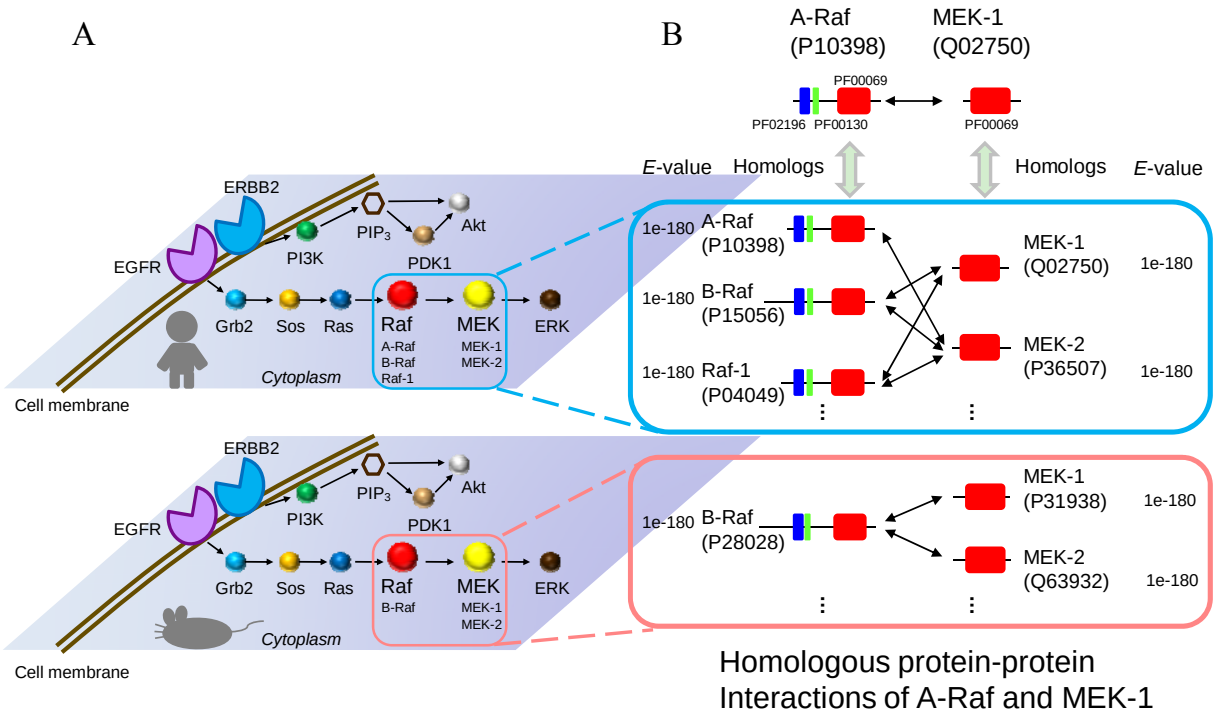


Figure 19. The homologous PPIs of the query PPI, A-Raf (P10398) and MEK-1 (Q02750). (**A**) The diagram of pathways of non-small cell lung cancer in human and mouse. The key component, interaction between Rafs and MEKs, is squared. (**B**) For this query, the PPISearch server identifies a PPI family including five homologous PPIs in human and two homologous PPIs in mouse that belong to Raf family proteins (Raf-1, A-Raf, and B-Raf), which phosphorylate and activate MEK-1 and MEK-2. These homologous PPIs share the same Pfam domain annotation with the query.

Chapter 5.

Conclusion

5.1 Summary

The interactions between proteins are critical to most biological processes. To identify and characterize protein-protein interactions 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, known 3D complexes, and interologs) have been proposed. Some PPI databases, such as IntAct, MIPS, DIP, MINT, and BioGRID 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 the above goal and analysis of newly identified interactions.

To address this issue, we have proposed 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. 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 demonstrate that this server achieves high agreements on interacting domain-domain pairs and function pairs between query protein pairs and their respective homologous PPIs.

This study demonstrated the utility and feasibility of the PPISearch server in identifying homologous PPIs and inferring conserved DDPs and MFPs from PPI families. By allowing users to input a pair of protein sequences, PPISearch is the first server that can identify homologous PPIs from annotated PPI databases and infer transferability of interacting domains and functions between homologous PPIs and a query. Our experimental results demonstrated that the query protein pair and its homologous PPIs achieved high agreement on conserved DDPs and MFPs. We believe that PPISearch is a fast homologous PPIs search server and is able to provide valuable annotations for a newly determined PPI.

5.2 Future works

5.2.1 Directions for future research

There are several directions for future research:

1. For supplying more biological evidence to support the concept of homologous PPIs, we will add the pathway-level and complex-level insights (described in **Section 3.4**).

2. Based on these PPI families constructed by our methodology, we will be able to investigate the evolutionary relationships between PPIs across multiple species.

3. The PPI families constructed by our methodology will be able to supply us with clues to study the evolutionary conservation of PPIs across multiple species.

4. Currently, we have preliminarily got evidence to support the feasibility of applying homologous PPIs to cross-species PPI prediction and network comparison (described in **Sections 4.1** and **4.2**). We will study these two issues more deeply.

5. We will add approaches, such as FASTA, into our methodology to correct the question caused by local sequence alignments (described in **Section 3.4**).

5.2.2 Combination of sequence-based and structure-based interolog mapping

Our laboratory has developed a concept "3D-domain interologs"[4, 54]. We will combine the sequence-based homologous PPIs with the structure-based method of interolog mapping. The detail of the concept of 3D-domain interologs is described as follows.

For studying the mechanisms of PPIs in multiple species, domain-domain interactions, which are regarded as key of PPIs, should be identified. As the rapid increasing of protein structures, to identify interacting domains from three-dimensional (3D) structural complexes is able to study domain-domain interactions. A known 3D structure of interacting proteins provides interacting domains and atomic details for thousands of direct physical interactions. Based on considering interacting domain of interacting protein pair, we have proposed a concept "3D-domain interolog mapping" to improve the generalized interologs mapping. Two physical interacting-domain sequences of a 3D-dimer protein structure are used as the queries to identify its 3D-domain interolog candidates by searching on protein sequences of genomes by utilizing PSI-BLAST. The proteins with both significant sequence similarity and the same interacting domains are considered as 3D-domain homologs forming a homolog family. Here, we define as 3D-domain homologs as: (1) candidates of both alignments with a significant PSI-BLAST *E*-value ($< 10^{-8}$); (2) candidates have 25% domain sequence identity in both sequences in the PSI-BLAST alignment; (3) candidates have 25% sequence identity in both sequences on contacted residues in the PSI-BLAST alignment. The 3D-domain interolog candidates are defined as the all protein pairs between two homolog families derived from two sequences of a structure 3D-dimer (Figure 20). We believe that 3D-domain interolog mapping is able to study

the evolution of the interacting domain through 3D-domain homologous family from multiple species.

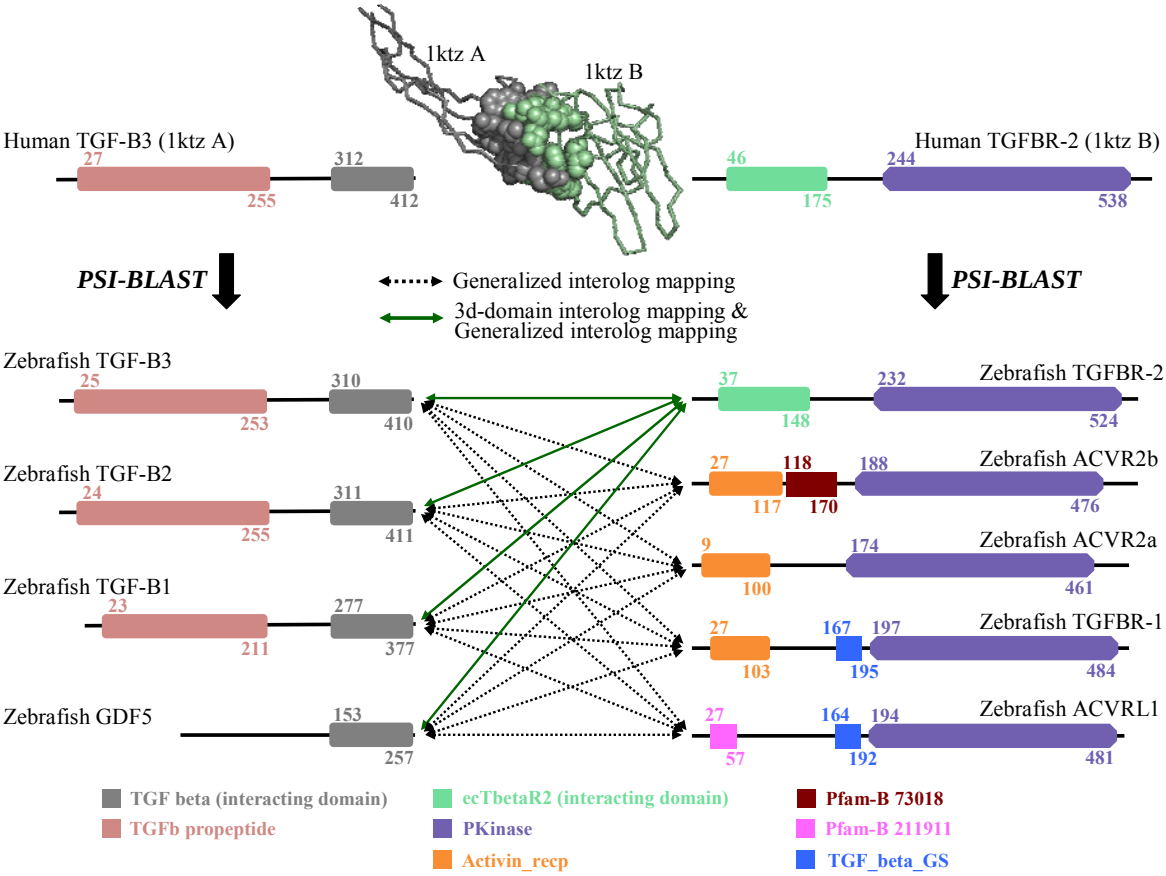


Figure 20. Architecture of 3D-domain interolog mapping. Human TGF-B3 and TGFBR-2 co-crystallize in PDB[55]. Four Zebrafish homologous proteins of Human TGF-B3 found are by PSIBLAST. Likewise, five Zebrafish proteins are homologous to Human TGFBR-2. Through generalized interologs mapping, all possible pairs between the two families are considered as the generalized interologs (show as black and green line with arrows). Moreover, we could find the interacting domains of TGF-B3－TGFBR-2 complex (TGF beta domain is showed as gray and ecTbetaR2 domain is shown as light green) by exploring the co-crystal structure. The pairs of proteins which contain these interacting domains are considered as 3D-domain interologs (show as green line with arrows).

Based on combination of homologous PPIs and 3D-domain interologs, we will develop a new scoring function to model protein interface. The scoring function is

$$E = E_{\text{interacting}} + E_{\text{consensus}} + E_{\text{similarity}}$$

The scoring function is composed of interacting force ($E_{\text{interacting}}$), consensus of residues ($E_{\text{consensus}}$) and template similarity ($E_{\text{similarity}}$). We have applied this function and 3D-domain interologs to measure the interaction changes during evolution and the effect of residue substitution on the binding interface.

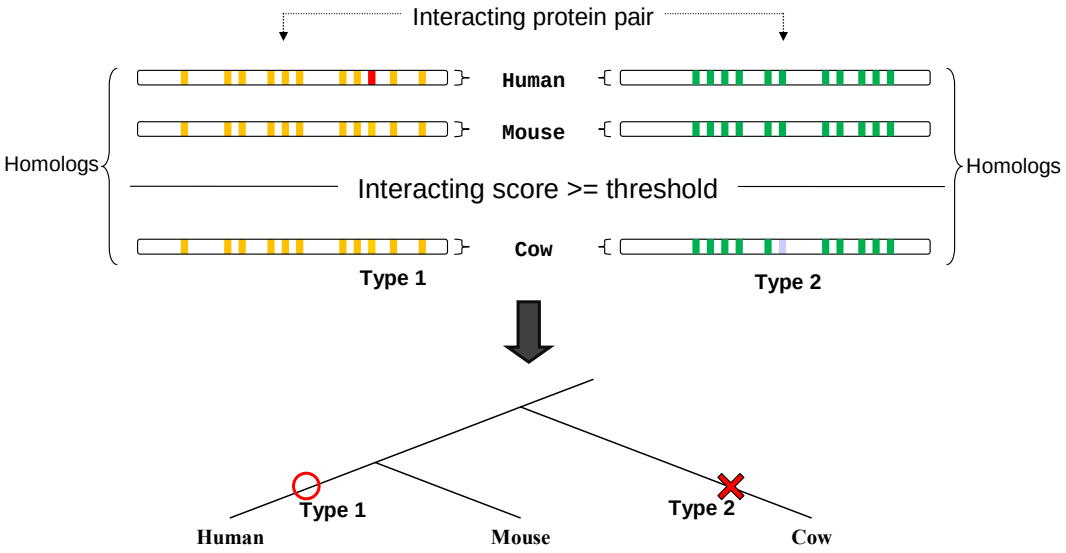


Figure 21. Overview of mutation analysis in protein-protein interactions.

In comparison of biochemical networks across species, protein-protein interactions may be conserved or non-conserved. Figure 21 shows an overview of how we analyze the causes of a protein-protein interaction would keep or lose in different organisms. In our study, we have acquired a reliable threshold of E (= $E_{\text{interacting}}$+$E_{\text{consensus}}$+$E_{\text{similarity}}$) to estimate that two proteins will interact with each other or not. The contacting residues of homologous proteins among organisms are colored by yellow and green.

In Figure 21, the protein-protein interaction exists in human and mouse (≥ threshold) but

not in cow (< threshold). This observation suggests that the mutation in human (colored red) may not disrupt the interaction (called Type 1 mutation), but the mutation in cow (colored blue) may cause the loss of this interaction (Type 2 mutation). This model could help us to perform large-scale analyses of changes in interacting modes and residues among multiple organisms. These analyses will support us to understand the causes of conservation and diversity in protein-protein interaction networks.

References

1. Watson, J. D., Laskowski, R. A. & Thornton, J. M. Predicting protein function from sequence and structural data. Current Opinion in Structural Biology 15, 275-284 (2005).
2. Yang, J.-M. & Tung, C.-H. Protein structure database search and evolutionary classification. Nucleic Acids Research 34, 3646-3659 (2006).
3. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. U. S. A. 96, 4285-4288 (1999).
4. 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).
5. Yu, H. Y. et al. Annotation transfer between genomes: Protein-protein interologs and protein-DNA regulogs. Genome Research 14, 1107-1118 (2004).
6. Shoemaker, B. A. & Panchenko, A. R. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. PLoS Computational Biology 3, 337-344 (2007).
7. Kerrien, S. et al. IntAct - open source resource for molecular interaction data. Nucleic Acids Research 35, D561-D565 (2007).
8. Mewes, H. W. et al. MIPS: analysis and annotation of genome information in 2007. Nucleic Acids Research 36, D196-D201 (2008).
9. Salwinski, L. et al. The Database of Interacting Proteins: 2004 update. Nucleic Acids Research 32, D449-D451 (2004).
10. Chatr-Aryamontri, A. et al. MINT: the molecular INTeraction database. Nucleic Acids Research 35, D572-D574 (2007).
11. Stark, C. et al. BioGRID: a general repository for interaction datasets. Nucleic Acids Research 34, D535-D539 (2006).
12. Patil, A. & Nakamura, H. Filtering high-throughput protein-protein interaction data using a combination of genomic features. BMC Bioinformatics 6, 100-112 (2005).
13. Saeed, R. & Deane, C. An assessment of the uses of homologous interactions. Bioinformatics 24, 689-695 (2008).
14. Scott, M. S. & Barton, G. J. Probabilistic prediction and ranking of human protein-protein interactions. BMC Bioinformatics 8, 239-259 (2007).
15. Michaut, M. et al. InteroPORC: automated inference of highly conserved protein interaction networks. Bioinformatics 24, 1625-1631 (2008).
16. Kelley, B. et al. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc. Natl. Acad. Sci. U. S. A. 100, 11394-11399 (2003).

17. Sharan, R. et al. Conserved patterns of protein interaction in multiple species. Proc. Natl. Acad. Sci. U. S. A. 102, 1974-1979 (2005).
18. 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).
19. Shoemaker, B. A. & Panchenko, A. R. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Computational Biology 3, e43 (2007).
20. Finn, R. D. et al. The Pfam protein families database. Nucleic Acids Research 36, D281-D288 (2008).
21. Hunter, S. et al. InterPro: the integrative protein signature database. Nucleic Acids Research 37, D211-D215 (2009).
22. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. Nature Genetics 25, 25-29 (2000).
23. Kersey, P. et al. Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. Nucleic Acids Research 33, D297-D302 (2005).
24. Andreeva, A. et al. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Research 32, D226-D229 (2004).
25. 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).
26. Bonifacino, J. S. & Traub, L. M. Signals for sorting of transmembrane proteins to endosomes and lysosomes. Annual Review of Biochemistry 72, 395-447 (2003).
27. Heldwein, E. E. et al. Crystal structure of the clathrin adaptor protein 1 core. Proc. Natl. Acad. Sci. U. S. A. 101, 14108-14113 (2004).
28. Lieb, J. D., Albrecht, M. R., Chuang, P. T. & Meyer, B. J. MIX-1: An essential component of the C-elegans mitotic machinery executes x chromosome dosage compensation. Cell 92, 265-277 (1998).
29. Hagstrom, K. A., Holmes, V. F., Cozzarelli, N. R. & Meyer, B. J. C. elegans condensin promotes mitotic chromosome architecture, centromere organization, and sister chromatid segregation during mitosis and meiosis. Genes & Development 16, 729-742 (2002).
30. Hirano, M. & Hirano, T. Hinge-mediated dimerization of SMC protein is essential for its dynamic interaction with DNA. EMBO Journal 21, 5733-5744 (2002).
31. Massague, J., Blain, S. W. & Lo, R. S. TGF-beta signaling in growth control, cancer, and heritable disorders. Cell 103, 295-309 (2000).
32. Laping, N. J. et al. Inhibition of transforming growth factor (TGF)-beta1-induced extracellular matrix with a novel inhibitor of the TGF-beta type I receptor kinase activity: SB-431542. Molecular Pharmacology 62, 58-64 (2002).

33. Groppe, J. et al. Cooperative assembly of TGF-beta superfamily signaling complexes is mediated by two disparate mechanisms and distinct modes of receptor binding. Molecular Cell 29, 1-13 (2008).
34. Kanehisa, M. et al. KEGG for linking genomes to life and the environment. Nucleic Acids Research 36, D480-D484 (2008).
35. Keseler, I. M. et al. EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic Acids Research 33, D334-337 (2005).
36. Juo, Z. S., Kassavetis, G. A., Wang, J., Geiduschek, E. P. & Sigler, P. B. Crystal structure of a transcription factor IIIB core interface ternary complex. Nature 422, 534-539 (2003).
37. Walhout, A. et al. Protein interaction mapping in C. elegans using proteins involved in vulval development. Science 287, 116-122 (2000).
38. Gadal, O. et al. A nuclear AAA-type ATPase (Rix7p) is required for biogenesis and nuclear export of 60S ribosomal subunits. EMBO Journal 20, 3695-3704 (2001).
39. Tirosh, I. & Barkai, N. Computational verification of protein-protein interactions by orthologous co-expression. BMC Bioinformatics 6, 40 (2005).
40. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25-29 (2000).
41. Wu, X., Zhu, L., Guo, J., Zhang, D. & Lin, K. Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. Nucleic Acids Research 34, 2137-2150 (2006).
42. Davy, A. et al. A protein-protein interaction map of the Caenorhabditis elegans 26S proteasome. EMBO Reports 2, 821-828 (2001).
43. Consortium, C. e. S. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282, 2012-2018 (1998).
44. Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. Basic local alignment search tool. Journal of Molecular Biology 215, 403-410 (1990).
45. Hermjakob, H. et al. IntAct: an open source molecular interaction database. Nucleic Acids Research 32, D452-D455 (2004).
46. Weng, S. et al. Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. Nucleic Acids Research 31, 216-218 (2003).
47. Edwards, A. et al. Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genetics 18, 529-536 (2002).
48. von Mering, C. et al. Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417, 399-403 (2002).
49. Jansen, R. et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302, 449-453 (2003).
50. Xenarios, I. et al. DIP, the Database of Interacting Proteins: a research tool for studying

cellular networks of protein interactions. Nucleic Acids Research 30, 303-305 (2002).
51. Birney, E. et al. An Overview of Ensembl. Genome Research 14, 925-928 (2004).
52. Kolch, W. Meaningful relationships: the regulation of the Ras/Raf/MEK/ERK pathway by protein interactions. Biochemical Journal 351, 289-305 (2000).
53. Wang, C. C., Cirit, M. & Haugh, J. M. PI3K-dependent cross-talk interactions converge with Ras as quantifiable inputs integrated by Erk. Molecular Systems Biology 5, 246 (2009).
54. Chen, Y.-C., Chen, H.-C. & Yang, J.-M. DAPID: a 3D-domain annotated protein-protein interaction database. Genome Informatics 17, 206-215 (2006).
55. Hart, P. J. et al. Crystal structure of the human TbetaR2 ectodomain -- TGF-beta3 complex. Nature Structural Biology 9, 203-208 (2002).

Appendix A

List of publications

Journal papers

1. Yang, J.-M. and **Chen, C.-C.** GEMDOCK: a generic evolutionary method for molecular docking. *Proteins* **55**, 288-304 (2004). (Impact factor: 4.429)

2. Yao, Y.-Y., Shrestha K.L., Wu, Y.-J., Tasi H.-J., **Chen, C.-C.**, Yang, J.-M., Ando, A., Cheng C.-Y. and Li, Y.-K. Structural simulation and protein engineering to convert an endo-chitosanase to an exo-chitosanase. *Protein Engineering Design & Selection* **21**, 561-566 (2008). (Impact factor: 2.662)

3. **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: 6.954)

Conference Papers

1. Huang J.-W., **Chen, C.-C.**, Yang, J.-M. (2008) Identifying critical positions and rules of antigenic drift for influenza A/H3N2 viruses. *The 2nd International Conference on Bioinformatics and Biomedical Engineering*, pp. 249-252

Appendix B

Journal papers

Yang, J.-M. and **Chen, C.-C.** GEMDOCK: a generic evolutionary method for molecular docking. *Proteins* **55**, 288-304 (2004).

Yao, Y.-Y., Shrestha K.L., Wu, Y.-J., Tasi H.-J., **Chen, C.-C.**, Yang, J.-M., Ando, A., Cheng C.-Y. and Li, Y.-K. Structural simulation and protein engineering to convert an endo-chitosanase to an exo-chitosanase. *Protein Engineering Design & Selection* **21**, 561-566 (2008).