

國立交通大學

生物資訊所

碩士論文

結構功能區域交互同源性為基之蛋白質功能區域
及交互作用預測



Inferring Domain Annotated Protein-Protein Interactions
through 3D-Domain Interologs

研究生：陳永強

指導教授：楊進木 教授

中華民國九十五年七月

結構功能區域交互同源性為基之蛋白質功能區域及交互作用預測

Inferring Domain Annotated Protein-Protein Interactions through
3D-Domain Interologs

研究生：陳永強

Student : Yung-Chiang Chen

指導教授：楊進木

Advisor : Jinn-Moon Yang



A Thesis Submitted to Institute of Bioinformatics

National Chiao Tung University in partial Fulfillment of the Requirements

for the Degree of Master in

Bioinformatics

July 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年七月

結構功能區域交互源性為基之蛋白質功能區域及交互作用預測

學生：陳永強

指導教授：楊進木

國立交通大學生物資訊所碩士班

摘 要

蛋白質間的交互作用在生物體內複雜反應途徑中扮演重要角色之一。在後基因體時代，具備大規模找尋蛋白質蛋白質交互作用的能力是深入了解蛋白質網路的主要途徑之一。Lu 等人提出”交互作用源性對應(interologs mapping)”，大規模預測蛋白質蛋白質交互作用 — 即利用計算比較基因體學的方法，將大量蛋白質交互作用註解從一個物種對應到另外一個未經實驗方法註解的物種上。然而，在蛋白質交互作用中，通常都是經由特定的功能區域(domain)作物理性接合進而執行功能。目前解蛋白質結晶結構的速度日益進步，這些實驗資料使得目前十分適合利用已知結構蛋白質複合體預測蛋白質-蛋白質交互作用。

在此研究中，我們提出一個新的觀念“結構功能區域交互源性對應(3D-domain interologs mapping)”，預測蛋白質功能區塊及交互作用。結構功能區域交互源性對應的定義為”在一個已知結構的蛋白質結構上，蛋白質 A 的功能區域 a 與蛋白質 B 的功能區域 b 作物理接合，則他們在同一個物種中的同源蛋白質 A'(具有功能區域 a)以及 B'(具有功能區域 b)可能會發生交互作用”。我們主要的創新在於能夠快速的在數百個物種中進行基因體規模的蛋白質交互作用預測，並且發展一個新的成對位置加權矩陣(pairPSSM)。這個矩陣能夠利用演化式側寫提供不同的胺基酸對出現在某個特定位置的統計意義，使記分系統更加準確。我們的方法在分辨真實蛋白質複合體及不具生物意義蛋白質對的測試中可以達到將近九成的正確率。另外我們也嘗試預測酵母菌的蛋白質交互作用，和過去方法相比我們能夠提昇將近一成的預測準確率，而且這些蛋白質交互作用的平均基因表現相關性明顯高於不會發生交互作用的蛋白質對。最後，我們在七個常見的物種中，包含人類(*Homo sapiens*)、家鼠(*Mus musculus*)、大鼠(*Rattus norvegicus*)、線蟲(*Caenorhabditis elegans*)、果蠅(*Drosophila melanogaster*)、酵母菌(*Saccharomyces cerevisiae*)以及大腸桿菌(*Escherichia coli*)進行大規模蛋白質交互作用預測，從這些物種中可以預測到約四十五萬對新的蛋白質蛋白質交互作用，同時我們還能在這些蛋白質交互作用中提供交互作用功能區塊及接觸胺基酸對的註解。綜合以上所述，我們認為”結構功能區域交互源性對應”及”成對位置加權矩陣”是一個具有實際應用價值的蛋白質蛋白質交互作用預測方法並能進一步研究蛋白質交互作用網路。

Inferring Domain Annotated Protein-Protein Interactions through 3D-Domain Interologs

Student: Yung-Chiang Chen

Advisor: Dr. Jinn-Moon Yang

Institute of Bioinformatics
National Chiao Tung University

ABSTRACT

The interaction between proteins is one of the most important features to most biological processes. In the postgenomic era, the ability to identify protein-protein interactions on a genomic scale is very important to determine networks of protein interactions. To predict protein interactions large-scale, Lu *et al.* presented “interologs mapping”, — predicting protein-protein interactions from one organism to another by using computational comparative genomics. However, behind protein interactions there are protein domains interacting physically with one another to perform the specific functions. According to the increasing number of solved structures involving protein complexes, it is ripe to test putative interactions on complexes of known 3D structures.

In this study, we proposed a new concept “3D-domain interologs mapping” to infer domain-annotated protein interactions. The 3D-domain interologs mapping is defined as “Domain *a* (in chain A) interacts with domain *b* (in chain B) in a 3D complex, their inferring protein pair A' (containing domain *a*) and B' (containing domain *b*) in the same species would be likely to interact with each other if both protein pairs (A' and A as well as proteins B and B') are homologous ” The key novelties of our method are fast genome-scale prediction across hundreds of organisms and construction of a pair Position Specific Scoring Matrix (pairPSSM). This matrix is able to provide statistical significance of residue pairs at various contact positions by evolutionary profiles, leading to a more sensitive scoring system. Our method successfully distinguishes the true protein complexes and unreasonable protein pairs with about 90% accuracy. We also evaluate our method in yeast proteome and get about 10% improvements than previous methods. The mean correlation of the gene expression profiles of our predictions is significantly higher than that for non-interacting protein pairs in *S. cerevisiae*. Finally, our method applies to seven organisms commonly used in molecular research, including *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. In these seven organisms, our method predicts ~450,000 new interactions in which the interacting domains and residues are automatically modeled. In conclusion, this study suggests that 3D-domain interologs mapping and pairPSSM are useful methods for predicting protein-protein interactions and detailed analyzing networks of protein interactions.

致 謝

在兩年碩士生涯中，我首先要感謝我的指導教授楊進木老師，不管在研究或是做人處世方面，都給我很多的指導與啟發，讓我對生物資訊領域有概括性的了解並且學習到作科學研究應該具備態度和方法。當我研究上遇到許多挫折的時候，感謝老師能持續的給我鼓勵並且適時提出合理解決方案。

其次我想感謝系統生物組的學長陳宏助、劉康平及學弟許文昌，因為有你們的建議還有系統程式方面的協助，使我的研究能夠順利進行。還有實驗室的同學們楊登凱、葛振寧、黃章維，在碩二這段最煎熬的時刻，陪我渡過一個又一個寂靜的夜晚，相信這段回憶會一直保留在我們腦海中吧。感謝俊辰學長常常和我討論人生的道理，使我對未來有不一樣的看法；感謝董花每天中午和晚上都陪我一起吃飯；感謝阿甫常幫我拿公文還有幫我處理電腦方面的問題。感謝 PIKI 和右儒提供我完整的 FTP 站台。感謝籃球的球友黃鎮剛老師，景盛，水雉，蔚倫，草霸，建華，立人，文宏，taco，小胖以及其他摩獸爭霸 dota 的隊友們幫助我排遣研究之餘的空閒時間。

另外我要感謝我的女朋友靜芬，謝謝你能體諒我這兩年不能常常陪在你身邊，對我所做的決定都給予最大的支持還有幫助，讓我可以沒有後顧之憂在新竹為自己的理想努力奮鬥。

最後我想感謝我的家人，儘管我沒有很多時間可以陪在你們身邊，謝謝你們還是無怨無悔的支持我，鼓勵我。謝謝媽媽每兩個禮拜都幫我準備愛心水果，讓我健健康康的在新竹唸書；謝謝爸爸不管早晚不論風雨都接送我到車站；謝謝哥哥姐姐常常聽我抱怨東抱怨西。感謝你們所為我付出的一切一切，在此我僅能以小小的研究成果獻給你們，感謝你們。

陳永強 謹誌

中華民國九十五年七月

CONTENTS

Abstract (in Chinese)	I
Abstract	II
Acknowledgements (in Chinese)	III
Contents	IV
List of Tables	VI
List of Figures	VII
Chapter 1. INTRODUCTION	1
1.1 Background	1
1.2 Related works	4
1.2.1 Generalized interologs mapping	4
1.2.2 Structural-based prediction of protein-protein interactions	5
1.3 Motivation	6
1.4 Thesis overview	7
Chapter 2. MATERIALS AND METHODS	9
2.1 Overview	9
2.2 Preparation of data sets	10
2.2.1 Database of 3D-dimer	10
2.2.2 Definition of protein domain	10
2.2.3 Data set of related 3D-dimers	11
2.2.4 Data set of true protein complexes and unreasonable protein pairs	12
2.2.5 Data set of yeast proteome	13
2.2.6 Data set of yeast gene expression	14
2.2.7 Performance criteria	14
2.3 Construction of pairPSSM	16
2.3.1 Building comprehensive and non-redundant protein database	16
2.3.2 Scoring matrix architecture	17
2.3.3 Construction of multiple sequence alignment	17
2.3.4 Target frequency estimation	18
2.3.5 Amino acid classification	19

2.3.6 pairPSSM evaluation.....	20
Chapter 3. RESULTS and DISCUSSIONS	21
3.1 Two issues in modeling protein-protein interactions by homology	21
3.1.1 Similar 3D-dimers imply similar interactive types	21
3.1.2 Sequence identity threshold of aligning contact residues	24
3.2 Verification in true protein complexes and unreasonable protein pairs.....	25
3.3 Verification in yeast proteome.....	27
3.4 A search example: <i>Ia2kAD</i>	28
3.5 Verification in yeast gene expression profiles	31
3.6 Application: Prediction of protein interactions in seven common organisms	31
3.7 Model human protein interactions by <i>IevtBD</i>	32
Chapter 4. CONCLUSIONS.....	34
4.1 Summary.....	34
4.2 Major contributions and future perspectives	35
References.....	76



List of Tables

Table 1. The compositions of protein sequence in RefSeq database (Release 16)	37
Table 2. The frequency of amino acid occurs in protein surface and whole protein	38
Table 3. 24 pairs of related heterodimers with > 30% identity but pair coverage < 0.4	39
Table 4. AP and FP on 182 queries where the unannotated candidates considered as negatives	40
Table 5. AP and FP on 101 queries where the unannotated candidates are removed	45
Table 6. The result of 1a2kAD to search yeast proteome	48
Table 7. The pairPSSM of protein complex 1a2kAD	50
Table 8. Statistic of our predictions for seven common organisms	52
Table 9. The result of 1evtBD to model seven FGF/receptor complexes	53



List of Figures

Figure 1. The 3D structure and domain architecture of protein complex P47/P97	54
Figure 2. The comparison of our and previous methods	55
Figure 3. The flow chart of our method.....	56
Figure 4. The standard deviations of contact residue potentials in the clusters of amino acid	57
Figure 5. The relationship between sequence identity and pair coverage of 459 pairs of related hetero dimers	58
Figure 6. The relationship between sequence identity and pair coverage of 1412 pairs of related homo dimers.....	59
Figure 7. The average pair coverage in different sequence identity interval	60
Figure 8. The relationship between sequence identity and pair coverage of the two-chain dimers.....	61
Figure 9. The interactive types of two hydrolase-antibody complexes 1op9AB and 1jttAL	62
Figure 10. Sequence identity threshold of aligning contact residues	63
Figure 11. Determining the threshold of specific interfacial energy on distinguishing the true protein complex and unreasonable protein pairs.....	64
Figure 12. Determining the threshold of general interfacial energy on distinguishing the true protein complex and unreasonable protein pairs.....	65
Figure 13. The relationship between number of contact residues in 3D-dimers and its specific interfacial energies which are calculated from pairPSSM	66
Figure 14. The relationship between number of contact residues in 3D-dimer and its general interfacial energies with are calculated from general empirical matrix	67

Figure 15. Determining the threshold of normalized specific interfacial energy on distinguishing the true protein complex and unreasonable protein pairs.....	68
Figure 16. The mean average positions and mean false positive rate of 182 queries. The unannotated candidates are considered as negatives.....	69
Figure 17. The mean average positions and mean false positive rate of 101 queries. The unannotated candidates are removed	70
Figure 18. Model for cycling transport factors proposed by Koepp and Silver	71
Figure 19. The multiple sequence alignment result of the 14 candidates to their corresponding template proteins of 1a2kAD	72
Figure 20. 3D-structure of 1a2kAD.....	73
Figure 21. Distributions of the correlation coefficients of gene expression profiles for four interacting protein sets	74
Figure 22. The 3D-structure of 1evtBD and multiple sequence alignment of seven homologous FGF receptors.....	75



Chapter 1

INTRODUCTION

1.1 Background

Many biological processes involve different types of interactions among proteins. Listing the proteins in the cell is not enough to fully understand the cellular machinery and all the interactions between them need to be delineated as well(1). Recently, systematic identification of protein-protein interactions had been constructed by high throughput experimental methods (large scale yeast two-hybrid analysis or proteomics immunoprecipitation e.g.) for diverse organisms, such as the yeast *Saccharomyces cerevisiae*, the fruitfly *Drosophila melanogaster* and the nematode worm *Caenorhabditis elegans*(2-4). Simultaneously, a lot of computational methods had also developed to predict protein-protein interactions genome-widely, such as gene fusion events(5), gene expression profiles(6), phylogenetic profiles(7), known 3D complexes(8,9), interologs mapping(10) (two proteins will interact with each other if their orthologous proteins do as well), domain-pair profiles(11), conservation of gene neighborhood(12) and co-evolution strategy(13). The interaction data obtained from these methods were being collected by DIP(14), BIND(15), MIPS(16) and STRING database(17).

In the postgenomic era, the ability to identify protein-protein interactions on a genomic scale is very important to determine protein interaction networks. Genome sequencing projects are in progress for more than 644 organisms, and complete sequences are now available for more than 160 prokaryotic and eukaryotic genomes. The NCBI Reference Sequence Project(18) collects 2,631,538 proteins for major research organisms. Most protein

sequences are without annotation of interaction. Facing the enormous protein sequences with unknown function, how to determine the protein interaction networks genome-scalely has become an important issue.

A research group presented “interologs mapping”(10) — the transfer of interaction annotation from one organism to another by using comparative genomics. For any given protein in one organism, all of its homologs in another organism are consider as a homolog family; both families of two interacting proteins are called interacting families and all possible protein pairs between two interacting families are considered as protein-protein interaction candidates.

Behind protein interactions there are protein domains interacting physically with one another to perform the necessary functions. Interactive domains can recruit the formation of multi-protein signaling complexes, and control the conformation, activity, and substrate specificity of enzymes(19). However, almost all large scale method to explore interacting proteins can not respond how a protein interacts with another one in molecular detail (which domains bind directly), whether experimental or computational methods. There are two major strategies to study domain-domain interactions. The first strategy was to identify the domain pairs that are highly correlated with interacting proteins pairs and estimated the domain-domain interaction probability by using known protein-protein interactions as training data(20,21). These estimated probabilities of domain-domain interactions may be used to predict the probabilities of protein-protein interactions. The other strategy was to identify interacting domain from three-dimensional structural information. They exploited structural information to provide interacting domains and atomic details for thousands of direct physical interactions between proteins(8,22). The knowledge about interacting domains of a given protein interaction is very important for predicting new protein interactions. For example, p97, a member of AAA+ family (ATPases associated with various cellular activities) are involved

in different cellular pathways by interacting with various adaptor proteins(23). The membrane fusion adaptor p47 forms a tight complex with p97 and mediates p97 binding to its t-SNARE (soluble NSF attachment protein receptor) syntaxin5 for another round of membrane fusion (24). The interaction between p47 and p97 in *Mus musculus* could be transferred to other species by interologs mapping (two proteins will interact with each other if their orthologous proteins do as well). According to the result of PSI-BLAST, we obtain the yeast Shp1p similar to p47 (*E value*: 10^{-114}) and the yeast Rix7p protein similar to p97 (*E value*: 10^{-122}). Both the two homologous proteins in *Saccharomyces cerevisiae* are very similar to the template (both *E value* smaller than 10^{-100} and sequence identity greater than 30%) and might interact with each other. Nevertheless, it may be not true in reality. Although both p97 and Rix7p belongs to type II AAA+ proteins which containing two ATPase domains, the Rix7p lacks the important binding domain – CDC48_N domain and not involved in the process of membrane fusion. Rix7p seems to be required for restructuring nucleoplasmic 60S pre-ribosomal particles to make them competent for nuclear export(25) (Figure 1). Therefore, the interaction between Rix7p and Shp1p should be an incorrect prediction result from lack of the knowledge about interactive domains.

According to the increasing number of solved structures involving protein complexes, it is ripe to test putative interactions on complexes of known 3D structure. In this study, we address these questions using a new concept, the “3D-domain interologs mapping” which is similar to “generalized interologs mapping”. The 3D-domain interologs mapping, the core idea of our method, is defined as “**Domain *a* (in chain A) interacts with domain *b* (in chain B) in a known 3D complex, their inferring protein pair A' (containing domain *a*) and B' (containing domain *b*) in the same species would be likely to interact with each other if both protein pairs (A' and A as well as proteins B and B') are homologous**”. Based on the concept of 3d-domain interologs mapping, we can infer lots of protein interactions across

different species quickly and automatically map interacting domains for our predicted interactions.

Our method can successfully distinguish the true protein complexes and non reasonable protein pair up to 90% accuracy. We evaluated our method in yeast proteome and get about 10% improve than previous methods. The mean correlation of the gene expression profiles of our predictions is significantly higher than that for non-interacting protein pairs in *S. cerevisiae*. Although our method uses structure information, it does not require that the structures of the modeling proteins be solved. For this reason, we can apply our method to predict protein interactions in the large protein database which contains several hundreds of complete genome sequences. Finally, the method applies to seven organisms commonly used in molecular research, including *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. In these seven organisms, our method predicts ~450,000 new protein interactions in which the interacting domains and residues (binding sites) are automatically modeled. These visualized interacting residues are useful for the detailed analysis of protein-protein interactions.

1.2 Related works

1.2.1 Generalized interologs mapping

One concept of interologs, first proposed by Walhout et al.(26), is if interacting proteins A and B in one organism have interacting orthologs A' and B' in another species, the pair of interactions A-B and A'-B' are called interologs. Protein-protein interactions can be predicted by maps known interactions in the source organism onto target organism. Yu et al.(10)

extended this idea to a large scale quantitative assessment on conservation of protein-protein interactions between proteins and organisms. They proposed “generalized interologs mapping: for any given protein in one organism, all of its homologs in another organism are considered as a homolog family; two families of two interacting proteins are called interacting families and all possible protein pairs between two interacting families are considered as protein-protein interaction candidates”. By using the interaction information of from *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Helicobacter pylori*, they quantitatively assessed the interactions can be reliably transferred between species as a function of the sequence similarity of the corresponding interacting proteins and find a joint sequence identity > 80% or a joint E -value < 10^{-70} to make a reliable transformation.

1.2.2 Structural-based prediction of protein-protein interactions

Some methods had paid attentions to known 3D-complexes in the PDB(27) to infer protein-protein interactions. Aloy et al.(8) used a 3D complex and alignments of homologues of the interacting proteins to access the fit of any possible interacting pairs on the complex by using empirical potentials. The MULTIPROSPECTOR, proposed by Lu et al.(22), attempts to study more distantly related protein sequences by threading sequences onto a library of interacting templates and scores based on the threading Z-score and the magnitude of the interfacial energy which is similar the first approach. The statistical interfacial pair potentials were developed from a dimer database (selecting the cocrystallized records from the PDB) with using of the following formula:

$$P(i, j) = -\log\left(\frac{N_{obs}(i, j)}{N_{exp}(i, j)}\right) \quad (1)$$

where $P(i, j)$ is the interfacial pair potential between interacting residue pair i and j ; $N_{obs}(i, j)$ is the observed number of interacting residue pairs of i, j between two chains; $N_{exp}(i, j)$ is the expected number of interacting residue pairs of i, j between two chains if there are no preferential interactions among them; The $N_{exp}(i, j)$ is calculated from:

$$N_{exp}(i, j) = X_i \cdot X_j \cdot N_{total} \quad (2)$$

where X_i is the mole fraction of residue i in total surface residues. N_{total} is the number of total interacting residue pairs. Both two methods applied the ratio of the observed frequencies to expected frequencies of pairings between two residue types to examine the two homologues if interacting with each other.

1.3 Motivations



The previous methods(8,22) focused on giving a pair of query proteins to search the 3D-dimer database library and then found a best fit template to score how well the pair of query proteins fit the template structure (Figure 2A). However, the technology of sequencing makes a might advance in the post-genomic era. There are more than two millions protein sequences across more than three thousands species in NCBI Reference Sequence database. It is hard to test all the protein pair (more than 1 billion) to search a dimer of known structure suitable to model them by homology. Here we combine the concept of generalized interologs mapping and structure based prediction of protein interactions to propose a new concept “3D-domain interologs mapping”, defined as “Domain a (in chain A) interacts with domain b (in chain B) in a known 3D complex, their inferring protein pair A' (containing domain a) and B' (containing domain b) in the same species would be likely to interact with each other if both protein pairs (A' and A as well as proteins B and B') are homologous. Using a dimer of

known structure as template, we can predict lots of domain annotated protein-protein interactions across different species (Figure 2B).

The statistical interfacial pair potentials were used to score how well the query protein pair fit the template structure by previous methods. This is a general empirical matrix for all the dimers of known structures to model the pair of query protein. However, although binding sites are mainly hydrophobic, protruding, and electrostatic complementary, no general patterns are observed(28). It had been found that the free energy of binding is not evenly distributed across interfaces; instead, there are hot spots of binding energy made up of a small subset of residues in the interface of complexes(29). Keskin et al. had found there is a correspondence between the experimental identified energy hotspot and the structurally conserved residues(28). Therefore, we consider the general empirical matrix cannot characterize all binding site correctly. Many researches had been proven that conservative residues may perform specific functional (e.g. catalysis, recognition, binding) role(30,31). In our study, we also develop a method to estimate the probabilities with which residue pairs occur at various contact positions by evolutionary profiles, leading to a more sensitive scoring system. We consider our pair Position Specific Scoring Matrix (pairPSSM) can automatically characterize each interface of complexes and achieve a good performance for predicting protein-protein interactions.

1.4 Thesis overview

In this study, we use “3D-domain interologs mapping” to predict domain annotated protein interactions. The flowchart shows in Figure 3. We collect dimers of known structure from Protein Databank. For each 3D-dimer, we estimate the probabilities with which residue pairs occur at various contact positions and construct a pairPSSM to assess the fit of any

possible interacting protein pairs. And then we use these dimers as queries to search target protein database and predict many candidates of protein-protein interaction. We evaluate our method on three datasets; one is the multiple complex structures with the same interacting SCOP domains(32); one is protein-protein interactions in yeast proteome; and the other is yeast gene expression profiles. We finally apply our method to seven organisms commonly used in molecular research, including *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. In these seven organisms, our method predicts ~450,000 new protein interactions in which the interacting domains and residues (binding sites) are automatically modeled. These visualized interacting residues are useful for the detailed analysis of protein-protein interactions.



Chapter 2

METHODS AND MATERIALS

2.1 Overview

We apply 3D-domain interologs mapping to infer domain annotated protein-protein interactions from 3D protein complexes. Step by Step of our method is showed as follows:

(i)preparing the database of dimer template: we consider two contact chains in a 3D protein complex as 3D-complex template. We identify the contact residues of a 3D-complex template (containing chains A and B) in the PDB and define domain boundary by the SCOP database;

(ii)for each dimer template, build a protein-protein interaction position-specific matrix (pairPSSM); (iii)generating candidates: we use a dimer template to search two protein families (**A'** and **B'**) which contain the corresponding domains (*a* and *b*) from the target protein database and consider all the protein pairs between the two family as candidates of interacting proteins; (iv)scoring: we project the contact residue positions from a dimer template to its homologous protein pair. Then, summation of energy of all contact residue pairs in the homologous protein pair based on pairPSSM is considered the interactive energy of the homologous protein pair. If the interactive energy exceed to a threshold, we predict the two proteins interact with each other.

2.2 Preparation of data sets

2.2.1 Database of 3D-dimer

The PDB database (Protein Data Bank, <http://www.rcsb.org/pdb/>)(27) stores many three-dimensional structure of macromolecules, some of which are cocrystallized records. We select the cocrystallized proteins with using the criteria listed below:

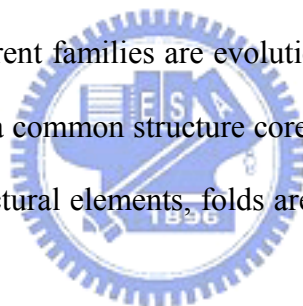
1. The resolution of the PDB records should be smaller than 3.0Å.
2. Each chain of the cocrystallized proteins should be comprised more than 35 amino acids to be considered as domains. If the protein is consisted of the cross-chain domain defined by SCOP, we also regard it.
3. The number of interacting residue pairs is set to be greater than 25 and each chain must contain more than 5 contact residues to make sure that the dimer is reasonably extensive(20). Interacting residue pairs are defined as a pair of residues from different chains that have at least one pair of heavy atoms within 4.5Å with each other.
4. Elimination of artificial packing complexes rather than biologically functional multimers by using PQS server(33), where adopt the reduction of solvent accessibility (ASA) due to oligomerisation.

From 35343 records in PDB (20060204), 29369 dimers of known structures are selected including 8018 heterodimers and 21351 homodimers. Then, we remove redundancy by sequence identity > 50% and leave 1122 heterodimers and 3514 homodimers, respectively.

2.2.2 Definition of protein domain

The protein domain definition is referenced by SCOP (Structure Classification of

Proteins, <http://scop.berkeley.edu/>(32). In 1.69 releases, there are 70859 domains from 25973 PDB entries. The SCOP database is based on evolutionary relationships and on the principles that govern their three-dimensional structure. Strong sequence similarity alone is considered to be sufficient evidence for common ancestry. Close structural and functional similarity together is also accepted as sufficient evidence for distant homology between proteins that lack significant sequence similarity. But neither structural nor functional similarity alone is considered to be strong evidence. Therefore, the SCOP database is organized on a number of hierarchical levels, with the principle ones being family, superfamily, fold and class. Families contain protein domains that share a clear common evolutionary origin, as evidenced by sequence identity or extremely similar structure and function. Superfamilies consist of families whose proteins share very common structure and function, and therefore there is reason to believe that the different families are evolutionarily related. Folds consist of one or more Superfamilies that share a common structure core structures. Depending on the type and organization of secondary structural elements, folds are grouped into four main classes (all α , all β , $\alpha+\beta$, α/β).



2.2.3 Data set of related 3D-dimers

We want to explore whether the two similar dimers possess the similar protein interaction type. Modeling protein interactions by homology is reasonable only when this hypothesis is valid. Here we defined the two 3D-dimers contain the same interacting SCOP domain (more than three contact residues within the domain boundary) are related-dimer pairs. From our database of 3D-dimer, we first remove the dimers with no definition of SCOP domain then leave 5553 heterodimers and 15026 homodimers. Second, the dimers are clustered by BlastCluster(34) according sequence identity more than 80% and both sequence

coverage more than 0.8. We choose one representative dimer from each cluster if the number of interacting residue pairs more than the mean of the cluster and the resolution of crystallization is smallest. In this way, a representative set of 3D-dimers is selected, which contains 999 heterodimers and 2837 homodimers. Third, the representative set of 3D-dimers is grouped based on the domain definition in SCOP. We group the dimers which possess the same interacting domain pair in family level. Totally, there are 540 types of interacting domain pair in the 999 heterodimers and 1425 types of domain-pair in the 2837 homodimers, respectively. 189 groups of heterodimer and 488 groups of homodimer contain more than 1 member. We choose one representative member for each group and pair the representative one for the all other members in the group. These pairs of dimer are considered as related dimer pairs. In this way, we select 459 related dimer pairs from the 189 groups of heterodimer and 1412 related dimer pairs from 489 groups of homodimer, respectively.

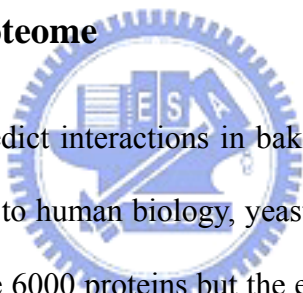
On the other hand, we define a subset from database of 3D-dimer which include the dimers should be in two-chain PDB records (there are only two proteins in the PDB entry). The rationale is interaction between two proteins may be bothered by other proteins if there are more than three proteins cocrystallized at the same time. There are 897 two-chain heterodimers and 3665 two-chain homodimers selected from the database of dimer template. Finally, we can get 114 pairs of two-chain heterodimer and 616 pairs of two-chain homodimers.

2.2.4 Data set of true protein complexes and unreasonable protein pairs

Our method is based on a 3D-dimer to model all potential protein interactions and used a specific pairPSSM to determine whether the two proteins interact in nature. Here we construct a data set include the protein pairs which really form complexes in living thing and the protein

pairs don't interact with each other (the unreasonable protein pairs). From 189 representative 3D-dimers (exclude the antibody-antigen complex), we used the PSI-BLAST to search our database of 3D-dimer which remove dimers with $> 80\%$ sequence identity and get homologous protein pair with *E value* threshold 0.1. If the protein pair is cocrystallized in PDB and it contains the same interactive SCOP domains for the query dimer, we consider the protein pair as positive that means this protein pair should be predicted by the 3D-dimer. In the other word, if the protein pair is not cocrystallized in PDB and it does not contain the same interactive SCOP domains for the query dimer, we consider the protein pair as negative. In this way, we can select 224 positive protein pairs and 282 negative protein pairs.

2.2.5 Data set of yeast proteome



We test our method to predict interactions in baker's yeast (*Saccharomyces cerevisiae*). Being a model system relevant to human biology, yeast has attracted special interest from the scientific community. There are 6000 proteins but the estimated number of actual interactions is smaller than 100,000 in *Saccharomyces cerevisiae*(1). Recently, high-throughput interaction screens of protein interactions in *S. cerevisiae* had been conducted by yeast two-hybrid(2,35) and affinity purification followed by mass spectroscopy(36,37). At the same time, many large-scale experiment of gene expression for *S. cerevisiae* proteome are also carried out(38,39). The public available data of functional genomics in *Saccharomyces cerevisiae* are most comprehensive.

The yeast proteome is obtained from the web site of the SGD (*Saccharomyces Genome Database*, <http://www.yeastgenome.org>)(40). The corresponding amino acid sequences of total 5877 open reading frames (ORFs) are subsequently downloaded. The comprehensive protein-protein interactions are downloaded from the DIP database (*Database of Interacting*

Protein, <http://dip.doe-mbi.ucla.edu/>)(14). The DIP database catalogs experimentally determined interactions between proteins. It combines information from a variety of sources to create a single, consistent set of protein-protein interactions. The data stored within the DIP database were curated, both, manually by expert curators and also automatically using computational approaches that utilize the knowledge about the networks of protein-protein interaction extracted from the most reliable, core subset of the DIP data. There are total 5882 reliable interactions downloaded from DIP and considered as positive set.

Because no directly information about which proteins do not interact, Jansen et al.(6) assumed that proteins in different compartments do not interact with each other and generate 2,708,746 non-interacting protein pairs from lists of proteins in separate subcellular compartments(41). We used these protein pairs as gold negative set.

2.2.6 Data set of yeast gene expression

The gene expression profiles of two interacting proteins were also used to access the accuracy of our method according to the basic assumption: “the gene pair with similar expression profiles is likely to encode an interacting protein pair”(42). The Rosetta compendium set consisting of the expression profiles of 300 deletion mutants and under chemical treatments(39) was used to measure the similarity of gene expression profiles of two genes

2.2.7 Performance criteria

We used two common metrics to assess the quality of our method, including mean average precision (MAP) and mean false positive rate (MFP). The mean average precision is

defined as:

$$AP = \left(\sum_{i=1}^A i / T_h^i \right) / A \quad (3a)$$

$$MAP = \left(\sum_{i=1}^M AP_i \right) / M \quad (3b)$$

where T_h^i is the number of compounds in a hit list containing i correct interactions; A is total number of true hits in the databases; M is the total number of template.

The mean false positive rate is defined as:

$$FP = \left(\sum_{h=1}^A (T_h - A_h) / (T - A) \right) / A \quad (4a)$$

$$MFP = \left(\sum_{i=1}^M FP_i \right) / M \quad (4b)$$

where A_h is the number of active ligands among the T_h highest ranking compounds; T is the total number of candidates from PSI-BLAST; A is total number of true hits in the T ; M is the total number of template.

In this study, the similarity of two gene-expression is defined by the Pearson correlation coefficient between the two gene-expression profiles (see 2.2.6). To test whether the mean of correlation coefficient for candidates of protein-protein interactions higher than that of non-interacting protein pairs, we calculate the T-score and the P-value for the null hypothesis of the sample mean (our prediction) smaller than the mean of gold negative set. We apply standard two sample T-test statistics:

$$T = \frac{u_1 - u_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \quad (5)$$

where u is mean of samples, and S is the standard deviations of the samples:

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - u)^2} \quad (6)$$

2.3 Construction of pairPSSM

Here we want to develop a method to estimate the probabilities with which residue pairs occur at various contact positions by evolutionary profiles, leading to a more sensitive scoring system. The probabilities with which residue pairs occur at various contact positions are transformed to energy of contact residue pair (pair Position Specific Scoring Matrix called pairPSSM). The energy calculated from pairPSSM is the specific interfacial energy. On the other hand, the energy calculated from empirical matrix is the general interfacial energy.



2.3.1 Building comprehensive and non-redundant protein database

To obtain the evolutionary profiles from multiple sequence alignment, our alignment result should be come from a comprehensive and non-redundant protein database. The protein database is obtained from the web site of the NCBI Reference Sequences database (Release 16, <http://www.ncbi.nlm.nih.gov/RefSeq/>). The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major research organisms. In Release 16, Refseq includes 2,273,764 protein sequences across 3244 organisms. Table 1 shows the protein sequence composition in Refseq database. Although RefSeq aims to provide a non-redundant set of sequences for users, two major source of redundancy occur in RefSeq. One is alternative splicing; the other is duplication of genes (paralog). Therefore, we used BlastCluster to remove redundancy with both the sequence identity and coverage as high as

90% in the same species. In this way, 2,109,945 protein sequences are selected into our non-redundant protein database.

2.3.2 Scoring matrix architecture

For a 3D-dimer with the number of contact residue pairs M , the empirical energy matrix of dimension 20×20 is replaced by a protein-protein interaction position-specific matrix (pair PSSM) of dimension $M \times 20$. The residue pair in a contact position is considered as a single symbol. The advantage of this matrix is estimation of the probabilities with which residue pairs occur at various contact positions, leading to a more sensitive scoring system.

2.3.3 Construction of multiple sequence alignment

To produce a multiple sequence alignment from the PSI-BLAST output, we simply collect all RefSeq sequence segments that have been aligned to the two proteins of 3D-dimer with E value below a threshold, by set to 10^{-9} to assure the members are similar enough to the 3D-dimer. The two proteins of 3D-dimer are used as a master for constructing two multiple sequence alignments, respectively. Any row that is $> 95\%$ identical to 3D-dimer is purged. To ensure that all the sequences found by PSI-BLAST are likely to be structurally related to the dimer template, we set a similarity threshold described by Batalov & Abagyan(43,44) was used, defined as follows:

$$t(L) \leq 31.8L^{-0.102} + m \times 17.4L^{-0.29} \quad (7)$$

where $t(L)$ is the percentage sequence identity threshold dependent on alignment length L , and m is the level confidence of this threshold (in standard deviations). In this work, we use $m = 3$

(the identity threshold at least above 25%) to make sure all the sequences found by PSI-BLAST with similar interactive type to their dimer template (see section 3.1).

Because most the interacting proteins belong to the same species, the proteins in the two multiple sequence alignments must be arranged in order by species. For example, the 1st protein in a multiple sequence alignment is a human protein, and the 1st protein in the other must be a human protein. The 2nd protein in a multiple sequence alignment is a mouse protein, and the 2nd protein in the other must be a mouse protein.

2.3.4 Target frequency estimation

Given two multiple sequence alignment from a 3D-dimer, we generate score matrices with the theoretical foundation is that the scores for a specific contact position be of the form $\log(Q_{ij}/P_{ij})$, where Q_{ij} is the estimated probability for contact residue pair $i&j$ to be found in the column and P_{ij} is the expect probability of $i&j$ to be found in the column. The estimate of Q_{ij} for a specific contact position should converse simply to the observed frequency of residue pair $i&j$ in that column. However, it is complicate estimating the Q_{ij} include small sample size and prior knowledge of relationships among the residues should be considered. We implemented the data-dependent pseudocount method introduced by Tatusov et al.(45). It is relative simplicity and often performs nearly as well comparing the Dirichlet mixtures(46). We slightly modify the data-dependent pseudocount method by using the prior knowledge of amino acid relationships embodied in the substitution matrix to generate residue pair pseudocount frequencies g_{ij} . For a given position pair of contact residues C , we construct pseudocount frequencies g_{ij} using the formula:

$$g_{ij} = P_i P_j e^{S_{ij}} \quad (8)$$

where S_{ij} is the interactive energy of residue $i&j$ contact in empirical energy matrix. P_i is the background probability of residue i . The rationale is that we use the prior knowledge of interactive energy between residue $i&j$ to estimate pseudo frequency. We then estimate Q_{ij} followed as:

$$Q_{ij} = \frac{\alpha f_{ij} + \beta g_{ij}}{\alpha + \beta} \quad (9)$$

where the α and β are the relative weights given to observed and pseudocount residue frequency. In our study, we let α = the number of different residue-pair types in column -1 and $\beta = 5$. If the β is larger, the greater the emphasis is given to prior knowledge.

P_{ij} is the expect probability of $i&j$ to be found in the column and is calculated from:

$$P_{ij} = P_i \cdot P_j \quad (10)$$

where P_i is the expect probability of residue i occurring in protein surface. The residue composition of the protein interface is obtained from Lu et al.(47)(Table 2).

2.3.5 Amino acid classification

The sequence variability at each contact position could be estimated from the two multiple sequence alignment of dimer template. However, by not making concessions for conservative mutations the scheme becomes too rigid. Unlike unconservative mutations, conservative ones preserve the essential nature of the side chain. Therefore, we make some tolerances for such mutations. Saha et al.(48) made a classification based on the similarity of the environment of each amino acid residue in protein structures to the nine groups:

(i) Ala and Val; (ii) Met, Leu and Ile; (iii) Gly, Ser and Thr; (iv) Pro, Phe, Tyr and Trp; (v) Cys;

(vi) His; (vii) Arg and Lys; (viii) Asp and Glu; (ix) Asn and Gln. We test this classification of amino acid whether suitable for access the contact residue potential by calculating the standard deviation of contact residue potential in the cluster of amino acid. Figure 4A shows that the three groups {A, V}, {P, F, Y, W} and {R, K} have high standard deviations of intra contact residue potential. Therefore, we slightly modify the group as follows:

(i) Ala and Gly; (ii) Val, Met, Leu and Ile; (iii) Pro, Ser and Thr; (iv) Phe, Tyr and Trp; (v) Cys; (vi) His and Arg; (vii) Lys; (viii) Asp and Glu; (ix) Asn and Gln. In this way, all the standard deviations of intra contact residue potential are smaller than 0.4 (Figure 4B). We consider this amino acid classification is more reasonable for measuring the contact residue potential.

2.3.6 pairPSSM evaluation

The pairPSSM is evaluated by two data sets. First, we test whether the energy calculated from pairPSSM could distinguish the true protein complexes and unreasonable protein pairs (data set described in section 2.2.4). Second, we apply our method to predict protein-protein interactions in yeast proteome (data set described in section 2.2.5) and used two common metrics (MAP and MFP described in 2.2.6) to assess the performance of pairPSSM and compare the empirical matrix used by previous method. Then, we test the similarity of gene expression profiles for the candidates of protein-protein interactions. Finally, we give two true biological examples to illustrate the operation and power of our method.

Chapter 3

RESULTS and DISCUSSIONS

In this study, we first explore the relationship between sequence similarity and interactive similarity in protein-protein interactions. Modeling protein interactions by homology is reasonable only when the correlation is high enough. Second, because the structures of proteins are unsolved, we are only able to use the method of sequence alignment to align the template and target proteins. The consistence ratio between sequence alignment and structure alignment must also be studied. Third, our method is verified in two data sets: one is true protein complexes and unreasonable protein pairs. The other is protein-protein interactions in yeast proteome. Finally, we applied 3D-domain interologs mapping to predict protein interactions for seven organisms commonly used in molecular research.

3.1 Two issues in modeling interactions by homology

3.1.1 Similar 3D-dimers imply similar interactive types

The data set (459 pairs of related heterodimers and 1412 pairs related homodimers) described above provide all instances of a particular interaction type occurring within different complex structures, that we then wish to compare to each other and correlate with sequence similarity. To compare the binding of different instances of the two dimers with the same interacting domains, we devise an index, pair coverage, to calculate binding site overlap from the number of shared interacting residue pairs. Given a pair of related dimers A-B and

A'-B', where A-A' and B-B' contain same SCOP domains. we use a structural alignment tool, CE(49), to align the A-A' and B-B', respectively. The pair coverage is defined as:

$$Pair\ coverage = \sqrt{\frac{NCPM^2}{NCP_{AB} \cdot NCP_{A'B'}}} \quad (11)$$

where the $NCPM$ is the matching number of contact residue pairs between the structural alignment of A-A' and B-B'; NCP_{AB} is the number contact residues pairs of dimer A-B; $NCP_{A'B'}$ is the number contact residues pairs of dimer A'-B'.

The value of pair coverage is range from 0 ~ 1. The interactive types of two dimers are very alike in the pair coverage of a pair of related dimers is greater than 0.4. This threshold is chosen after visual inspection of many pairs of related dimers. On the other hand, the percentage sequence identity is calculated by the number of identical residues divided by the number of structurally equivalent residues. In following discussion, the sequence identity between two dimers is defined as the minimum of sequence identity in A&A' and B&B'. The rationale is that the interacting partners with the lower sequence identity would tend to be the better indicator for the diversity of the interaction.

Figure 5 shows the relationship between sequence identity and pair coverage of 459 pairs of related hetero dimers. It is clear that the interactions tend to be similar when sequence identity is above 30%. The pairs of related dimers in the gray box are the exceptions of the related dimer pair with > 30% sequence identity but pair coverage < 0.4. In the 280 out of 459 pairs of related heterodimers with > 30% sequence identity, there are only 24 pairs with pair coverage < 0.4. The rate of exception is 8%.

On the other hand, Figure 6 shows the relationship between sequence identity and pair coverage of 1412 pairs of related homodimers. The trend which gives a guide to the degree of sequence similarity needed to be confident in a similar interaction is also observed. However,

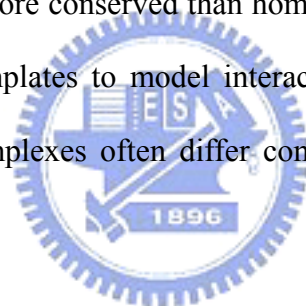
there are much more exceptions in the pairs of homodimers than heterodimers. In the 640 out of 1412 pairs of related homodimers with $> 30\%$ sequence identity, there are 189 pairs with pair coverage < 0.4 . The rate of exception is about 30%. Figure 7 shows the average pair coverage in different sequence identity. It is clear that the pair coverage of heterodimers higher than that of homodimers in difference sequence identity. It means the specificity of interaction in heterodimer is more conserved than homodimers. For this reason, we think heterodimers are more suitable used to template to model protein interactions than homodimers.

There are many PDB record with more than two proteins cocrystallized. Here we want to study whether the interactions bothered by other proteins. 114 pairs of related two-chain heterodimers and 616 pairs of related two-chain homodimers are selected (see 2.2.3). Figure 8 shows that the relationship between sequence identity and pair coverage is not much different from heterodimers or homodimers. 4 out of 90 of related two-chain heterodimers and 78 out of 255 related two-chain homodimers with $> 30\%$ sequence identity but with pair coverage < 0.4 . The rates of exception are 4.5% and 30.5%, respectively. From the above result, we think the conservative of interactive specificity is bothered slightly by other proteins.

There are 24 exceptions out of 280 pairs of related heterodimers with sequence identity $> 30\%$ but pair coverage < 0.4 . Table 3 shows the type of interacting domains in these 24 cases. Surprisingly, 11 out of 24 cases contain the b.1.1.1 domain (V set antibody variable domain). That means the interactive types of antigen-antibody complex need not conserve. Figure 9 shows an example of two much similar antigen-antibody complex but their interactive types are completely different. 1op9 (PDB id) is a hydrolase-antibody complex in human and 1jtt (PDB id) also is a hydrolase-antibody in chicken. The A chain of 1op9 and A chain of 1jtt both contain the V set antibody variable domain (SCOP id, b.1.1.1) and the sequence identity between the two proteins is as high as 77%. The B chain of 1op9 and B chain of 1jtt both

contain the C-type lysozyme domain (SCOP id, d.2.1.2) and the sequence identity between the two proteins is as high as 61%. By study the interactive site of the two complexes, we discover both the binding area of 1op9A and 1jttA at the variable region of V set antibody variable domain. However, the binding sites of 1op9B and 1jttL are very different. We superimpose the 1op9B and 1jttL and discover the binding site of the two proteins at two different sides (Figure 9). Therefore, the pair coverage between the two complexes is very low.

In summary, we find the related dimers indeed keep similar interactive type. Sequence similarity needed to be confident in a similar interaction. We suggest one must be careful with identity below than 30% to model interactions by homology. Because the specificity of interaction in heterodimer is more conserved than homodimers, we consider the heterodimers are more suitable used to templates to model interactions. Finally, we find the interactive types of antigen-antibody complexes often differ completely, they may be not suitable for used to as templates.



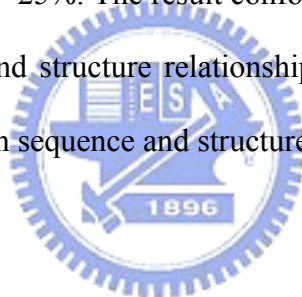
3.1.2 Sequence identity threshold of aligning contact residues

To model interactions for the proteins with unknown 3D structures, we need use the method of sequence alignment to align protein sequences between the template (3D-dimer) and target proteins (unknown 3D structure). Here we would likely to study if the performance of PSI-BLAST (a method of sequence alignment)(34) is as better as the performance of CE (a method of structure alignment)(49) in matching of contact residues. We devise an index, consistence ratio, to calculate the accuracy of PSI-BLAST in match of contact residue by using alignment result of CE as reference. The consistence ratio is defined as:

$$\text{Consistence Ratio} = \frac{NCM_{BLAST}}{NCM} \quad (12)$$

where the NCM_{CE} is the number of contact residue matching in CE alignment; NCM_{BLAST} is the number of contact residue matching occur both in PSI-BLAST and CE alignment.

286 pairs of related heterodimers whose pair coverage greater than 0.5 are used for our study to ensure the contact residue equivalent with biological meaningful. Figure 10A shows the relationship between the consistence ratio and sequence identity. The higher sequence identity, the higher consistence between PSI-BLAST and CE alignment. Figure 10B shows the mean of consistence ratios in different sequence identity. When the sequence identity is greater than 25%, the consistence is very high. Furthermore, there is a twilight zone between sequence identity between 20%~25%. The result conforms to the traditional twilight zone that already known for sequence and structure relationships(50). If the sequence identity below than 20%, it is much different in sequence and structure alignment result.



3.2 Verification in true protein complexes and unreasonable protein pairs

To verify our method, the two dataset should be collected. The one is positive dataset which contains the protein pairs indeed interact with each other and the other one is negative dataset which contains the protein pairs which do not interact. Here we would likely to study whether the specific interfacial energy calculated from pairPSSM could distinguish the two datasets. From above sections, 224 pairs of homologous, non-identical 3D-heterodimers are collected into positive dataset and 282 pairs of homologous protein pair are collected into negative dataset. We can use one dimer to score the other one by our specific empirical matrix (pairPSSM) and general empirical matrix.

Figure 11A shows the frequency of positive set and negative set occurred in different

specific interfacial energy intervals. The threshold is determined by result in minimum false positives and minimum false negatives. We calculate the error rate by averaging the number of false positive divided by number of positive set and the number of false negative divided by number of negative set. Figure 11B shows that when the specific interfacial energy is set to 50, we can obtain the minimum error rate 18%. We also apply the general interfacial energy to the positive set and negative set. The result is in Figure 12B. When the general interfacial energy is set to 10, we can obtain the minimum error rate 17%.

We are interested in the distribution of positive and negative dataset is not high concentrated in the two sides (positives in high energy and negatives in low energy) when using the specific interfacial energy (Figure 11A). That is why the error rate higher in using specific interfacial energy than in using general interfacial energy. We find there it is a high correlation between the specific interfacial energy and the number of contact residues in 3D-dimers (Figure 13). The correlation coefficient is 0.9321. The correlation between general interfacial energy and the number of contact residues (Figure 14) is not as high as the correlation between the specific interfacial energy and the number of contact residues. The correlation coefficient is only 0.6753.

Because specific interfacial energy is highly dependent on the characteristic of dimer template, we design a method to normalize the specific interfacial energy. When a homologous protein pair is modeled by a 3D-dimer and gets a specific interfacial energy scored by pairPSSM, we normalize the energy defined as follow:

$$E_{normalized} = \frac{E_{predict}}{E_{template}} \quad (13)$$

where the $E_{predict}$ is the specific interfacial energy of the homologous protein pair and the $E_{template}$ is the specific interfacial energy of the dimer template. By using the normalized

interfacial energy, we can find the distribution of positive and negative dataset is much more concentrated in the two sides (Figure 15A) than the unnormalized (Figure 11A) and the error rate reduce from 18% to 13% (Figure 15B). For this reason, we consider the normalized specific interfacial energy equal set to 0.4 is a good threshold for predicting protein interactions.

3.3 Verification in yeast proteome

The yeast (*Saccharomyces cerevisiae*) is a simple, unicellular eukaryote developed to a unique powerful model system for biological research. Its prominent useful features are the cheap and easy cultivation, short generation times, the detailed genetic and biochemical knowledge accumulated in many years of research. Therefore, this organism provides a highly suitable system to study basic biological processes that are relevant for many other higher eukaryotes including human. There are about 6000 proteins in this organism. Up to present, 5882 reliable protein-protein interactions in yeast are collected in DIP database (see 2.2.5). Here we predict protein interactions in this organism and used two indices, average precisions and false positive rate, to verify our method. The two indices are common used to evaluate the quality of database searching. From the dataset of 8018 heterodimers, we remove the redundancy with sequence identity $> 50\%$ and then select 1122 representative heterodimers as queries to search database of yeast proteome by PSI-BLAST. We defined the proteins searched out with *E value* smaller than 10^{-3} is homologous to the query protein. Given a query of heterodimer A-B, A' is the homologous protein for A and B' is the homologous protein for B. All the homologous protein pairs A'-B' are considered as candidates of protein-protein interactions. The known interactive protein pairs among the candidates are considered as positives and the others are considered as negatives.

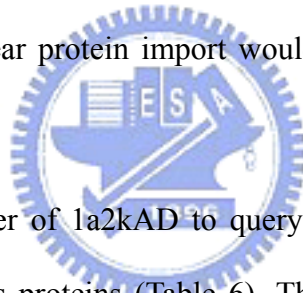
182 out of 1122 queries have both positive candidates and negative candidates, and then these queries could calculate the average precisions and false positive rate. All the detail result of the 182 queries is listed in Table 4. Figure 16 shows the mean average precisions (MAP) and mean false positive rate (MFP) of the 182 queries. The MAP is 0.42 and the MFP is 0.31 by using specific interfacial energy. On the other hand, the MAP is 0.35 and the MFP is 0.37 by using the general interfacial energy. In order to avoid our method merely predict protein interactions with high sequence identity, we set the sequence identity limit to remove the candidates if one protein of candidates with sequence identity > sequence identity limit. Figure 16 shows our method using pairPSSM is much better than the general empirical matrix even though in predicting remote protein interactions.

In the above section, the candidates which are not included in the known interactive protein pairs are considered as negatives. However, it may be somewhat unreasonable because many candidates are indeed interacting proteins in nature but have not proven by experimental methods in the past. Therefore, we only consider the candidates overlapping with 2708,746 (see 2.2.5) non-interacting protein pairs defined by Jasen et al. as negatives. The candidates without any annotations are removed for calculate average precisions and false positive rates. In this way, our method using pairPSSM is about 10 % improvement than the general empirical matrix (Figure 17).

3.4 A search example: *1a2kAD*

We give an example using the 3D-dimer, 1a2kAD, to search database of yeast proteome and illustrate the accuracy and operation of our method. The A chain of 1a2k is a rat nuclear transport factor 2 (NTF2) and the D chain of 1a2k is a dog GTP binding protein ran(51). The transportation between nucleus requires to the nuclear pore complexes (NPC) in the nuclear

envelope and several key factors including importin α and β , which recognize proteins with a nuclear localization sequences (NLS), the small GTP binding protein ran and nuclear transport factor (NTF)(52,53). Both RNA export and nuclear protein import depend on ran. The molecular details of the export of transport factors had been speculated by Koepp and Silver(52) in Figure 18. Once inside the nucleus, importin α must dissociate from the NLS-bearing substrate, which may be accomplished by competition with RNA-binding proteins. Ran may move out of the nucleus as a complex of Ran-GTP–importin β . Dissociation of these two proteins could be a result of the GAP activity of Rna1p, either inside the NPC or on the cytoplasmic face of the NPC. There is evidence that Rna1p can interact with importin β (53). The precise signal for an irreversible step of export is unclear, but it is possible that free importin β could dissociate importin α from RNA-binding proteins. Thus, the key players in nuclear protein import would be regenerated for another round of transport.



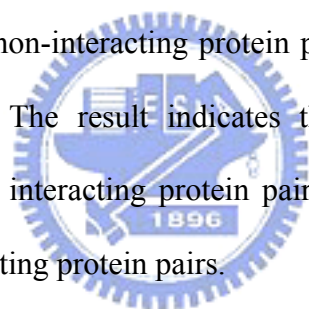
Here we use the 3D-dimer of 1a2kAD to query database of yeast proteome and then obtain 14 pairs of homologous proteins (Table 6). The two protein pair, NTF2&GSP1 and NTF2&GSP2, has been proven to bind with each other by yeast two hybrid test(54,55) and the other twelve protein pairs are non-interacting proteins due to locate in different compartments The specific interfacial energies calculated from pairPSSM (Table 7) of the two positive protein pairs are both above the threshold 0.4 and the twelve negative protein pairs are below the threshold (Table 6). It shows that our score is good for predicting protein interactions. However, the general interfacial energies of the two positive protein pairs are both above the threshold -15 (the more negative are more favor to bind)(47). And ten out of twelve negative protein pairs are above the threshold. In summary, 12 out of 14 protein pairs are predicted incorrectly with general interfacial energy an all the 14 interactions are predicted correctly by our specific interfacial energy.

Figure 19 shows the multiple sequence alignment result of the 14 candidates to their corresponding template, A chain of 1a2k or D chain of 1a2k. The interface involves primarily the putative switch II loop of ran (residue 65 to 78, Figure 19B orange box) and the hydrophobic cavity surrounding surface of NTF2(51). The interaction made by the switch II loops accounts for the ability of NTF2 to discriminate between GDP and GTP bound forms of Ran. A striking feature of the interactive interface was the aromatic ring of Phe72 of ran (Figure 19B, orange star site and Figure 20B, orange residue). It inserts into the hydrophobic cavity of NTF2 where it was surrounded by the hydrophobic side chains of Trp41, Leu59, Phe61, Ile64, Leu89, Ala91, Met97, Phe119 and Leu121. The GSP1 and GSP2 of two positive protein pairs are conservative in this important site (Figure 19). On the other hand, the interactive interface on NTF2 involved this molecule's characteristic hydrophobic cavity. Hydrophobic residues in the upper portion of the NTF2 cavity, together with negatively charged residues, Glu42, Asp92 and Asp94, are surrounding the cavity (Figure 19A, yellow boxes and Figure 18A, yellow residues) made significant contributions to the interface with GDP-Ran. The three important negative residues are conservative from A chain of 1a2k (rat NTF2) to the yeast NTF2. However, the three important sites are mutated to Threonine in BRE5 (Figure 19A). The BRE5 is an ubiquitin protease cofactor which forms deubiquitination complex with ubp3p that coregulates anterograde and retrograde transport between the Endoplasmic Reticulum and Golgi compartments. The three important residue mutated may be resulted in BRE5 does not interact with GSP1 and GSP2. Encouragingly, we give poor score to the 2 candidates (0.08) and successfully identify the true interactions GSP1&NTF2 and GSP2&NTF2.

3.5 Verification in yeast expression profiles

Recently, many scientists consider that genes with similar expression profiles are likely to encode interacting proteins(56). Therefore, we compare the distribution of gene expression profiles for the two gold standard sets and our predicted protein-pairs by 3D-domain interolog mapping with the score exceeding 0.4 and 0.6 (Figure 20). The protein pairs composed of the same protein are not used to calculate the gene expression profiles because their expression profiles must be identical and should not be taken account of. Figure 21 shows that the distribution of the correlation coefficients of our predicted protein pairs is similar to the core set of DIP (Positives) and right shift to non-interacting protein pairs (Negatives).

Then we used standard two sample T-test to test the mean of correlation coefficient for our predicted protein-pairs to non-interacting protein pairs. The *E values* of the two sets are 10^{-30} and 10^{-26} , respectively. The result indicates that the prediction based 3D-domain interolog yields many reliable interacting protein pairs indicates whose mean is significant higher than that for non-interacting protein pairs.



3.6 Application: Prediction of protein interactions in seven common organisms

In the above section, we have verified our method in two data sets and obtained a reasonable threshold for normalized specific interfacial energy about 0.4~0.5. Here we apply 3D-domain interologs mapping to prediction protein interactions in seven organisms commonly used in molecular research, including *Mus musculus* (house mouse), *Homo sapiens* (Human), *Rattus norvegicus* (Norway rat), *Caenorhabditis elegans* (nematode), *Drosophila melanogaster* (fruit fly), *Saccharomyces cerevisiae* (baker's yeast) and *Escherichia coli*. By

set the threshold to 0.5, we obtain about 450,000 protein interactions from the seven common organisms (Table 8). Comparing our predictions and the interactions deposited in DIP database, there is a large difference for number of interactions in the same organism. For example, we predict 1850 interactions in *Saccharomyces cerevisiae* but DIP collects 18225 interactions. On the other hand, we predict 112114 interactions in *Homo sapiens* but DIP collects only 292 interactions. There are two reason for large drop, one is the large-scale experimental method (such as large scale yeast two-hybrid analysis or proteomics-immunoprecipitation) is hard to apply in mammalian organisms and results in the interactions deposited in DIP are few in human, mouse or rat. The other reason is gene duplication and alternative splicing often occurred in the mammalian organisms and result in some redundancy protein in protein database. In these organisms, we may over estimate the number of prediction of protein-protein interactions.

Structural genomics projects are generating new structures at an unprecedented rate—a benefit of recent developments in high-throughput technologies(57). As a result, the number of protein structures in the Protein Data Bank (PDB) is increasing rapidly. For each new determined 3D-dimer, we can apply our method to predict all the candidates in thousands of organisms quickly. It helps the biologists to further detail analysis the network of protein interactions.

3.7 Model human protein interactions by *levtBD*

Another example for illustrate the power of our method can apply not only to yeast proteome but also to any other organisms. Interactions between the fibroblast growth factors (FGFs) and their receptors had been intensive studied(58,59). FGFs play key roles in morphogenesis, development, angiogenesis, and wound healing. These FGF-simulated

processes are mediated by for FGF receptor tyrosine kinase. There are more than 20 human protein FGFs that bind to one or more of 7 FGF receptors (FGFR1b, -1c, -2b, -2c, -3b, -3c, -4), where the c and b denote isoforms IIIc & IIIb formed by alternative splicing. The complex of FGF1/FGFR1 (Figure 22A) had been dissolved by Plotnikov et al. in PDB (accession number: 1evt)(60). Ornitz et al. perform a study of FGFR specificity by measuring mitogenic activity of FGFR-inducible BaF3 cell-line(61). Table 9 shows the binding affinity of the seven FGF/receptor complexes (from FGF4 to 7 receptors, FGFR1b, -1c, -2b, -2c, -3b, -3c, -4). The experimental determined binding affinity relative to the FGF-1 is $< 10\%$ defined as low affinity and $> 10\%$ defined as high affinity.

In our study, we used the 1evtBD to model interactions for the seven FGF/receptor complexes. 6 out of 7 FGF/receptor complexes are high affinity and our method give high interfacial energy for the six complexes. However, the other one, FGF4/FGFR3b complexes, with very low binding affinity (1.0%) but our method give a high normalized interfacial energy (0.84). For a detail sequence analysis (Figure 22B), we find that most contact positions in FGFR3b are much conservative except some residues in D3 immunoglobulin (Ig)-like domains (Figure 21B, orange box). This result may mean some other factors involved in determining the strength of the FGFR interactions. In conclusion, we successfully predict 6 out of 7 FGF/receptor complexes. There is a good agreement between the specific interfacial energy and binding affinity even though still with an incorrect case.

Chapter 4

CONCLUSION

4.1 Summary

We develop a new method “3D-domain interologs mapping” to infer domain annotated protein-protein interactions across several commonly organisms. We also develop a method to estimate the probabilities with which residue pairs occur at various contact positions by evolutionary profiles, leading to a more sensitive scoring system. In this study, we get some critical conclusions as follows:

1. Similar dimers indeed keep similar interactive type. We suggest one must be careful with identity below than 30% to model interactions by homology.
2. The method of sequence alignment is reliable in alignment of contact positions when the identity $> 20 \sim 25\%$.
3. The specific interfacial energy calculated from pairPSSM can successfully distinguish the true protein complexes and non reasonable protein pair with about 90% accuracy.
4. The pairPSSM outperforms general empirical matrix about 10% improvements even though for the distantly related protein sequences.
5. The mean correlation of the gene expression profiles of our predictions is significantly higher than that for non-interacting protein pairs in *S. cerevisiae*.

Although our method uses structure information, it does not require that the structures of the modeling proteins be solved. For this reason, our method can predict protein-protein

interactions in the large protein sequence database which contains several hundreds of complete genome sequence. We applied to seven organisms commonly used in molecular research, including *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. In these seven organisms, our method predicts ~450,000 new protein interactions in which the interacting domains and residues (binding sites) are automatically modeled. These visualized interacting residues are useful for the detailed analysis of protein-protein interactions.

4.2 Major contributions and future perspectives

We have developed a new method to predict protein interactions genome-scalely across a lot of organisms and constructed pairPSSM for each dimer, leading to a more sensitive scoring system. In post genomic era, Structural genomics projects are generating new structures at an unprecedented rate—a benefit of recent developments in high-throughput technologies. As a result, the number of protein structures in the Protein Data Bank (PDB) is increasing rapidly. We can use the more and more protein complexes of known 3D-structures to predict enormous protein interactions across several hundreds of genome sequence. Combining our predictions and several large protein-protein interaction databases, such DIP, BIND, MIPS or STRING etc., we can construct more completed networks of protein interactions for several organisms commonly used in molecular research. It is useful to makes biologists to realized biological system in details.

Some important issues will be discussed in the future. The protein-protein interactions are associated with processes such as cell signaling, enzymatic activity, immunological recognition, DNA repair and replication, vesicular traffic etc. Although binding sites are mainly hydrophobic, protruding, and electrostatic complementary, no general patterns are

observed. For this reason, we want to explore whether the characteristics of different interfaces between proteins could be identified by pairPSSM.

A method, alanine scanning mutagenesis, experimentally probes the energetic contributions of individual side chains to protein bindings. By using this technique, Wells and his colleagues had discovered that single residues can contribute a large fraction of the binding free energy(62). The completed dataset for energetics of sidechain interactions determined by alanine-scanning mutagenesis are collected in ASEdb, including 91 protein-protein complexes and 2915 mutations(63). Keskin et al. had discovered that there is a correspondence between the experimental identified energy hotspot and the structurally conserved residues(28). In the future, we will want to explore the relationship between the conservation in contact residue pairs and experimental identified energy hotspots. On the other hand, we will modify the nr protein database, amino acid classification or the usage of pseudo count to improve the accuracy of our predictions.

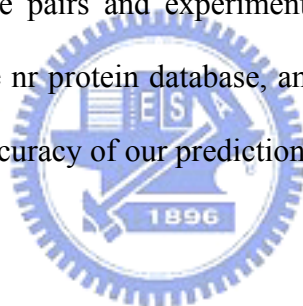


Table1. The compositions of protein sequence in NCBI RefSeq database (Release 16).

Taxonomy Class	Number of species	Number of proteins
Fungi	62	121,142
Invertebrate	182	87,244
Microbial	743	1,572,058
Mitochondrion	794	11,595
Vertebrate other	374	58,273
Plasmid	47	71,511
Plastid	51	5,274
Protozoa	65	114,427
Viral	1,578	45,001
Vertebrate mammalian	157	203,320



Table2. The frequency of amino acid occurs in protein surface and whole protein. The data of whole protein is downloaded from SWISSPROT database (<http://tw.expasy.org/sprot/relnotes/relstat.html>).

Amino acid	Surface (%)	Whole protein (%)
ILE	2.0	5.9
VAL	2.8	6.7
LEU	3.3	9.6
PHE	1.7	4.0
CYS	0.3	1.5
MET	1.3	2.4
ALA	6.2	7.8
GLA	7.5	6.9
THR	6.1	5.4
SER	7.0	6.8
TRP	0.5	1.1
TYR	2.1	3.1
PRO	6.0	4.8
HIS	2.2	2.3
ASN	6.7	4.2
GLN	5.9	4.0
ASP	9.6	5.3
GLU	11.7	6.7
LYS	10.7	5.9
ARG	6.2	5.4

Table 3. 24 pairs of related heterodimers with > 30% sequence identity but with pair coverage < 0.4.

Template	Related dimer	Pair coverage	IdeA ^a	IdeB ^b	DomainA ^c	DomainB ^d
1kxqBG	1kxtEF	0.00	100.0	69.4	c.1.8.1	b.1.1.1
1kxqBG	1kxvAC	0.00	100.0	62.1	c.1.8.1	b.1.1.1
1op9AB	1jtpAL	0.00	76.7	58.9	b.1.1.1	d.2.1.2
1ewyAC	1gaqAB	0.00	52.4	70.2	b.43.4.2, c.25.1.1	d.15.4.1
1op9AB	1p2cBC	0.00	49.1	60.5	b.1.1.1	d.2.1.2
1op9AB	1j1oHY	0.00	47.3	60.5	b.1.1.1	d.2.1.2
1jb0AD	1jb0BD	0.00	47.3	100.0	f.29.1.1	d.187.1.1
1jb0AC	1jb0BC	0.00	47.3	100.0	f.29.1.1	d.58.1.2
1jb0AF	1jb0BF	0.00	47.3	100.0	f.29.1.1	f.23.16.1
1op9AB	1bvkBC	0.00	46.4	60.5	b.1.1.1	d.2.1.2
1c17HL	1deeAD	0.00	42.2	60.2	b.1.1.1	b.1.1.1, b.1.1.2
1k3zAD	1iknCD	0.00	54.0	38.3	b.1.18.1	d.211.1.1
1op3HK	1uweLV	0.00	35.7	54.9	b.1.1.2	b.1.1.2
1bzqAL	1h0dBC	0.00	34.6	66.4	d.5.1.1	b.1.1.1
1mdaHM	2bbkJM	0.00	31.4	77.7	b.69.2.1	g.21.1.1
1dxrCL	1eysCM	0.02	48.2	32.0	a.138.1.2	f.26.1.1
1bqhAG	1bqhDK	0.03	100.0	100.0	b.1.1.2	b.1.1.1
1hezAE	1hezCE	0.03	100.0	100.0	b.1.1.1	d.15.7.1
1s6bAB	1oqsAB	0.05	45.6	43.5	a.133.1.2	a.133.1.2
1op9AB	1fbiHX	0.12	47.9	56.6	b.1.1.1	d.2.1.2
1bd2AD	1mi5AD	0.14	84.3	56.0	d.19.1.1	b.1.1.1
1r8sAE	1re0AB	0.30	81.5	37.7	c.37.1.8	a.118.3.1
1abrAB	1m2tAB	0.38	40.3	52.7	d.165.1.1	b.42.2.1, b.42.2.1
1hcfAX	1wwwWX	0.39	51.4	44.4	g.17.1.3	b.1.1.4

^a The sequence identity between the first chain of template and the first chain of protein of the related dimer.

^b The sequence identity between the second chain of template and the second chain of protein of the related dimer.

^c The interacting domains in the first chain of template.

^d The interacting domains in the second chain of template.

Table 4. Average precisions and false positive rates of specific interfacial energy and general interfacial energy on 182 queries. The unannotated candidates are considered as negative.

Query #	PDB ID	No. of Candidates	No. of Positives	No. of Negatives	AP (specific)	FP (specific)	AP (general)	FP (general)
1	1a0rBP	515	2	513	0.83	0.00	0.20	0.01
2	1a2kAD	74	2	72	0.25	0.06	0.03	0.66
3	1a6dAB	66	1	65	0.07	0.20	0.07	0.20
4	1a9nAB	264	2	262	0.53	0.05	0.51	0.23
5	1agrDH	18	1	17	0.50	0.06	0.08	0.71
6	1aisAB	4	2	2	0.58	0.50	0.50	0.75
7	1auiAB	117	5	112	0.81	0.19	0.58	0.22
8	1b34AB	10	5	5	0.51	0.52	0.61	0.52
9	1b7tAZ	30	8	22	0.73	0.16	0.30	0.47
10	1bi8AB	2200	6	2194	0.00	0.44	0.00	0.66
11	1buhAB	110	1	109	1.00	0.00	0.05	0.17
12	1c9bMN	6	2	4	0.83	0.13	0.75	0.25
13	1d3bEF	36	11	25	0.57	0.26	0.36	0.38
14	1dceAB	9	3	6	0.57	0.50	0.63	0.33
15	1dkgAD	15	1	14	0.50	0.07	0.50	0.07
16	1dn1AB	28	4	24	0.25	0.43	0.42	0.38
17	1doaAB	37	3	34	1.00	0.00	0.30	0.13
18	1e79AG	4	2	2	0.83	0.25	1.00	0.00
19	1eesAB	105	3	102	0.28	0.05	0.53	0.11
20	1eqzAB	6	1	5	0.25	0.60	0.25	0.60
21	1f3mAC	318	1	317	0.01	0.24	0.01	0.59
22	1f5qCD	1100	13	1087	0.03	0.21	0.04	0.33
23	1fbvAC	60	1	59	0.03	0.51	0.04	0.46
24	1finAB	1430	17	1413	0.14	0.35	0.03	0.40
25	1foeAB	185	3	182	0.16	0.14	0.06	0.18
26	1fq1AB	550	4	546	0.01	0.36	0.01	0.73
27	1fqvOP	14	8	6	0.88	0.27	0.88	0.27
28	1fxtAB	70	1	69	0.14	0.09	0.50	0.01
29	1g0uBJ	103	34	69	0.34	0.54	0.51	0.43
30	1g0uHI	103	34	69	0.28	0.59	0.37	0.45
31	1g0uLM	103	34	69	0.32	0.50	0.51	0.40
32	1g3nAB	2310	6	2304	0.01	0.33	0.00	0.45
33	1g3nEG	1210	14	1196	0.09	0.17	0.01	0.59
34	1g65DE	103	34	69	0.36	0.46	0.35	0.50
35	1g65IJ	103	34	69	0.32	0.53	0.36	0.47
36	1g65IZ	103	34	69	0.41	0.47	0.43	0.40
37	1g65KW	103	34	69	0.41	0.44	0.32	0.51

Query #	PDB ID	No. of Candidates	No. of Positives	No. of Negatives	AP (specific)	FP (specific)	AP (general)	FP (general)
38	1g65KX	103	34	69	0.48	0.36	0.37	0.46
39	1g65OP	103	34	69	0.38	0.50	0.36	0.54
40	1g65UV	103	34	69	0.29	0.56	0.55	0.34
41	1gl2AB	30	5	25	0.46	0.33	0.52	0.20
42	1gl2BC	6	4	2	1.00	0.00	0.82	0.50
43	1gotAB	1030	2	1028	0.25	0.47	0.00	0.40
44	1grnAB	374	6	368	0.17	0.10	0.03	0.36
45	1gw5AM	198	10	188	0.07	0.41	0.08	0.46
46	1gw5AS	110	5	105	0.08	0.42	0.12	0.31
47	1gw5BM	198	10	188	0.24	0.33	0.21	0.24
48	1gw5BS	110	5	105	0.21	0.31	0.16	0.40
49	1gw5MS	35	4	31	0.44	0.31	0.29	0.22
50	1h2tCZ	40	4	36	0.35	0.42	0.17	0.39
51	1h8eDG	4	2	2	0.75	0.50	0.75	0.50
52	1hq3AG	9	1	8	0.25	0.38	0.11	1.00
53	1hq3FH	4	1	3	0.25	1.00	0.25	1.00
54	1hr6EF	34	2	32	0.61	0.11	0.39	0.09
55	1i2mAB	148	1	147	0.50	0.01	0.14	0.04
56	1i50AB	154	2	152	0.42	0.02	0.29	0.02
57	1i50AK	77	2	75	0.58	0.01	0.58	0.01
58	1i50BI	6	2	4	0.50	0.38	0.75	0.25
59	1i7qAB	18	1	17	0.50	0.06	0.14	0.35
60	1ibrCD	666	7	659	0.40	0.06	0.13	0.16
61	1iruAG	105	34	71	0.44	0.42	0.39	0.48
62	1iruBC	105	34	71	0.38	0.48	0.38	0.50
63	1iruCD	105	34	71	0.38	0.46	0.33	0.52
64	1iruDE	105	34	71	0.40	0.43	0.36	0.48
65	1iruFG	104	34	70	0.36	0.49	0.35	0.50
66	1iruFN	104	34	70	0.39	0.42	0.48	0.36
67	1iruH2	105	34	71	0.36	0.49	0.31	0.55
68	1iruI1	105	34	71	0.27	0.61	0.29	0.56
69	1iruJK	105	34	71	0.58	0.27	0.33	0.48
70	1iruJ1	105	34	71	0.34	0.51	0.39	0.41
71	1iruKL	105	34	71	0.42	0.45	0.39	0.44
72	1iruLM	105	34	71	0.32	0.54	0.40	0.44
73	1iruNW	105	34	71	0.37	0.46	0.39	0.44
74	1iruOP	105	34	71	0.37	0.51	0.35	0.52
75	1iruS1	105	34	71	0.30	0.53	0.40	0.49
76	1iru12	105	34	71	0.27	0.60	0.33	0.48
77	1iw7CD	9	3	6	0.67	0.39	0.61	0.44

Query #	PDB ID	No. of Candidates	No. of Positives	No. of Negatives	AP (specific)	FP (specific)	AP (general)	FP (general)
78	1j2qDL	105	34	71	0.42	0.39	0.40	0.44
79	1j7dAB	119	4	115	0.28	0.54	0.27	0.59
80	1jatAB	116	4	112	0.04	0.55	0.03	0.72
81	1jfiAB	12	2	10	0.70	0.15	0.42	0.20
82	1jm7AB	11	2	9	0.63	0.33	0.37	0.28
83	1jr3BE	456	23	433	0.08	0.47	0.25	0.28
84	1jr3CD	31	5	26	0.41	0.32	0.48	0.18
85	1k5dDF	555	4	551	0.57	0.16	0.24	0.17
86	1k5dJK	111	3	108	0.32	0.06	0.56	0.02
87	1k83AI	231	3	228	0.51	0.01	0.47	0.01
88	1k8kAB	65	3	62	0.49	0.13	0.11	0.37
89	1k8kAD	11	2	9	0.58	0.11	0.25	0.50
90	1k8kAE	10	2	8	0.58	0.13	0.64	0.31
91	1k8kBF	10	2	8	0.83	0.06	0.58	0.13
92	1k8kBG	10	2	8	0.83	0.06	1.00	0.00
93	1k8kCF	81	1	80	1.00	0.00	1.00	0.00
94	1keeGH	80	1	79	0.25	0.04	0.33	0.03
95	1kfuLS	24	1	23	1.00	0.00	1.00	0.00
96	1ki1AB	180	3	177	0.24	0.05	0.09	0.16
97	1kx5EF	6	1	5	0.25	0.60	0.25	0.60
98	1kyoAB	34	2	32	0.45	0.06	0.24	0.14
99	1l4aBD	6	1	5	0.33	0.40	0.17	1.00
100	1lb1CD	148	3	145	0.18	0.08	0.08	0.16
101	1ldjAB	44	5	39	0.94	0.01	0.86	0.03
102	1ltxAR	6	1	5	1.00	0.00	0.50	0.20
103	1mljEF	49	3	46	0.07	0.55	0.05	0.85
104	1m2vAB	13	3	10	0.25	0.60	0.37	0.33
105	1n1jAB	20	2	18	0.63	0.17	0.38	0.19
106	1n4pCD	6	3	3	0.64	0.33	0.81	0.22
107	1ni4AD	13	1	12	0.13	0.58	0.10	0.75
108	1nt2AB	3	2	1	0.83	0.50	0.83	0.50
109	1nvwRS	148	2	146	0.38	0.02	0.04	0.24
110	1oe9AB	35	8	27	0.83	0.13	0.53	0.19
111	1ofhCI	50	5	45	0.22	0.33	0.10	0.68
112	1p22AB	2	1	1	1.00	0.00	1.00	0.00
113	1pp9AB	36	2	34	0.38	0.10	0.18	0.19
114	1q5qAI	69	19	50	0.48	0.26	0.34	0.43
115	1qbkBC	1110	9	1101	0.20	0.08	0.02	0.23
116	1qdlAB	16	1	15	1.00	0.00	0.33	0.13
117	1qgkAB	23	2	21	0.58	0.26	0.70	0.07

Query #	PDB ID	No. of Candidates	No. of Positives	No. of Negatives	AP (specific)	FP (specific)	AP (general)	FP (general)
118	1qs0AB	13	1	12	0.09	0.83	0.14	0.50
119	1qviAY	35	8	27	0.49	0.19	0.32	0.37
120	1r4mHL	56	1	55	0.14	0.11	0.14	0.11
121	1rypBI	103	34	69	0.33	0.51	0.33	0.63
122	1rypCD	103	34	69	0.39	0.46	0.41	0.52
123	1rypFN	103	34	69	0.45	0.41	0.40	0.42
124	1rypHI	103	34	69	0.35	0.59	0.38	0.45
125	1rypH2	103	34	69	0.40	0.42	0.37	0.50
126	1rypI1	103	34	69	0.32	0.52	0.38	0.49
127	1rypI2	103	34	69	0.57	0.33	0.38	0.39
128	1rypLM	103	34	69	0.33	0.50	0.36	0.44
129	1rypS1	103	34	69	0.59	0.32	0.55	0.35
130	1s3sFG	4	1	3	1.00	0.00	1.00	0.00
131	1s63AB	6	3	3	0.81	0.22	0.83	0.33
132	1sfcAD	10	3	7	0.57	0.43	0.39	0.52
133	1sfcBC	6	2	4	0.83	0.13	1.00	0.00
134	1sfcEF	30	5	25	0.19	0.58	0.24	0.41
135	1sxjAB	850	38	812	0.31	0.36	0.28	0.30
136	1sxjAE	31	1	30	0.33	0.07	0.50	0.03
137	1sxjBC	813	38	775	0.36	0.34	0.26	0.34
138	1sxjCD	811	40	771	0.30	0.34	0.32	0.35
139	1sxjDE	688	31	657	0.28	0.31	0.30	0.26
140	1t2kCD	9	1	8	0.25	0.38	0.20	0.50
141	1tafAB	2	1	1	1.00	0.00	1.00	0.00
142	1tcoAC	52	2	50	0.06	0.45	0.05	0.59
143	1tt5AB	44	4	40	0.42	0.49	0.36	0.50
144	1tvkAB	15	1	14	0.20	0.29	0.50	0.07
145	1u7eAB	216	5	211	0.32	0.15	0.19	0.11
146	1ukvGY	34	3	31	0.59	0.09	0.83	0.03
147	1umcCD	13	1	12	0.09	0.83	0.08	1.00
148	1ur6AB	30	1	29	0.10	0.31	0.09	0.34
149	1v11AB	13	1	12	0.09	0.83	0.11	0.67
150	1vg0AB	70	13	57	0.87	0.04	0.55	0.13
151	1vrqAB	64	1	63	0.02	0.98	0.02	1.00
152	1w0jCD	10	3	7	0.58	0.43	0.67	0.33
153	1w85CD	13	1	12	0.08	0.92	0.09	0.83
154	1w98AB	1199	17	1182	0.21	0.17	0.05	0.35
155	1wa5AC	370	7	363	0.03	0.39	0.02	0.50
156	1wa5BC	184	1	183	0.02	0.33	0.05	0.11
157	1wq1RG	148	2	146	0.01	0.82	0.01	0.82

Query #	PDB ID	No. of Candidates	No. of Positives	No. of Negatives	AP (specific)	FP (specific)	AP (general)	FP (general)
158	1xcgAB	180	3	177	0.17	0.06	0.04	0.30
159	1xewXY	133	6	127	0.59	0.02	0.64	0.02
160	1xo2AB	1320	17	1303	0.26	0.22	0.02	0.45
161	1y56AB	90	1	89	0.04	0.28	0.01	1.00
162	1y8qCD	45	4	41	0.33	0.54	0.14	0.70
163	1y8rBC	18	2	16	0.64	0.16	0.36	0.25
164	1ya7CJ	105	34	71	0.42	0.41	0.34	0.52
165	1z2cAB	76	3	73	0.72	0.02	0.15	0.42
166	1z5sAB	30	1	29	0.11	0.28	0.25	0.10
167	2b4sAB	452	5	447	0.05	0.15	0.03	0.19
168	2ba0AD	10	2	8	0.35	0.56	0.23	0.69
169	2ba0FI	19	4	15	0.60	0.22	0.52	0.22
170	2ba1AD	15	3	12	0.64	0.28	0.74	0.31
171	2ba1BG	15	4	11	0.36	0.45	0.26	0.70
172	2bcjAQ	1080	5	1075	0.01	0.54	0.00	1.00
173	2bkiAB	45	8	37	0.52	0.22	0.38	0.38
174	2bkuAB	888	6	882	0.08	0.15	0.13	0.20
175	2bl0AB	8	2	6	0.42	0.42	0.29	0.58
176	2bl0AC	9	2	7	0.33	0.43	0.27	0.57
177	2br2EF	19	4	15	0.51	0.32	0.29	0.40
178	2btfAP	11	1	10	1.00	0.00	0.50	0.10
179	2bykCD	16	2	14	0.63	0.21	0.39	0.21
180	2c35EF	2	1	1	1.00	0.00	1.00	0.00
181	2ey4AE	2	1	1	1.00	0.00	1.00	0.00
182	3gtuAB	9	1	8	0.14	0.75	0.13	0.88

Table 5. Average precisions and false positive rate of specific interfacial energy and general interfacial energy on 101 queries. The unannotated candidates are removed.

Query #	Template	No. of Candidates	No. of Positives	No. of Negatives	AP (specific)	FP (specific)	AP (general)	FP (general)
1	1a0rBP	130	2	128	1.00	0.00	0.58	0.01
2	1a2kAD	14	2	12	0.75	0.04	0.11	1.04
3	1a6dAB	18	1	17	1.00	0.00	1.00	0.00
4	1a9nAB	43	2	41	0.60	0.10	0.53	0.33
5	1auiAB	51	5	46	0.82	0.20	0.65	0.22
6	1b7tAZ	13	8	5	0.97	0.08	0.69	0.45
7	1bi8AB	181	6	175	0.22	0.35	0.03	0.63
8	1buhAB	46	1	45	1.00	0.00	0.11	0.18
9	1dkgAD	12	1	11	1.00	0.00	1.00	0.00
10	1dn1AB	8	4	4	0.75	0.44	0.68	0.44
11	1doaAB	15	3	12	1.00	0.00	0.64	0.08
12	1eesAB	28	3	25	0.70	0.05	0.83	0.04
13	1f3mAC	76	1	75	0.10	0.12	0.05	0.24
14	1f5qCD	224	13	211	0.10	0.28	0.22	0.38
15	1finAB	363	17	346	0.24	0.32	0.10	0.37
16	1foeAB	16	3	13	0.40	0.38	0.23	0.51
17	1fq1AB	112	4	108	0.04	0.52	0.03	0.75
18	1fxtAB	19	1	18	1.00	0.00	1.00	0.00
19	1g3nAB	203	6	197	0.12	0.28	0.04	0.43
20	1g3nEG	226	14	212	0.18	0.23	0.05	0.66
21	1gl2AB	12	5	7	0.68	0.31	0.87	0.14
22	1gl2BC	5	4	1	1.00	0.00	0.89	0.50
23	1gotAB	164	2	162	0.51	0.47	0.04	0.40
24	1grnAB	69	6	63	0.61	0.10	0.16	0.33
25	1gw5AM	46	10	36	0.32	0.37	0.32	0.44
26	1gw5AS	29	5	24	0.28	0.38	0.37	0.23
27	1gw5BM	52	10	42	0.44	0.39	0.44	0.32
28	1gw5BS	33	5	28	0.45	0.29	0.33	0.51
29	1h2tCZ	11	4	7	0.50	0.61	0.44	0.46
30	1i2mAB	48	1	47	1.00	0.00	0.25	0.06
31	1i50AB	28	2	26	1.00	0.00	1.00	0.00
32	1i50AK	15	2	13	1.00	0.00	1.00	0.00
33	1i7qAB	6	1	5	0.50	0.20	0.25	0.60
34	1ibrCD	134	7	127	0.62	0.04	0.24	0.18
35	1j7dAB	41	4	37	0.32	0.58	0.31	0.70
36	1jatAB	41	4	37	0.13	0.47	0.08	0.76
37	1jfiAB	5	2	3	1.00	0.00	1.00	0.00

Query #	Template	No. of Candidates	No. of Positives	No. of Negatives	AP (specific)	FP (specific)	AP (general)	FP (general)
38	1jm7AB	3	2	1	1.00	0.00	1.00	0.00
39	1jr3BE	80	23	57	0.39	0.52	0.64	0.39
40	1jr3CD	14	5	9	0.65	0.27	0.85	0.16
41	1k5dDF	117	4	113	0.64	0.19	0.68	0.21
42	1k5dJK	46	3	43	0.40	0.08	0.64	0.02
43	1k83AI	29	3	26	1.00	0.00	1.00	0.00
44	1k8kAB	6	3	3	0.76	0.33	0.56	0.56
45	1k8kAD	5	2	3	0.58	0.33	0.37	0.83
46	1k8kBF	3	2	1	1.00	0.00	1.00	0.00
47	1k8kBG	3	2	1	1.00	0.00	1.00	0.00
48	1k8kCF	31	1	30	1.00	0.00	1.00	0.00
49	1keeGH	8	1	7	1.00	0.00	1.00	0.00
50	1kfuLS	7	1	6	1.00	0.00	1.00	0.00
51	1kilAB	15	3	12	0.67	0.14	0.25	0.50
52	1l4aBD	2	1	1	1.00	0.00	0.50	1.00
53	1lb1CD	15	3	12	0.36	0.39	0.25	0.50
54	1ldjAB	7	5	2	0.94	0.20	0.88	0.40
55	1mljEF	18	3	15	0.23	0.44	0.13	0.82
56	1nljAB	7	2	5	0.83	0.10	0.83	0.10
57	1ni4AD	7	1	6	0.20	0.67	0.14	1.00
58	1nvwRS	14	2	12	0.75	0.04	1.00	0.00
59	1oe9AB	15	8	7	0.92	0.11	0.80	0.18
60	1qbkBC	186	9	177	0.36	0.08	0.16	0.18
61	1qdlAB	6	1	5	1.00	0.00	0.50	0.20
62	1qs0AB	7	1	6	0.17	0.83	0.25	0.50
63	1qviAY	15	8	7	0.60	0.36	0.53	0.52
64	1r4mHL	11	1	10	1.00	0.00	1.00	0.00
65	1s3sFG	4	1	3	1.00	0.00	1.00	0.00
66	1sfcAD	7	3	4	0.67	0.42	0.42	0.75
67	1sfcEF	12	5	7	0.66	0.46	0.73	0.23
68	1sxjAB	137	38	99	0.56	0.39	0.63	0.33
69	1sxjBC	137	38	99	0.64	0.37	0.60	0.38
70	1sxjCD	138	40	98	0.60	0.35	0.62	0.37
71	1sxjDE	115	31	84	0.59	0.32	0.67	0.31
72	1tcoAC	23	2	21	0.20	0.29	0.11	0.57
73	1tt5AB	7	4	3	0.77	0.50	0.77	0.50
74	1tvkAB	5	1	4	0.50	0.25	0.50	0.25
75	1u7eAB	25	5	20	0.50	0.31	0.55	0.23
76	1ukvGY	14	3	11	0.67	0.15	0.87	0.06
77	1umcCD	7	1	6	0.20	0.67	0.14	1.00

Query #	Template	No. of Candidates	No. of Positives	No. of Negatives	AP (specific)	FP (specific)	AP (general)	FP (general)
78	1ur6AB	8	1	7	1.00	0.00	0.33	0.29
79	1v11AB	7	1	6	0.17	0.83	0.20	0.67
80	1vg0AB	32	13	19	0.92	0.06	0.66	0.17
81	1vrqAB	21	1	20	0.05	0.95	0.05	1.00
82	1w85CD	7	1	6	0.14	1.00	0.17	0.83
83	1w98AB	273	17	256	0.42	0.18	0.13	0.39
84	1wa5AC	50	7	43	0.15	0.49	0.12	0.64
85	1wa5BC	19	1	18	0.10	0.50	0.14	0.33
86	1wq1RG	13	2	11	0.33	0.27	0.27	0.36
87	1xcgAB	15	3	12	0.50	0.17	0.22	0.56
88	1xewXY	10	6	4	1.00	0.00	1.00	0.00
89	1xo2AB	317	17	300	0.48	0.17	0.16	0.42
90	1y56AB	27	1	26	0.14	0.23	0.04	1.00
91	1y8qCD	7	4	3	0.71	0.50	0.62	0.75
92	1y8rBC	4	2	2	0.75	0.50	0.83	0.25
93	1z5sAB	8	1	7	0.25	0.43	0.50	0.14
94	2b4sAB	95	5	90	0.26	0.13	0.17	0.16
95	2bcjAQ	124	5	119	0.06	0.46	0.02	1.02
96	2bkiAB	15	8	7	0.82	0.21	0.80	0.32
97	2bkuAB	187	6	181	0.35	0.13	0.48	0.19
98	2bl0AB	6	2	4	0.50	0.38	0.42	0.50
99	2bl0AC	6	2	4	0.42	0.50	0.42	0.50
100	2bykCD	6	2	4	0.83	0.13	1.00	0.00
101	3gtuAB	3	1	2	0.33	1.00	0.33	1.00

Table 6. The result of 1a2kAD to search yeast proteome.

Homologs of 1a2kA	Homologs of 1a2kD	Exp ^a	SP energy ^b	SP energy (normal)	GE energy ^c	IDE1 ^d	IDE ^e	Function1 ^f	Function 2 ^g
NTF2	GSP1	P	68.4	0.5	20.7	43	80	Nuclear envelope protein in nucleocytoplasmic transport	GTP binding protein involve in nuclear organization
NTF2	GSP2	P	68.4	0.5	20.7	43	79	Nuclear envelope protein in nucleocytoplasmic transport	GTP binding protein involve in nuclear organization
BRE5	YPT6	N	15.6	0.11	-31.2	21	26	Ubiquitin protease cofactor, coregulate vesicle transport	GTPase, involved in the secretory pathway
BRE5	GSP1	N	11	0.08	-22.2	21	80	Ubiquitin protease cofactor, coregulate vesicle transport	GTP binding protein involve in nuclear organization
BRE5	GSP2	N	11	0.08	-22.2	21	79	Ubiquitin protease cofactor, coregulate vesicle transport	GTP binding protein involve in nuclear organization
BRE5	YPT7	N	10.6	0.08	-24	21	24	Ubiquitin protease cofactor, coregulate vesicle transport	GTPase, required for homotypic fusion event
BRE5	RHO3	N	7.7	0.06	-18.6	21	24	Ubiquitin protease cofactor, coregulate vesicle transport	Non-essential small GTPase involved in the establishment of cell polarity.
BRE5	RHO2	N	6.7	0.05	-17.2	21	25	Ubiquitin protease cofactor, coregulate vesicle transport	Non-essential small GTPase of involved in microtubule assembly
BRE5	YPT31	N	5.7	0.04	-18.8	21	25	Ubiquitin protease cofactor, coregulate vesicle transport	GTPase, involved in the exocytic pathway;
BRE5	YPT11	N	4.4	0.03	-23.4	21	15	Ubiquitin protease cofactor, coregulate vesicle transport	Rab-type small GTPase mediate distribution of mitochondria to daughter cells
BRE5	TEM1	N	4.2	0.03	-27	21	21	Ubiquitin protease cofactor, coregulate vesicle transport	GTP-binding protein involved in termination of M-phase
BRE5	VPS21	N	2.7	0.02	-20.7	21	27	Ubiquitin protease cofactor, coregulate vesicle transport	GTPase required for transport during endocytosis
BRE5	SAR1	N	-23.2	-0.17	-7	21	20	Ubiquitin protease cofactor, coregulate vesicle transport	GTPase, component of COPII coat of vesicles
BRE5	MSS1	N	-35.2	-0.25	3.9	21	15	Ubiquitin protease cofactor, coregulate vesicle transport	Mitochondrial protein, modify the wobble uridine

^a Exp means the functional annotations for the candidates, ***P*** represent known interacting proteins interaction and ***N*** represent the non interacting proteins defined by Jasen et al.

^b SP energy is the abbreviation of the “specific interfacial energy” which is calculated from pairPSSM of 1a2kAD.

^c GE energy is the abbreviation of the “general interfacial energy” which is calculated from general empirical matrix.

^d IDE1 means the sequence identity percentage between the candidate protein and 1a2k A chain.

^e IDE2 means the sequence identity percentage between the candidate protein and 1a2k D chain.

^f The functional annotation for the protein of candidate homologous to 1a2k A chain.

^g The functional annotation for the protein of candidate homologous to 1a2k D chain.



Table 7. The pairPSSM of protein complex 1a2kAD. There are 47 pairs of contact residues. Each row represents 45 types of energy in a specific contact position (9 clusters in 20 amino acids result in 45 types cluster pair). The abbreviation A represent residue {Ala and Gly}; B: {Val, Met, Leu and Ile}; C: {Pro, Ser and Thr}; D: {Phe, Tyr and Trp}; E: {Cys}; F: {His and Arg}; G: {Lys}; H: {Asp and Glu}; I: {Asn and Gln}.

Contact position	AA	AB	AC	AD	AE	AF	AG	AH	AI	BB	BC	BD	BE	BF	BG	BH	BI	CC	CD	CE	CF	CG	CH	CI	DD	DE	DF	DG	DH	DI	EE	EF	EG	EH	EI	FF	FG	FH	FI	GG	GH	GI	HH	HI	II
0 DD	-1.0	0.1	-1.0	0.5	1.0	-0.5	-1.7	-1.7	-1.1	1.7	0.0	1.9	2.0	0.5	-0.6	-0.6	0.0	-0.9	0.6	1.0	-0.5	-1.5	-1.4	-1.0	5.3	2.3	0.9	-0.1	-0.1	0.3	4.2	1.5	0.3	-0.2	0.6	-0.1	-1.3	-0.3	-0.5	-2.2	-1.1	-1.6	-2.2	-1.5	-0.9
1 HD	-1.2	-0.1	-1.2	0.3	0.8	-0.8	-1.9	-1.9	-1.4	1.5	-0.2	1.6	1.7	0.3	-0.8	-0.4	-0.3	-1.2	0.5	0.8	-0.7	-1.7	-1.4	-1.2	1.9	2.1	0.7	1.0	3.3	0.1	4.0	1.3	0.0	-0.4	0.4	-0.3	-1.5	-0.6	-0.8	-2.5	-1.3	-1.8	-2.4	-1.7	-1.2
2 HA	-1.5	-0.4	-1.1	-0.1	0.5	-1.1	-2.3	1.3	-1.7	1.1	-0.5	1.2	1.4	-0.1	-1.1	-1.2	-0.6	-1.5	0.0	0.4	-1.0	-2.1	-0.5	-1.5	1.6	1.7	0.3	-0.7	-0.7	-0.2	3.7	1.0	-0.3	-0.8	0.1	-0.7	-0.2	2.3	-1.1	-0.5	-0.8	-1.1	-1.6	-2.1	-1.5
3 HF	-1.5	-0.4	-1.6	-0.1	0.5	-1.1	-1.2	0.6	-1.7	1.1	-0.5	1.2	1.4	-0.1	-1.1	-0.8	-0.6	-1.5	0.0	0.4	-0.6	-0.9	1.3	-1.5	1.6	1.7	0.3	-0.7	-0.7	-0.2	3.7	1.0	-0.3	-0.8	0.1	-0.7	-0.7	1.8	-1.1	-2.8	-0.8	-2.1	-2.7	-2.1	-1.5
4 BG	-1.4	-0.3	-1.4	0.1	0.6	-0.9	-2.1	-2.1	-1.6	1.3	-0.4	1.4	1.6	3.2	2.2	-0.7	-0.5	-1.3	0.1	0.6	-0.9	-1.9	-1.8	-1.4	1.7	1.9	1.4	0.2	-0.5	-0.1	3.8	1.1	-0.1	-0.6	0.2	-0.5	-1.7	-0.7	-0.4	-2.6	-1.5	-1.1	-2.6	-1.9	-1.3
5 CG	-1.4	-0.4	-1.5	0.0	0.5	-0.1	-1.2	-2.2	-1.6	1.2	-0.4	1.3	1.5	0.8	-0.5	-1.1	-0.5	-1.4	0.0	0.5	2.5	1.5	-1.5	-1.5	1.6	1.8	0.4	-0.6	-0.6	-0.2	3.8	1.0	-0.2	-0.7	0.1	-0.6	-1.7	-0.8	-1.0	-2.7	-1.6	-1.1	-2.7	-2.0	-1.4
6 DG	-1.2	-0.1	-1.2	0.3	0.8	-0.8	-1.9	-1.9	-1.4	1.5	-0.2	1.6	1.7	0.3	-0.8	-0.8	-0.3	-1.2	0.3	0.8	0.1	-0.9	-1.6	-1.2	1.9	2.1	3.8	2.9	0.0	0.1	4.0	1.3	0.0	-0.4	0.4	-0.3	-1.5	-0.6	-0.8	-2.5	-1.3	-1.8	-2.4	-1.7	-1.2
7 DD	-1.0	0.1	-1.0	0.5	1.0	-0.5	-1.7	-1.7	-1.1	1.7	0.0	1.9	2.0	0.5	-0.6	-0.6	0.0	-0.9	1.2	1.0	-0.5	-1.5	-1.4	-1.0	5.3	2.3	0.9	-0.1	-0.1	0.3	4.2	1.5	0.3	-0.2	0.6	-0.1	-1.3	-0.3	-0.5	-2.2	-1.1	-1.6	-2.2	-1.5	-0.9
8 BD	-1.1	0.0	-1.1	0.4	0.9	-0.6	-1.8	-1.8	-1.3	1.8	0.0	3.9	1.9	0.4	-0.7	-0.7	-0.2	-1.1	0.4	0.9	-0.6	-1.6	-1.5	-1.1	2.0	5.9	0.8	-0.3	-0.2	0.2	4.1	1.4	0.2	-0.3	0.5	-0.2	-1.4	-0.5	-0.6	-2.4	-1.2	-1.7	-2.3	-1.6	-1.0
9 BD	-1.0	0.1	-1.0	0.5	1.0	-0.5	-1.7	-1.7	-1.1	1.8	0.1	3.9	2.0	0.5	-0.6	-0.6	0.0	-0.9	0.5	1.0	-0.5	-1.5	-1.4	-1.0	2.1	2.3	0.9	-0.1	-0.1	0.3	4.2	1.5	0.3	-0.2	0.6	-0.1	-1.3	-0.3	-0.5	-2.2	-1.1	-1.6	-2.2	-1.5	-0.9
10 AD	-1.3	-0.2	-1.3	3.3	0.7	-0.8	-2.0	-2.0	-1.5	1.5	0.0	2.7	1.7	0.2	-0.9	-0.9	-0.4	-1.3	1.4	0.7	-0.8	-1.8	-1.7	-1.3	1.8	5.8	0.6	-0.5	-0.5	0.0	3.9	1.2	0.0	-0.5	0.3	-0.4	-1.6	-0.7	-0.9	-2.6	-1.4	-1.9	-2.5	-1.8	-1.2
11 HG	-1.2	-0.1	-1.2	0.3	0.8	1.3	0.0	-1.9	-1.4	1.5	-0.2	1.6	1.7	0.3	-0.8	-0.8	-0.3	-1.2	0.3	0.8	-0.7	-1.7	-1.6	-1.2	1.9	2.1	0.7	-0.4	-0.4	0.1	4.0	1.3	0.0	-0.4	0.4	-0.3	-1.5	2.1	-0.8	-2.5	1.2	-1.8	-1.7	-1.7	-1.2
12 HA	-0.6	-0.1	-1.2	0.3	0.8	-0.8	-1.9	2.0	0.5	1.5	-0.2	1.6	1.7	0.3	-0.8	-0.8	-0.3	-1.2	0.3	0.8	-0.7	-1.7	-1.4	-1.2	1.9	2.1	0.7	-0.4	-0.4	0.1	4.0	1.3	0.0	1.4	0.4	-0.3	-1.5	-0.6	-0.8	-2.5	-1.3	-1.8	-2.4	-1.7	-1.2
13 HA	-0.6	0.0	-1.1	0.4	0.9	-0.6	-1.8	1.9	0.4	1.6	-0.1	1.7	1.9	0.4	-0.7	-0.7	-0.2	-1.1	0.4	0.9	-0.6	-1.6	-1.3	-1.1	2.0	2.2	0.8	-0.3	-0.2	0.2	4.1	1.4	0.2	-0.3	0.5	-0.2	-1.4	-0.5	-0.6	-2.4	-1.2	-1.7	-2.3	-1.6	-1.0
14 HG	-1.3	-0.2	-1.3	0.2	0.7	-0.8	-1.2	-2.0	-1.5	1.4	-0.3	1.5	1.7	0.2	-0.9	-0.9	-0.4	-1.3	0.2	0.7	-0.8	-1.8	-1.7	-1.3	1.8	2.0	0.6	-0.5	-0.5	0.0	3.9	1.2	0.0	-0.5	0.3	-0.4	-1.6	2.3	0.7	-2.6	1.3	-0.1	-1.7	-1.8	-1.2
15 CD	-1.9	-0.8	-1.3	0.2	0.1	-1.4	-2.6	-2.6	-0.6	0.8	-0.5	0.9	1.1	-0.4	-1.5	-1.5	-1.0	0.4	2.4	2.0	1.5	-1.4	-2.3	1.4	1.2	1.4	1.6	0.1	-1.0	2.3	3.3	0.6	-0.6	-1.1	-0.3	-1.0	-2.2	-1.3	-0.1	-3.2	-2.0	-2.5	-3.1	-2.4	0.4
16 CH	-0.6	-0.5	0.6	-0.1	0.4	-0.2	-2.3	-0.9	-0.7	1.1	-0.6	1.1	1.3	-0.1	-1.2	-1.2	-0.7	-1.0	-0.1	0.4	-1.1	-2.2	1.3	-1.6	1.5	1.7	0.3	-0.8	-0.8	-0.3	3.6	0.9	-0.4	-0.8	0.0	-0.7	-1.9	1.3	-1.2	-2.9	-0.8	-2.2	-2.8	0.8	-1.6
17 CA	0.3	-0.3	1.9	0.1	0.6	1.7	-0.7	-2.1	1.2	1.3	-0.4	1.4	1.6	0.1	-1.0	-1.0	-0.5	-1.3	0.1	1.8	-0.9	-1.9	-1.8	-1.0	1.7	1.9	0.5	-0.5	-0.5	-0.1	3.8	1.1	-0.1	-0.6	0.2	-0.5	-1.7	-0.7	-0.9	-2.6	-1.5	-2.0	-2.6	-1.9	-1.3
18 BD	-1.8	-0.7	0.5	1.1	0.2	-1.4	-2.6	-2.6	-0.6	1.1	0.8	2.5	2.7	0.1	-1.4	-1.5	2.0	0.1	2.5	0.1	-0.4	-2.4	-2.3	0.1	1.2	1.4	0.0	-1.0	-1.0	0.2	3.4	0.7	-0.6	-1.1	-0.2	-1.0	-2.1	-1.2	-1.4	-3.1	-2.0	-2.4	-3.0	-2.4	-1.8
19 BD	-1.8	-0.7	-1.8	-0.3	0.2	-1.3	-2.5	-2.5	-2.0	1.2	1.7	2.9	1.2	-0.3	-1.4	-1.4	1.8	-1.7	-0.3	0.2	-0.7	-1.4	-2.2	-1.8	1.3	1.5	2.7	0.1	-0.9	1.8	3.4	0.7	2.4	-1.0	-0.2	0.8	-2.1	-1.2	0.8	-3.0	-1.9	0.4	-3.0	-2.3	-1.7
20 BA	-1.3	2.6	-1.3	0.2	0.7	1.8	0.5	-2.0	0.4	1.4	-0.1	1.5	1.7	0.2	-0.9	-0.9	-0.4	-1.3	0.2	0.7	-0.8	-1.8	-1.7	-1.3	1.8	2.0	0.6	-0.5	-0.5	0.0	3.9	1.2	0.0	-0.5	2.0	-0.4	-1.6	-0.7	-0.9	-2.6	-1.4	-1.9	-2.5	-1.8	-1.2
21 BA	-1.2	2.5	-1.2	0.3	0.8	1.7	0.3	-1.9	0.4	1.5	-0.2	1.6	1.7	0.3	-0.8	-0.8	-0.3	-1.2	0.3	0.8	-0.7	-1.3	-1.6	-1.2	1.9	2.1	0.7	-0.4	-0.4	0.1	4.0	1.3	0.0	-0.4	0.4	-0.3	-1.5	-0.6	-0.8	-2.5	-1.3	-1.8	-2.4	-1.7	-1.2
22 BI	-1.3	-0.2	-1.3	0.2	0.7	-0.8	-2.0	-2.0	-1.5	1.4	-0.3	1.5	1.7	0.2	-0.4	-0.9	2.7	-1.3	0.2	0.7	-0.8	-1.8	-1.7	-1.3	1.8	2.0	0.6	-0.5	-0.5	0.0	3.9	1.2	0.0	-0.5	0.3	-0.4	-0.7	-0.7	1.8	-2.6	-1.4	0.5	-2.5	-1.8	1.2
23 DD	-1.0	0.1	-1.0	0.5	1.0	-0.5	-1.7	-1.7	-1.1	1.7	0.0	1.9	2.0	0.5	-0.6	-0.6	0.0	-0.9	0.6	1.0	-0.5	-1.5	-1.4	-1.0	5.3	2.3	0.9	-0.1	-0.1	0.3	4.2	1.5	0.3	-0.2	0.6	-0.1	-1.3	-0.3	-0.5	-2.2	-1.1	-1.6	-2.2	-1.5	-0.9
24 FF	-1.2	-0.1	-1.2	0.3	0.8	1.5	-1.9	-1.9	-1.4	1.5	-0.2	1.6	1.7	0.5	-0.8	-0.8	-0.3	-1.2	0.3	0.8	2.0	-1.7	-1.6	-1.2	1.9	2.1	0.7	-0.4	-0.4	0.1	4.0	1.3	0.0	-0.4	0.4	3.1	0.4	-0.6	-0.8	-2.5	-1.3	-1.8	-2.4	-1.7	-1.2
25 BI	-1.1	0.0	-1.1	0.4	0.9	-0.6	-1.8	-1.8	-1.3	1.6	-0.1	1.7	1.9	0.4	-0.4	-0.7	2.9	-1.1	0.4	0.9	-0.6	-1.6	-1.5	-1.1	2.0	2.2	1.0	-0.3	-0.2	1.4	4.1	1.4	0.2	-0.3	0.5	-0.2	-1.4	-0.5	-0.6	-2.4	-1.2	-1.7	-2.3	-1.6	-1.0
26 BD	-1.1	0.0	-1.1	0.4	0.9	-0.6	-1.8	-1.8	-1.3	1.8	0.0	4.0	1.9	0.4	-0.7	-0.7	-0.2	-1.1	0.4	0.9	-0.6	-1.6	-1.5	-1.1	3.3	2.2	0.8	-0.3	-0.2	0.2	4.1	1.4	0.2	-0.3	0.5	-0.2	-1.4	-0.5	-0.6	-2.4	-1.2	-1.7	-2.3	-1.6	-1.0
27 BF	-1.4	1.5	-1.5	0.9	0.5	-1.0	-2.2	-2.2	-1.6	1.4	2.1	1.3	1.5	2.6	-0.1	-1.1	-0.5	-1.4	1.0	0.5	-0.9	-2.0	-1.9	-1.5	1.6	1.8	0.8	-0.6	-0.6	-0.2	3.8	1.0	-0.2	-0.7	0.1	-0.6	-1.7	-0.8	-1.0	-2.7	-1.6	-2.1	-2.7	-2.0	-1.4
28 AC	-1.2	-0.1	1.6	0.3	0.8	-0.8	-1.9	-1.9	-1.4	1.5	0.5	1.6	1.7	0.3	-0.8	-0.8	-0.3	-1.2	0.3	1.9	-0.7	-1.7	-1.0	1.5	1.9	2.1	0.7	-0.4	-0.4	0.1	4.0	1.3	0.0	-0.4	0.4	-0.3	-1.5	-0.6	-0.8	-2.5	-1.3	-1.8	-2.4	-1.7	-1.2
29 BA	-1.2	2.3	-1.2	0.6	0.8	-0.8	-1.9	-1.9	-1.4	1.5	1.9	1.6	1.7	1.4	-0.8	-0.8	0.3	-1.2	0.3	0.8	-0.7	-1.7	-1.6	-1.2	1.9	2.1	0.7	-0.4	-0.4	0.1	4.0	1.3	0.0	-0.4	0.4	-0.3	-1.5	-0.6	-0.8	-2.5	-1.3	-1.8	-2.4	-1.7	-1.2
30 BC	-0.8	0.3	-0.9	0.6	1.2	-0.4	-1.6	-1.6	-1.0	1.8	2.2	1.9	2.1	0.6	-0.5	-0.5	0.1	-0.8	0.8	1.1	-0.3	-1.4	-1.3	-0.8	2.2	2.4	1.0	0.0	0.0	0.5	4.4	1.7	0.4	-0.1	0.7	0.0	-1.1	-0.2	-0.4	-2.1	-1.0	-1.4	-2.0	-1.4	-0.8

Contact position	AA	AB	AC	AD	AE	AF	AG	AH	AI	BB	BC	BD	BE	BF	BG	BH	BI	CC	CD	CE	CF	CG	CH	CI	DD	DE	DF	DG	DH	DI	EE	EF	EG	EH	EI	FF	FG	FH	FI	GG	GH	GI	HH	HI	II	
31	BB	-0.8	0.3	-0.9	0.6	1.2	-0.4	-1.6	-1.6	-1.0	3.6	0.2	2.0	2.1	0.6	-0.5	-0.5	0.1	-0.8	0.7	1.1	-0.3	-1.4	-1.3	-0.8	2.2	2.4	1.0	0.0	0.0	0.5	4.4	1.7	0.4	-0.1	0.7	0.0	-1.1	-0.2	-0.4	-2.1	-1.0	-1.4	-2.0	-1.4	-0.8
32	BA	-0.8	2.5	-0.9	0.8	1.2	-0.4	-1.6	-1.6	-1.0	1.8	0.2	1.9	2.1	0.6	-0.5	-0.5	0.1	-0.8	0.7	1.1	-0.3	-1.4	-1.3	-0.8	2.2	2.4	1.0	0.0	0.0	0.5	4.4	1.7	0.4	-0.1	0.7	0.0	-1.1	-0.2	-0.4	-2.1	-1.0	-1.4	-2.0	-1.4	-0.8
33	BI	-1.0	0.1	-1.0	0.5	1.0	-0.5	-1.7	-1.7	-1.1	1.7	0.0	1.8	2.0	0.7	-0.6	-0.6	2.8	-0.9	0.5	1.0	-0.5	-1.5	-1.4	-1.0	2.1	2.3	0.9	-0.1	-0.1	0.6	4.2	1.5	0.3	-0.2	0.6	-0.1	-1.3	-0.3	-0.5	-2.2	-1.1	-1.6	-2.2	-1.5	-0.9
34	BF	-1.2	1.6	-1.2	0.3	0.8	-0.8	-1.9	-1.9	-1.4	1.6	2.0	1.6	1.7	2.4	-0.8	-0.8	-0.3	-1.2	0.3	0.8	-0.7	-1.7	-1.6	-1.2	1.9	2.1	1.0	-0.4	-0.4	0.1	4.0	1.3	0.0	-0.4	0.4	-0.3	-1.5	-0.6	-0.8	-2.5	-1.3	-1.8	-2.4	-1.7	-1.2
35	FC	-0.8	0.3	-0.9	0.6	1.2	-0.4	-1.6	-1.6	-1.0	1.8	0.2	1.9	2.1	0.6	-0.5	-0.5	0.1	-0.8	0.7	1.1	2.2	-1.4	-1.3	-0.3	2.2	2.4	1.0	0.0	0.0	0.5	4.4	1.7	0.4	-0.1	0.7	0.0	-1.1	-0.2	-0.4	-2.1	-1.0	-1.4	-2.0	-1.4	-0.8
36	FB	-1.0	0.1	-1.0	0.5	1.0	-0.5	-1.7	-1.7	-1.1	1.7	0.0	1.8	2.0	3.2	-0.6	-0.6	0.6	-0.9	0.5	1.0	-0.5	-1.5	-1.4	-1.0	2.1	2.3	1.2	-0.1	-0.1	0.3	4.2	1.5	0.3	-0.2	0.6	-0.1	-1.3	-0.3	-0.5	-2.2	-1.1	-1.6	-2.2	-1.5	-0.9
37	FF	-1.4	-0.3	-1.4	0.1	0.6	1.7	-2.1	-2.1	-0.8	1.3	-0.4	1.4	1.6	0.5	-1.0	-1.0	-0.5	-1.3	0.1	0.6	2.2	-1.9	-1.8	-0.5	1.7	1.9	0.5	-0.5	-0.5	-0.1	3.8	1.1	-0.1	-0.6	0.2	3.1	-1.7	-0.7	-0.3	-2.6	-1.5	-2.0	-2.6	-1.9	-1.3
38	IB	-1.1	0.0	-1.1	0.4	0.9	-0.6	-1.8	-1.8	-1.3	1.6	-0.1	1.7	1.9	0.9	-0.7	0.1	2.9	-1.1	0.4	0.9	-0.6	-1.6	-1.5	-1.1	2.0	2.2	0.8	-0.3	-0.2	0.6	4.1	1.4	0.2	-0.3	0.5	-0.2	-1.4	-0.5	-0.6	-2.4	-1.2	-1.7	-2.3	-1.6	-1.0
39	IF	-1.5	-0.4	-1.6	-0.1	0.5	-0.4	-2.3	-2.3	-0.4	1.1	-0.5	1.2	1.4	-0.1	-1.1	-1.2	-0.1	-1.5	0.0	0.4	-0.5	-2.1	-1.1	1.9	1.6	1.7	0.3	-0.7	-0.7	-0.2	3.7	1.0	-0.3	-0.8	0.1	1.4	-1.8	-0.1	1.9	-2.8	-1.7	-2.1	-2.7	-2.1	-1.5
40	IH	-1.6	-0.5	-1.7	-0.1	0.4	-1.2	-2.3	-2.3	-0.4	1.1	-0.6	1.1	1.3	-0.1	-1.2	-1.2	-0.7	-1.6	-0.1	0.4	0.1	-2.2	-1.4	1.6	1.5	1.7	0.3	-0.8	-0.8	-0.3	3.6	0.9	-0.4	-0.8	0.0	0.5	-1.9	0.1	1.3	-2.9	-1.7	-2.2	-1.4	1.0	2.2
41	IA	-1.4	-0.4	-1.5	0.0	0.5	-0.4	-2.2	-1.0	1.9	1.2	-0.4	1.3	1.5	0.7	-1.1	-0.3	1.6	-1.4	0.0	0.5	-0.1	-2.0	-1.9	1.6	1.6	1.8	0.4	-0.6	-0.6	-0.2	3.8	1.0	-0.2	-0.7	0.1	-0.6	-1.7	-0.8	-1.0	-2.7	-1.6	-2.1	-2.7	-2.0	-1.4
42	ID	-1.1	0.0	-1.1	0.4	0.9	-0.6	-1.8	-1.8	-1.3	1.6	-0.1	1.7	1.9	0.4	-0.7	-0.7	-0.2	-1.1	0.4	0.9	-0.6	-1.6	-1.2	-1.1	2.0	2.2	1.8	-0.3	0.5	3.6	4.1	1.4	0.2	-0.3	0.5	-0.2	-1.4	-0.5	-0.6	-2.4	-1.2	-1.7	-2.3	-1.6	-1.0
43	DB	-0.8	0.3	-0.9	0.6	1.2	-0.4	-1.6	-1.6	-1.0	1.8	0.2	3.7	2.1	0.6	-0.5	-0.5	0.1	-0.8	0.7	1.1	-0.3	-1.4	-1.3	-0.8	2.5	2.4	1.0	0.0	0.0	0.5	4.4	1.7	0.4	-0.1	0.7	0.0	-1.1	-0.2	-0.4	-2.1	-1.0	-1.4	-2.0	-1.4	-0.8
44	DA	-1.0	0.1	-1.0	2.6	1.0	-0.5	-1.7	-1.7	-1.1	1.7	0.0	2.8	2.0	0.5	-0.6	-0.6	0.0	-0.9	2.4	1.0	-0.5	-1.5	-1.4	-1.0	2.1	2.3	0.9	-0.1	-0.1	0.3	4.2	1.5	0.3	-0.2	0.6	-0.1	-1.3	-0.3	-0.5	-2.2	-1.1	-1.6	-2.2	-1.5	-0.9
45	DB	-0.8	0.3	-0.9	0.6	1.2	-0.4	-1.6	-1.6	-1.0	1.8	0.2	2.6	2.1	0.6	-0.5	-0.5	0.1	-0.8	0.7	1.1	-0.3	-1.4	-1.3	-0.8	2.2	2.4	3.6	0.0	0.0	0.5	4.4	1.7	0.4	-0.1	0.7	0.0	-1.1	-0.2	-0.4	-2.1	-1.0	-1.4	-2.0	-1.4	-0.8
46	AC	-0.7	0.4	1.3	0.8	1.3	-0.2	-1.4	-1.4	-0.9	2.0	0.3	2.1	2.3	0.8	-0.3	-0.3	0.2	-0.6	0.8	1.3	-0.2	-1.2	-1.1	-0.7	2.4	2.6	1.2	0.1	0.2	0.6	4.5	1.8	0.6	0.1	0.9	0.2	-1.0	-0.1	-0.2	-1.9	-0.8	-1.3	-1.9	-1.2	-0.6



Table 8. Statistic of our predictions for seven organisms commonly used in molecular research projects.

Species	Number of proteins	Number of interactions (our prediction)	Number of interactions (DIP)
<i>Mus musculus</i> (house mouse)	56924	223151	292
<i>Homo sapiens</i> (Human)	29571	112114	1407
<i>Rattus norvegicus</i> (Norway rat)	24115	71407	109
<i>Caenorhabditis elegans</i> (nematode)	22729	17242	4030
<i>Drosophila melanogaster</i> (fruit fly)	19620	41665	20988
<i>Saccharomyces cerevisiae</i> (baker's yeast)	5877	1850	18225
<i>Escherichia coli</i>	4850	477	7408

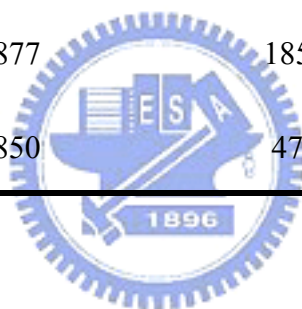


Table 9. The result of 1evtBD to model seven FGF/receptor complexes.

Homologs of 1evtB	Homologs of 1evtD	Binding affinity	SP energy ^a	SP energy (normal)	GE energy ^b	IDE1 ^c	IDE2 ^d
FGF4	FGFR2c	94.3	105.7	0.92	-1.6	35.8	76.3
FGF4	FGFR3c	69.1	104.8	0.91	-1	35.8	73
FGF4	FGFR4	108	103.4	0.9	0.5	35.8	66
FGF4	FGFR1c	102.3	102.5	0.89	0.5	35.8	99.1
FGF4	FGFR2b	14.9	101.2	0.88	-1.9	35.8	69.3
FGF4	FGFR1b	15.6	98	0.85	-0.4	35.8	86
FGF4	FGFR3b	1	97.3	0.84	1.3	35.8	61.9

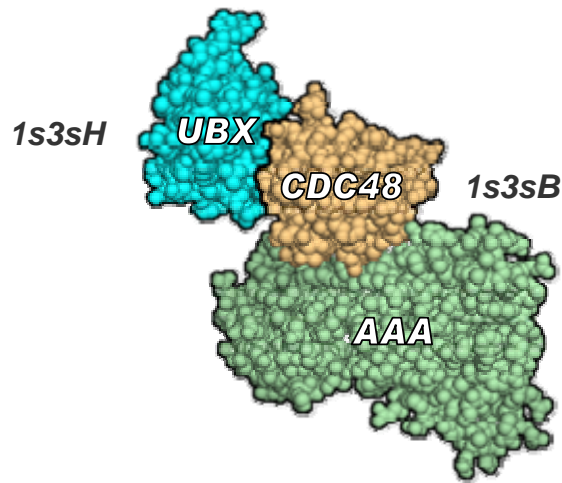
^a SP energy is the abbreviation of the “specific interfacial energy” which is calculated from pairPSSM of 1evtBD.

^b GE energy is the abbreviation of the “general interfacial energy” which is calculated from general empirical matrix.

^c IDE1 means the sequence identity percentage between the candidate protein and 1evt B chain.

^d IDE2 means the sequence identity percentage between the candidate protein and 1evt D chain.

(A)



(B)

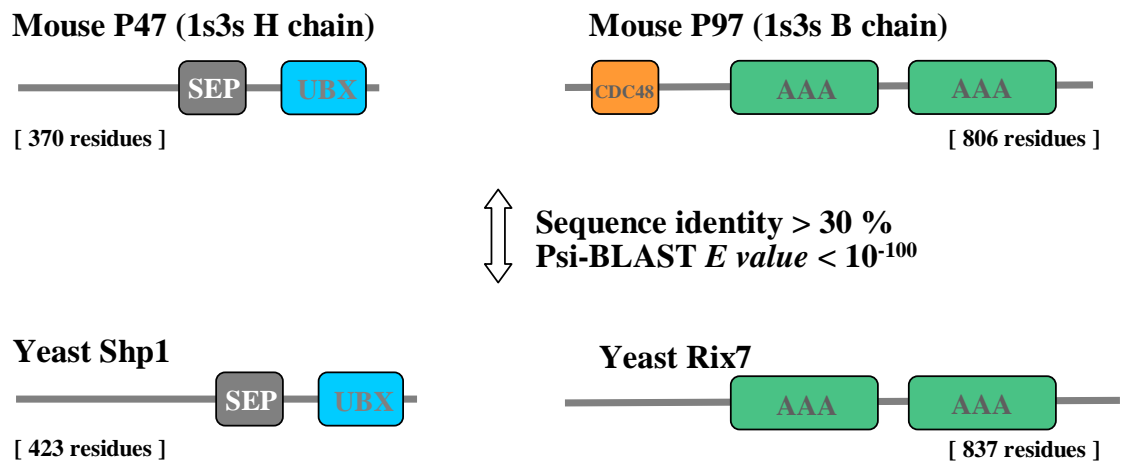
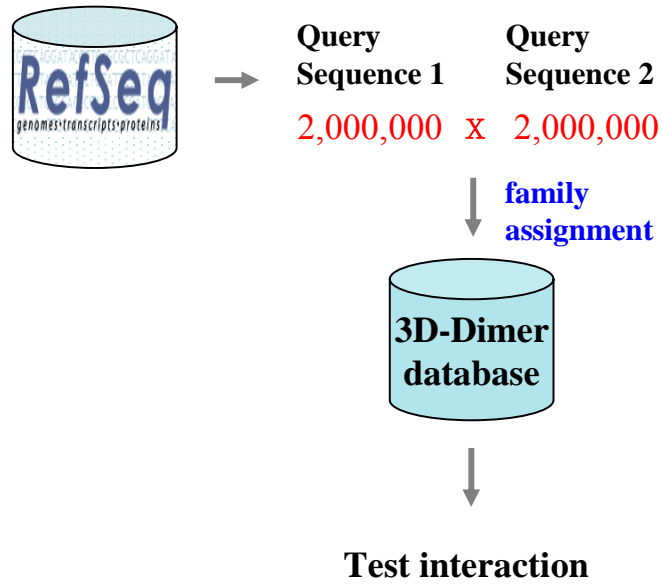


Figure 1. The 3D structure and domain architecture of protein complex P47/P97. (A) The 3D structure of protein complex P47/P97. (B) The domain architecture of two mouse proteins P47&P97 and two yeast proteins Shp1&Rix1. Both homologous protein pairs (P47 to Shp1 and P97 to Rix7) are with sequence identity > 30% and PSI-BLAST *E value* < 10^{-100} . The color boxes are represented as functional domains.

(A)



(B)

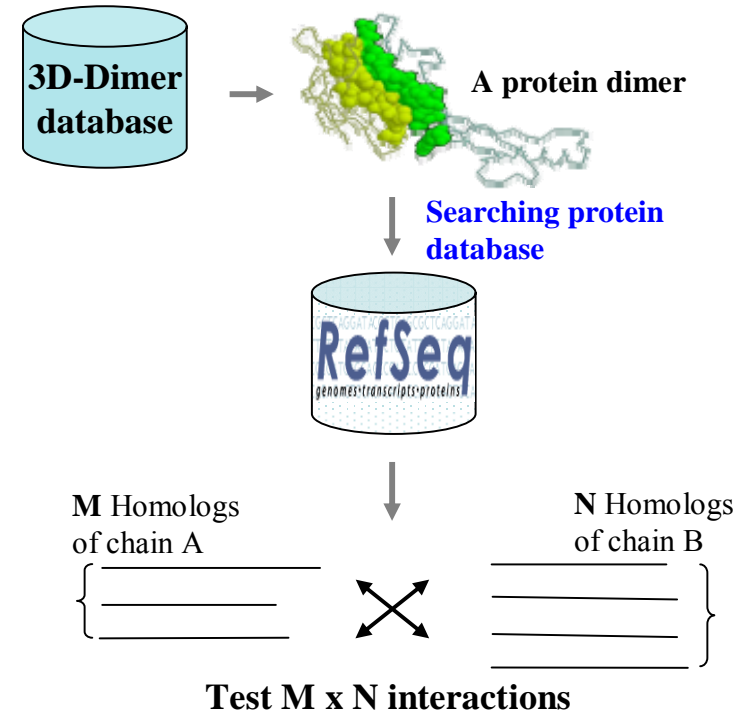


Figure 2. The comparison of our and previous methods. (A) The previous method takes family assignment for 2 query proteins and fits the two proteins on the complex of known structure. (B) Our method modified by the concept of interologs. We use a complex of known structure to search protein database and test any possible homologous interacting protein pairs on the complex.

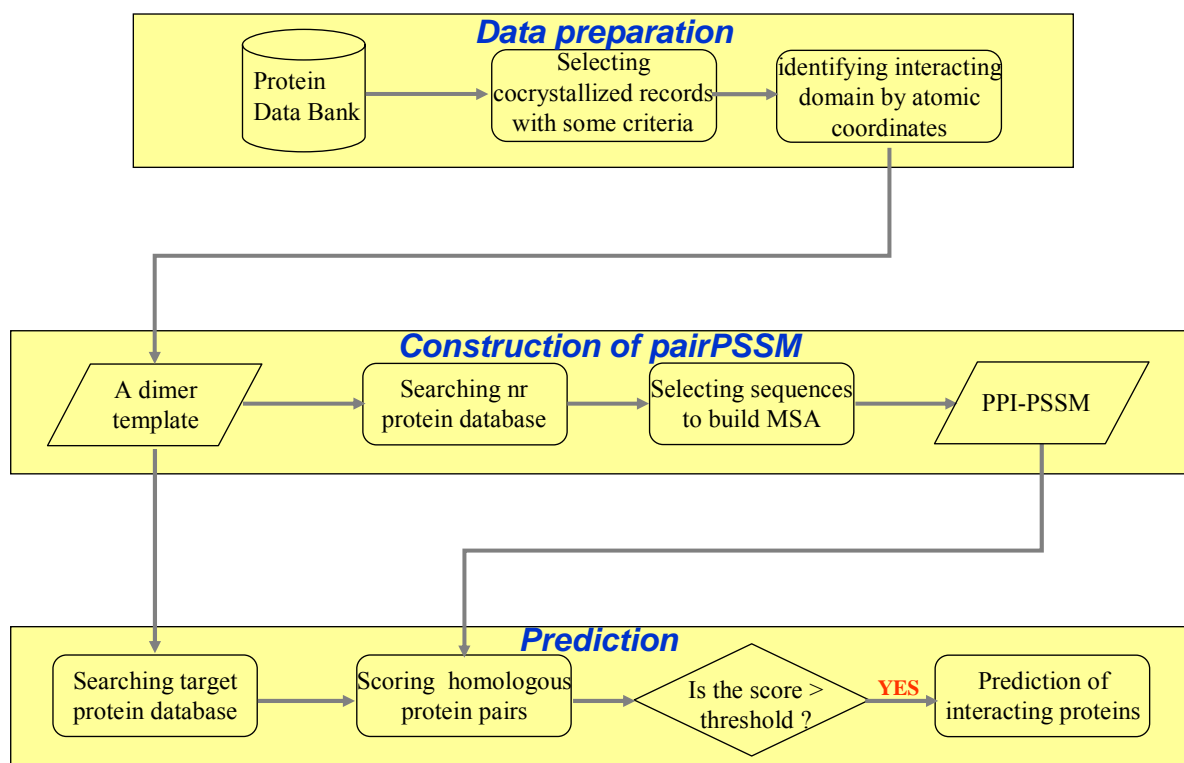


Figure 3. The flow chart of our method. In this study, we first collect dimers of known structure from Protein Databank and identify interacting domains. For each 3D-dimer, we estimate the probabilities with which residue pairs occur at various contact positions and construct a pairPSSM to assess the fit of any possible interacting protein pairs. And then we use these dimers as queries to search target protein database and predict many candidates of protein-protein interaction.

(A)

	AV	LIM	GST	PFYW	C	H	RK	DE	NQ
AV	0.62	0.46	0.35	0.83	0.35	0.40	0.75	0.38	0.38
LIM	0.46	0.31	0.17	0.69	0.09	0.09	0.48	0.14	0.16
GST	0.35	0.17	0.29	0.64	0.16	0.17	0.46	0.19	0.07
PFYW	0.83	0.69	0.64	0.95	0.63	0.77	0.80	0.66	0.63
C	0.35	0.09	0.16	0.63	0.00	0.00	0.70	0.10	0.25
H	0.40	0.09	0.17	0.77	0.00	0.00	0.55	0.05	0.15
RK	0.75	0.48	0.46	0.80	0.70	0.55	0.79	0.43	0.45
DE	0.38	0.14	0.19	0.66	0.10	0.05	0.43	0.38	0.08
NQ	0.38	0.16	0.07	0.63	0.25	0.15	0.45	0.08	0.33

(B)



	AG	LIMV	STP	FYW	C	HR	K	DE	NQ
AG	0.35	0.30	0.11	0.14	0.10	0.07	0.10	0.08	0.08
LIMV	0.30	0.35	0.21	0.22	0.13	0.20	0.13	0.17	0.17
STP	0.11	0.21	0.26	0.13	0.16	0.21	0.14	0.19	0.12
FYW	0.14	0.22	0.13	0.33	0.12	0.23	0.00	0.19	0.24
C	0.10	0.13	0.16	0.12	0.00	0.15	0.00	0.10	0.25
HR	0.07	0.20	0.21	0.23	0.15	0.37	0.05	0.19	0.11
K	0.10	0.13	0.14	0.00	0.00	0.05	0.00	0.05	0.00
DE	0.08	0.17	0.19	0.19	0.10	0.19	0.05	0.38	0.08
NQ	0.08	0.17	0.12	0.24	0.25	0.11	0.00	0.08	0.33

Figure 4. The standard deviations of contact residue potentials in the clusters of amino acid. The deviation > 0.5 is colored by dark gray and the deviation between 0.3 and 0.5 is colored by gray. (A) The amino acid classification is defined by Saha et al. (B) The classification is slightly modified by us.

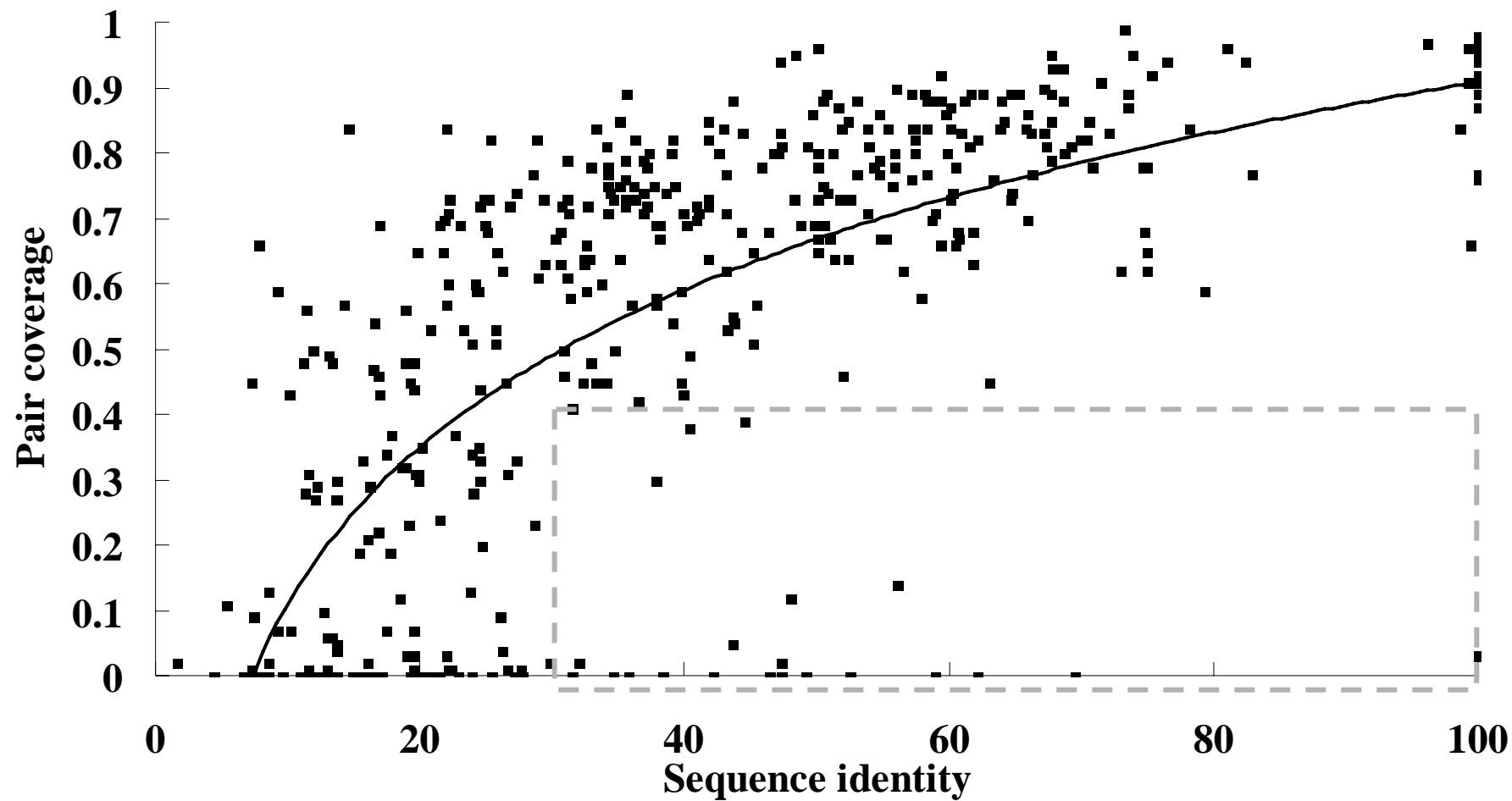


Figure 5. The relationship between sequence identity and pair coverage of 459 pairs of related hetero dimers. The dots in gray box are the exceptions of the pairs of dimer with $> 30\%$ sequence identity but pair coverage < 0.4 .

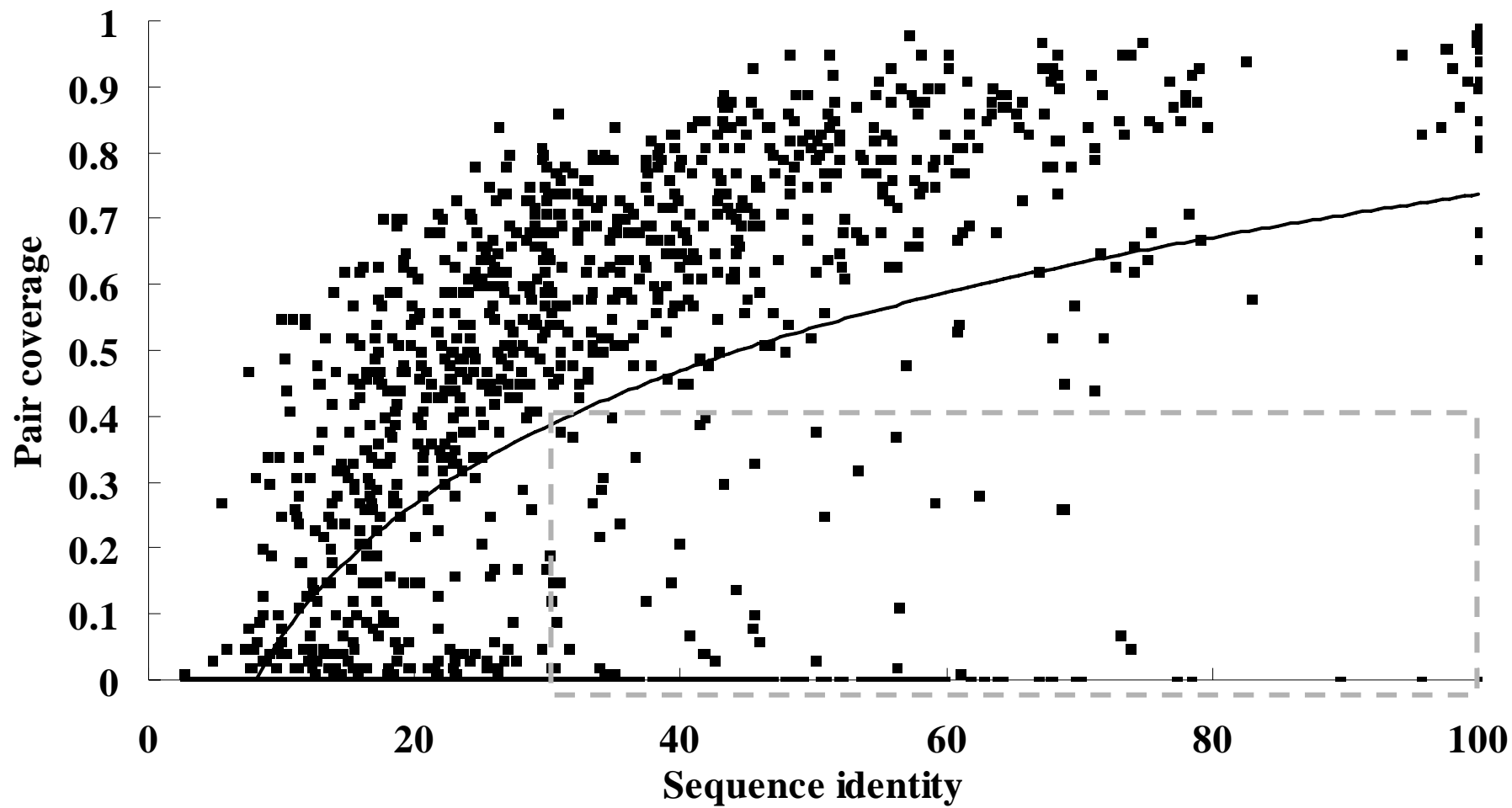


Figure 6. The relationship between sequence identity and pair coverage of 1412 pairs of related homo dimers. The dots in gray box are the exceptions of the pairs of dimer with >30% sequence identity but pair coverage < 0.4.

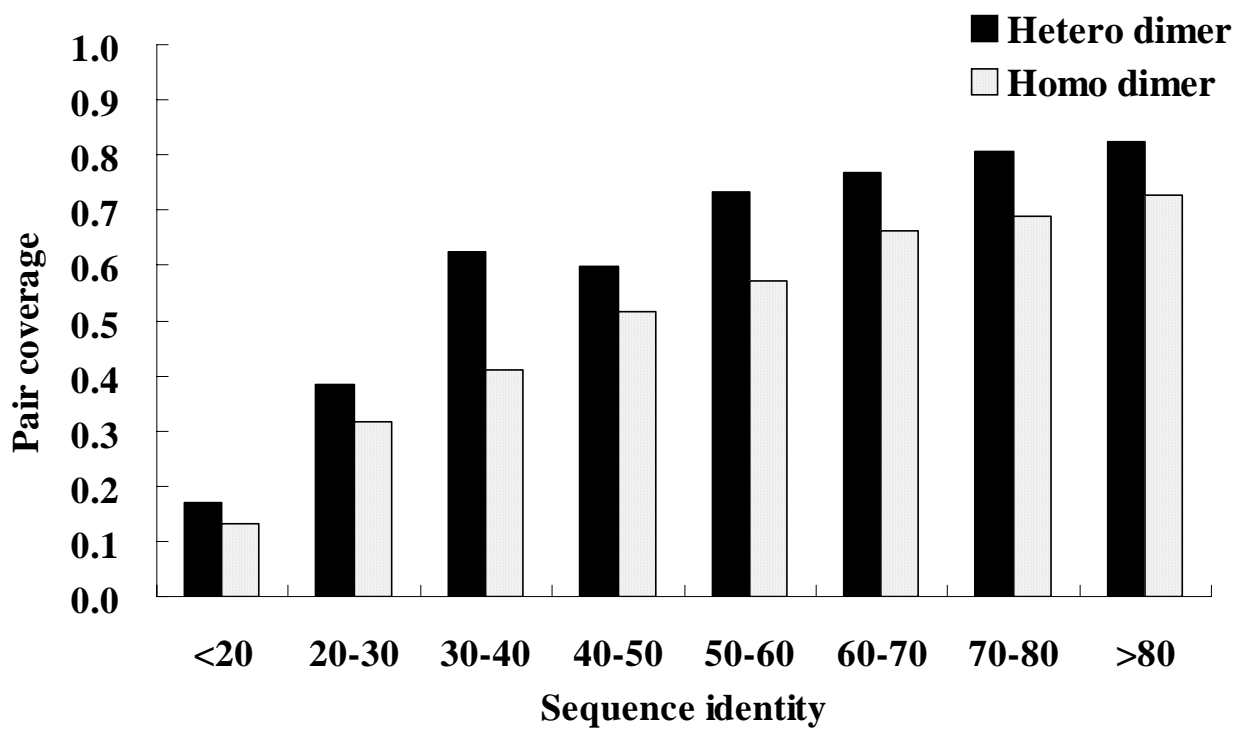
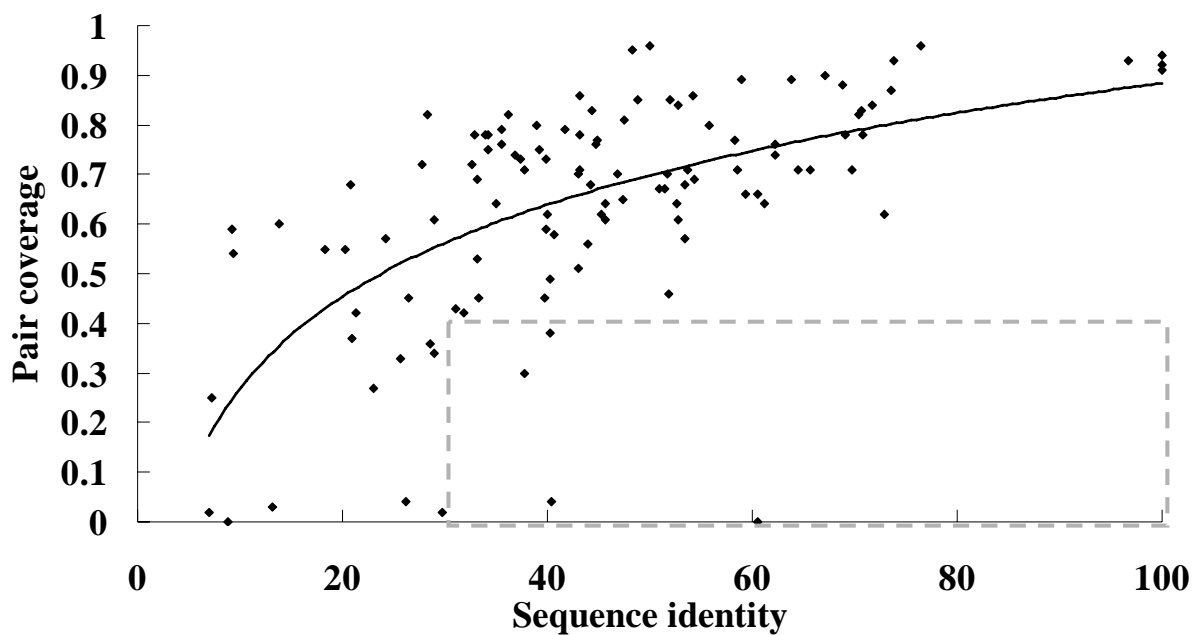


Figure 7. The average pair coverage in different sequence identity interval. The black bar is for heterodimers and the gray bar is for homodimers.

(A) Two-chain heterodimers



(B) Two-chain homodimers

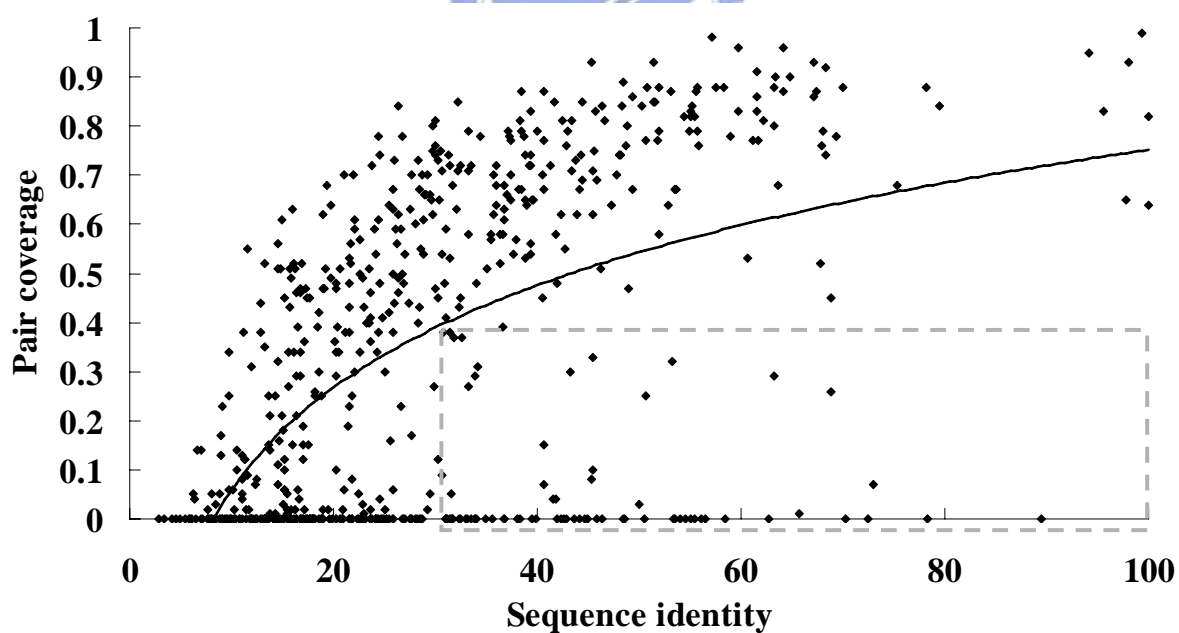


Figure 8. The relationship between sequence identity and pair coverage of the two-chain dimers. (A) 114 pairs of related two-chain heterodimers. (B) 616 pairs of related two-chain homodimers. The dots in gray box are the exceptions of the pairs of dimer with $>30\%$ sequence identity but pair coverage < 0.4 .

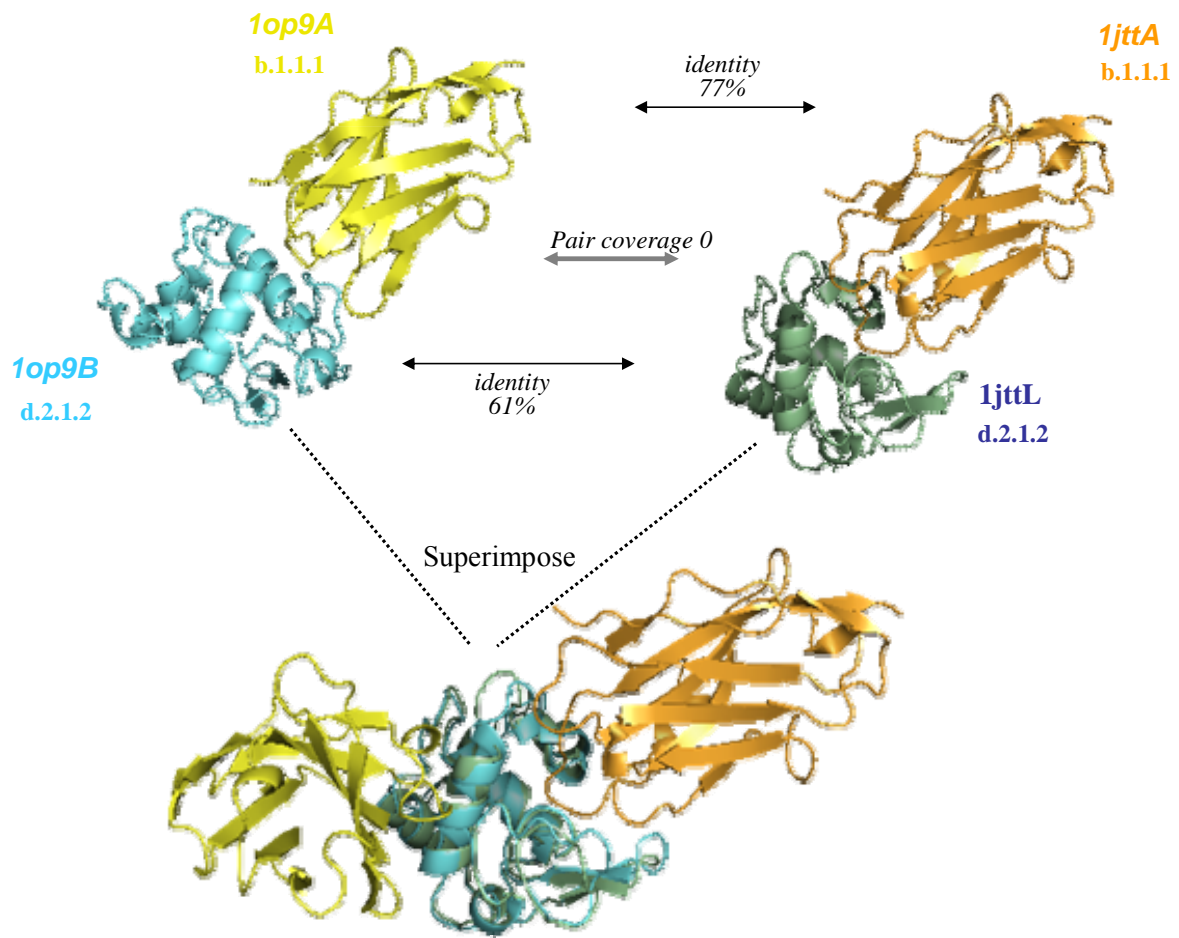
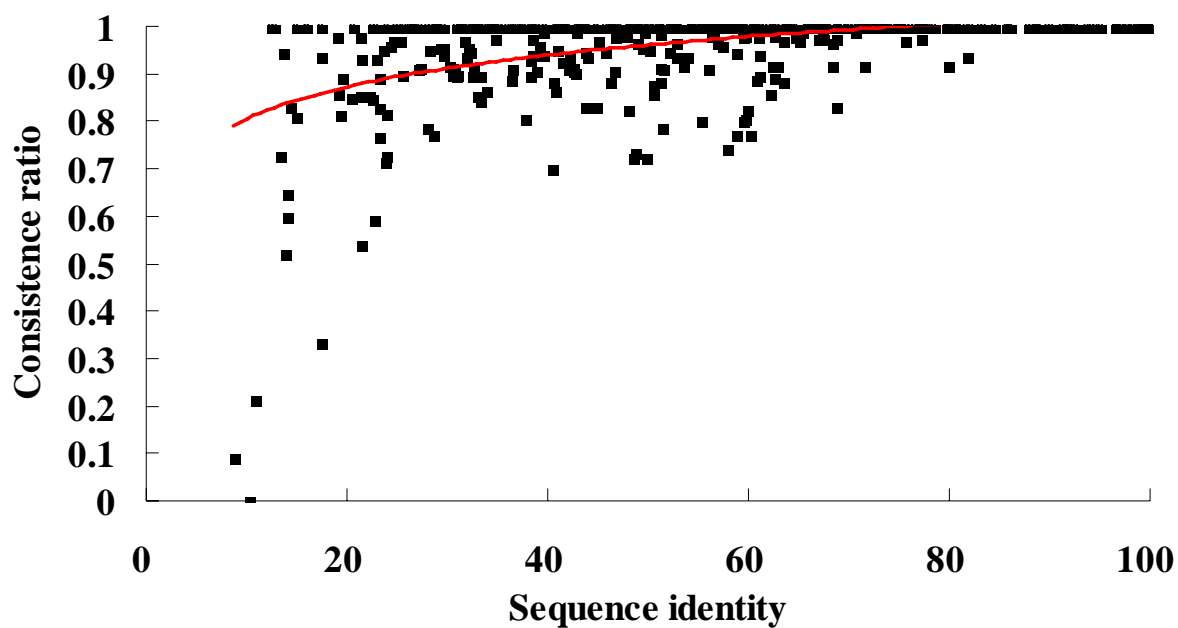


Figure 9. The interactive types of two hydrolyase-antibody complexes 1op9AB and 1jttAL. We superimpose the 1op9B and 1jttL and discover the binding site of the two proteins at two different sites (bottom).

(A)



(B)

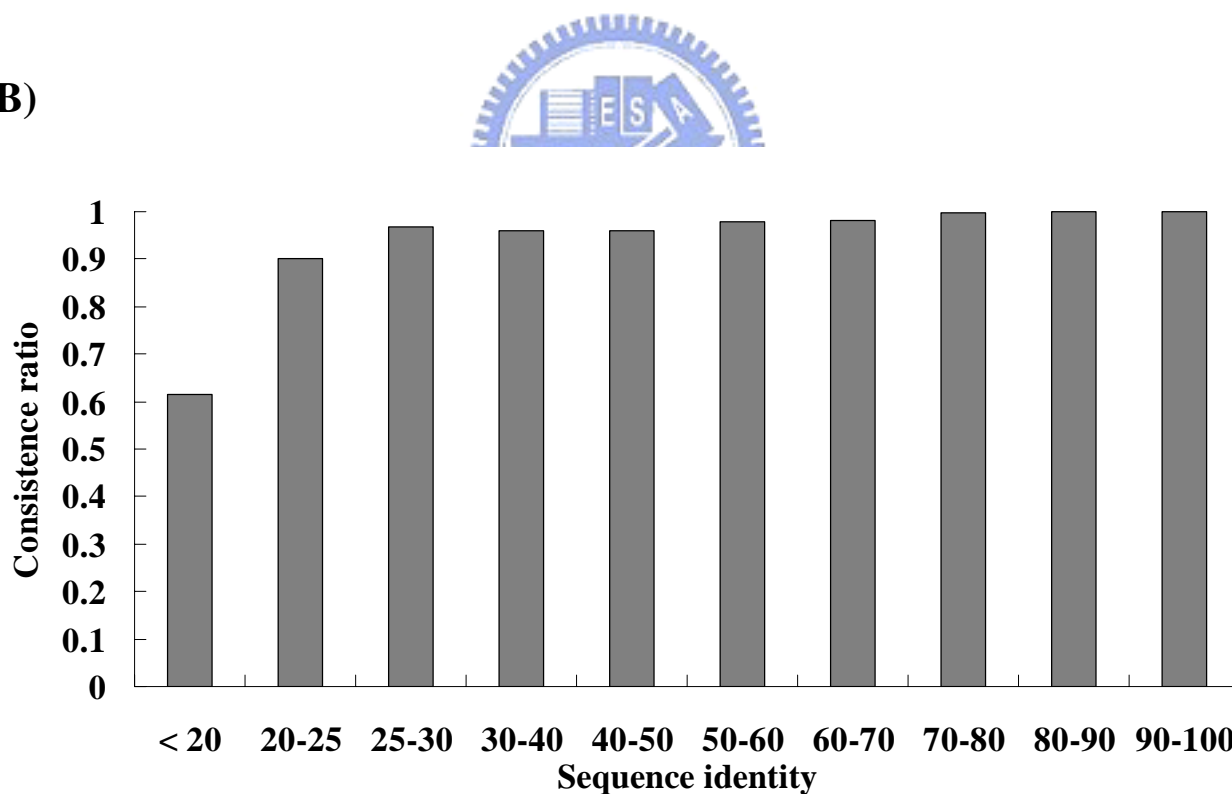
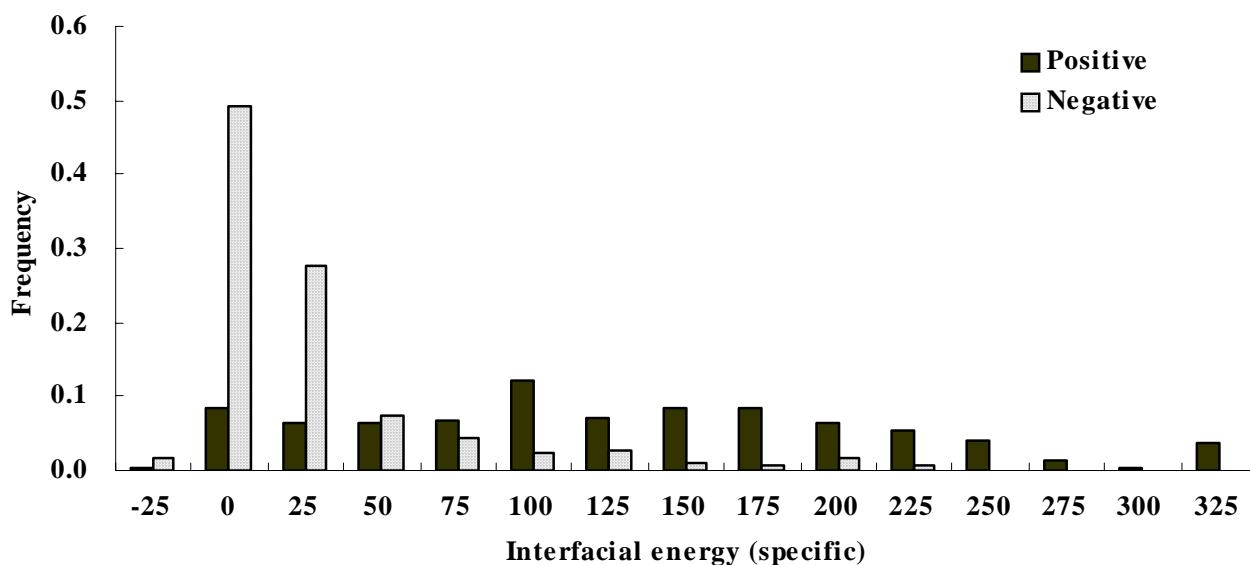


Figure 10. Sequence identity threshold of aligning contact residues. (A) The relationship between the consistence ratio and sequence identity. (B) The mean of consistence ratios in different sequence identity.

(A)



(B)

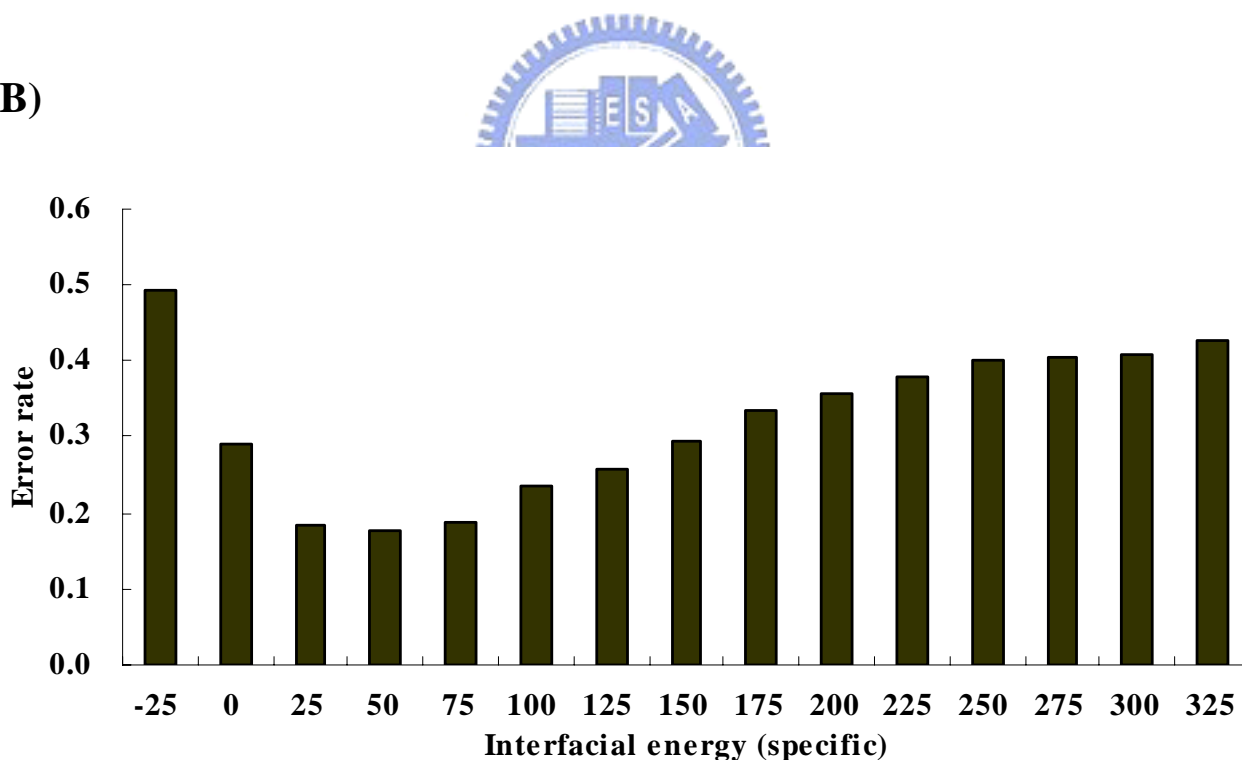
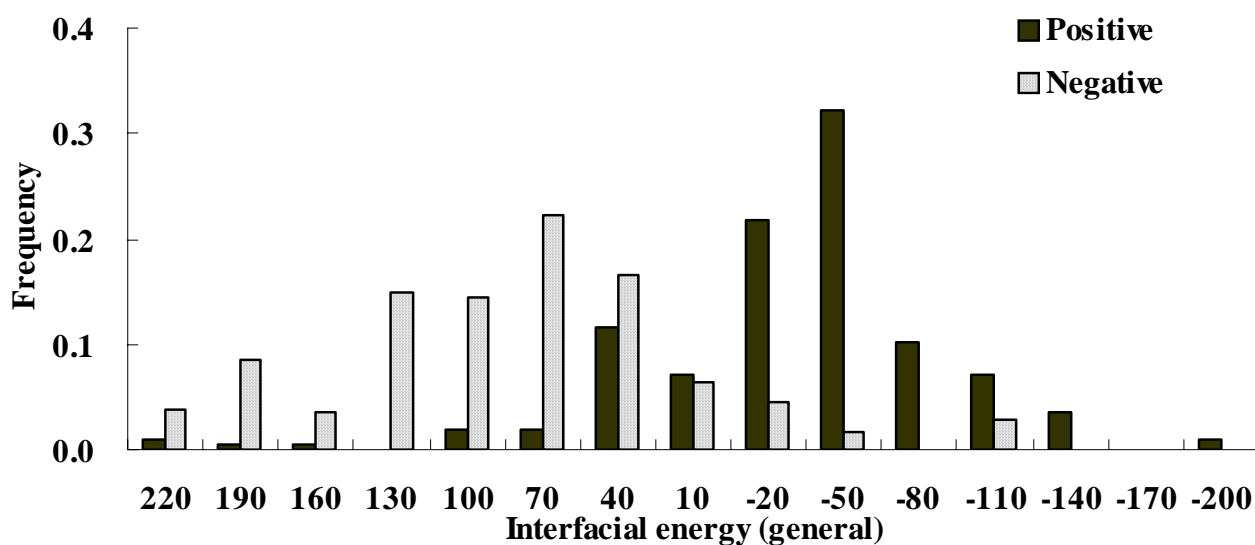


Figure 11. Determining the threshold of specific interfacial energy on distinguishing the true protein complex and unreasonable protein pairs. The specific interfacial energy is calculated from pair PSSM. (A) The frequency of positives and negatives in different interfacial energy intervals. (B) The error rate of prediction at different thresholds. A threshold of 50 is consequently set from this histogram.

(A)



(B)

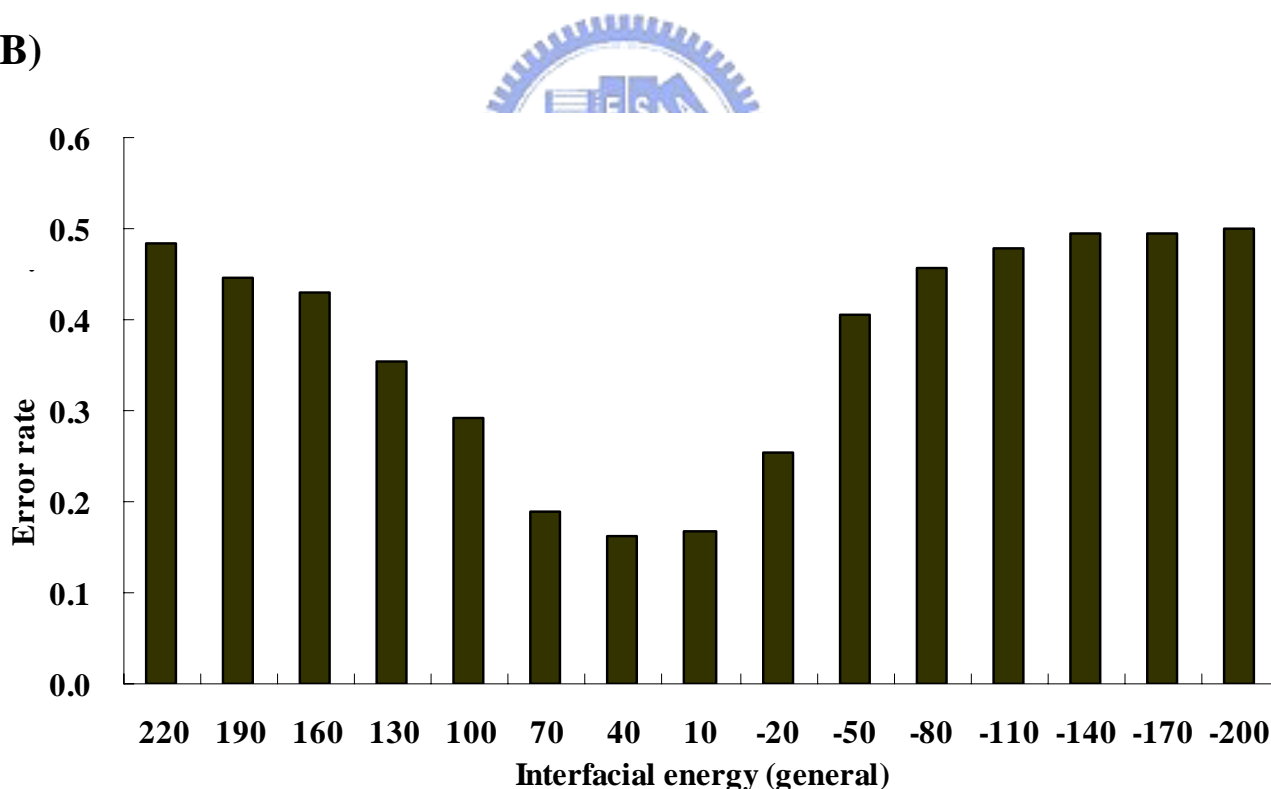


Figure 12. Determining the threshold of general interfacial energy on distinguishing the true protein complex and unreasonable protein pairs. The general interfacial energy is calculated from general empirical matrix. (A) The frequency of positives and negatives in different interfacial energy intervals. (B) The error rate of prediction at different thresholds. A threshold of 10 is consequently set from this histogram.

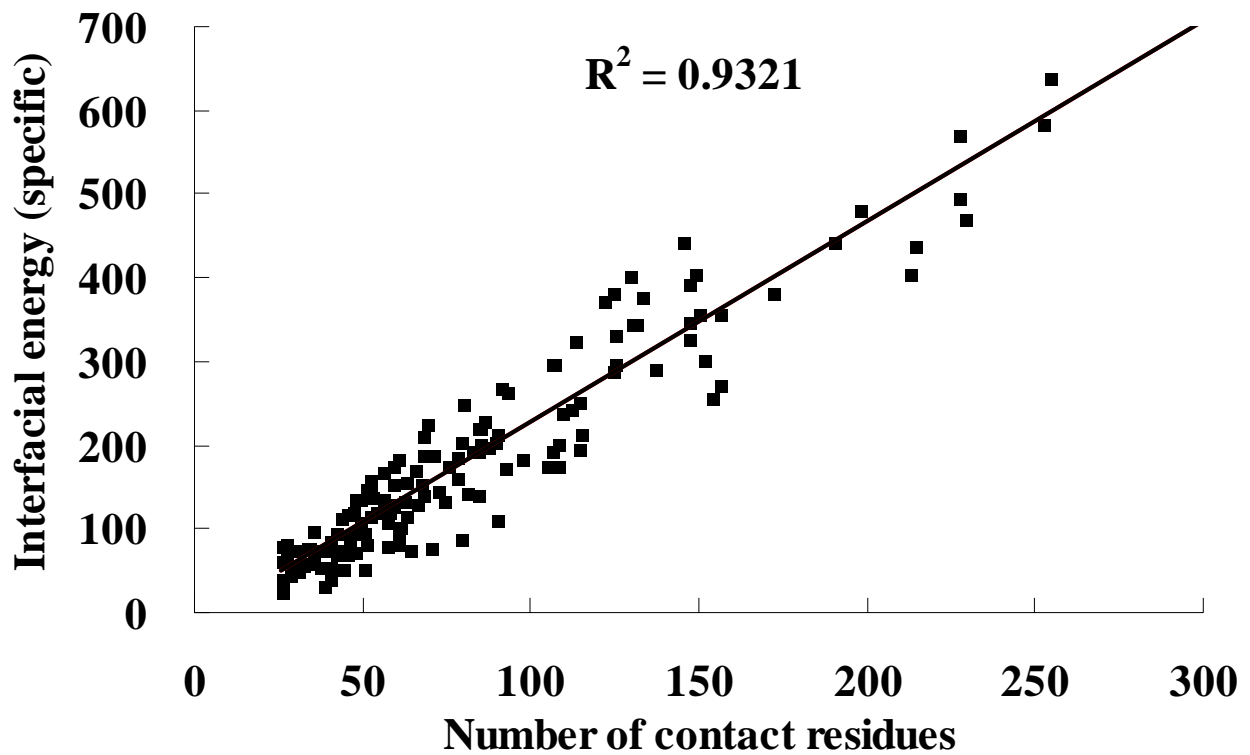


Figure 13. The relationship between number of contact residues in 3D-dimers and its specific interfacial energies which are calculated from pairPSSM. The correlation coefficient is 0.9321.

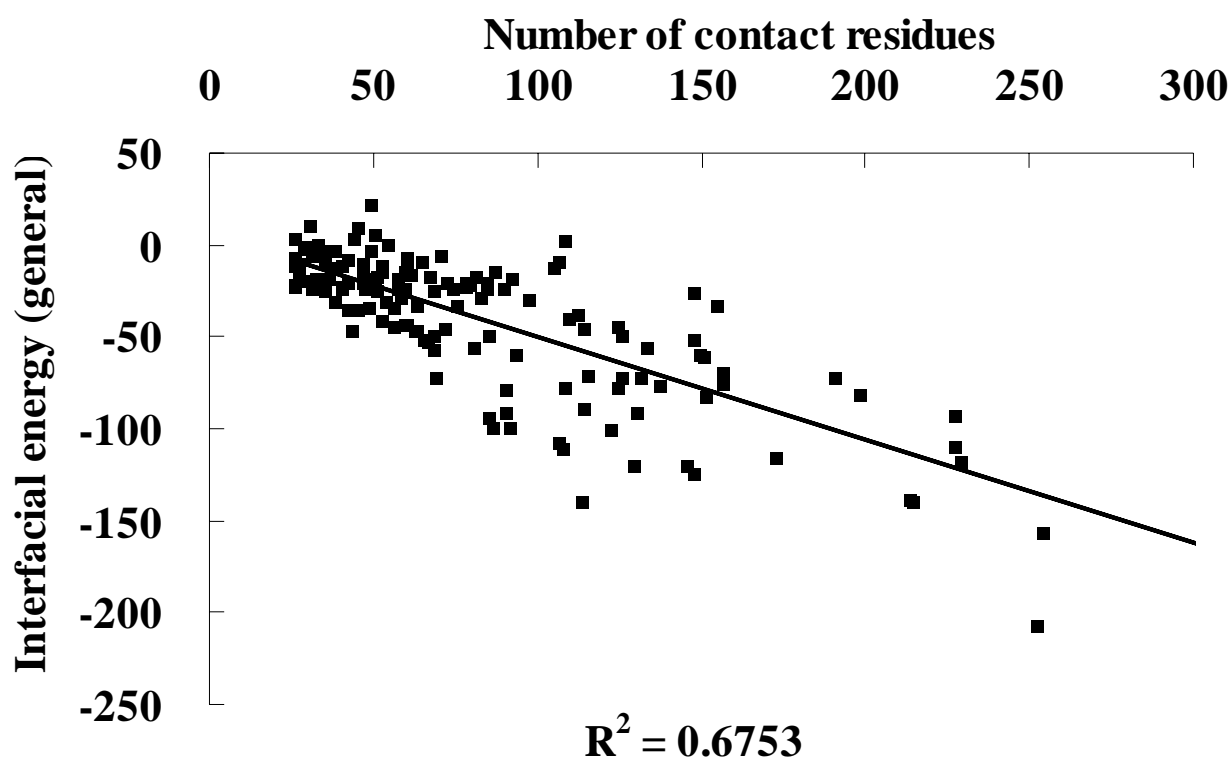
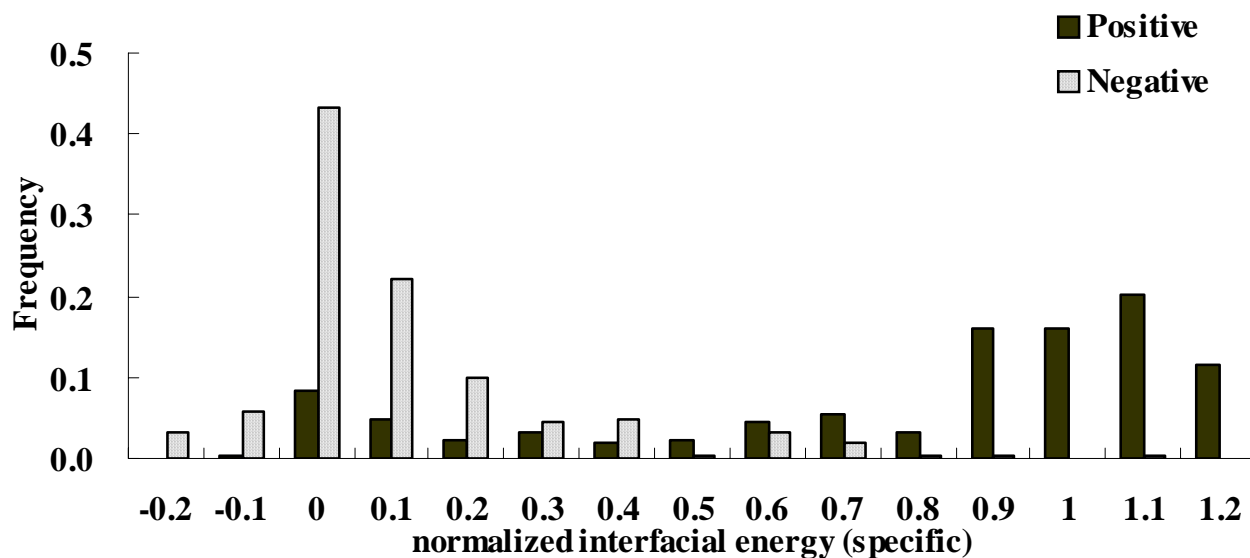


Figure 14. The relationship between number of contact residues in 3D-dimer and its general interfacial energies with are calculated from general empirical matrix. The correlation coefficient is 0.6753.

(A)



(B)

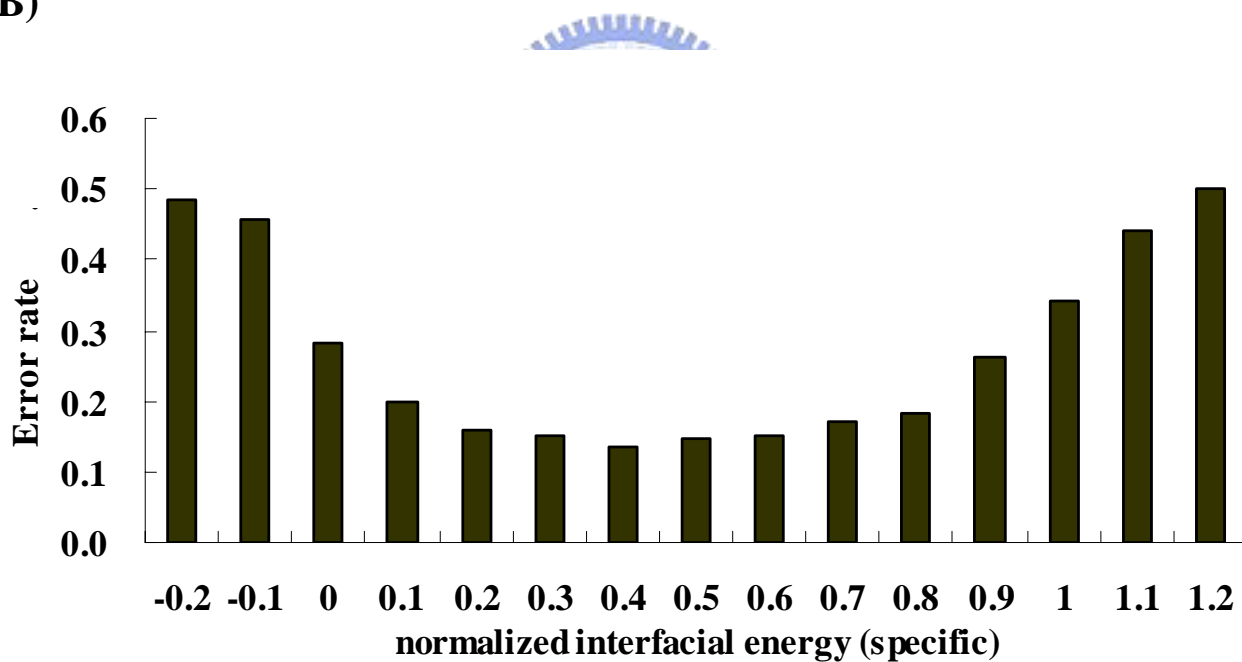


Figure 15. Determining the threshold of normalized specific interfacial energy on distinguishing the true protein complex and unreasonable protein pairs. The method to calculate normalized specific interfacial energy describes in text. (A) The frequency of positives and negatives in different interfacial energy intervals. (B) The error rate of prediction at different thresholds. A threshold of 0.4 is consequently set from this histogram.

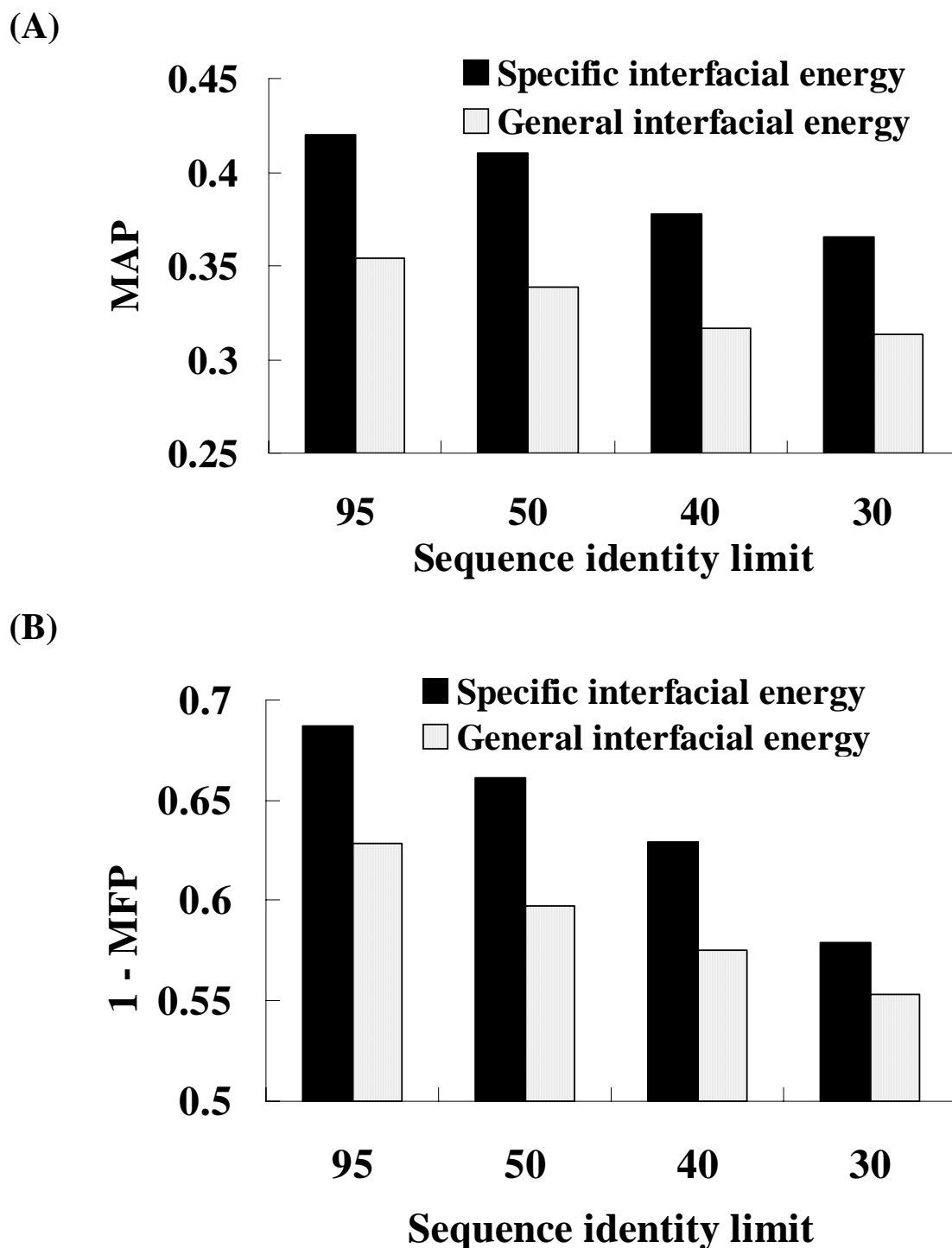


Figure 16. The mean average positions and mean false positive rate of 182 queries. The unannotated candidates are considered as negatives. Sequence identity limit means that if one protein of candidate with sequence identity > sequence identity limit, the candidate is removed. (A) Result of MAP in sequence identity limit with 95%, 50%, 40% and 30%, respectively. (B) Result of MFP in sequence identity limit with 95%, 50%, 40% and 30%, respectively.

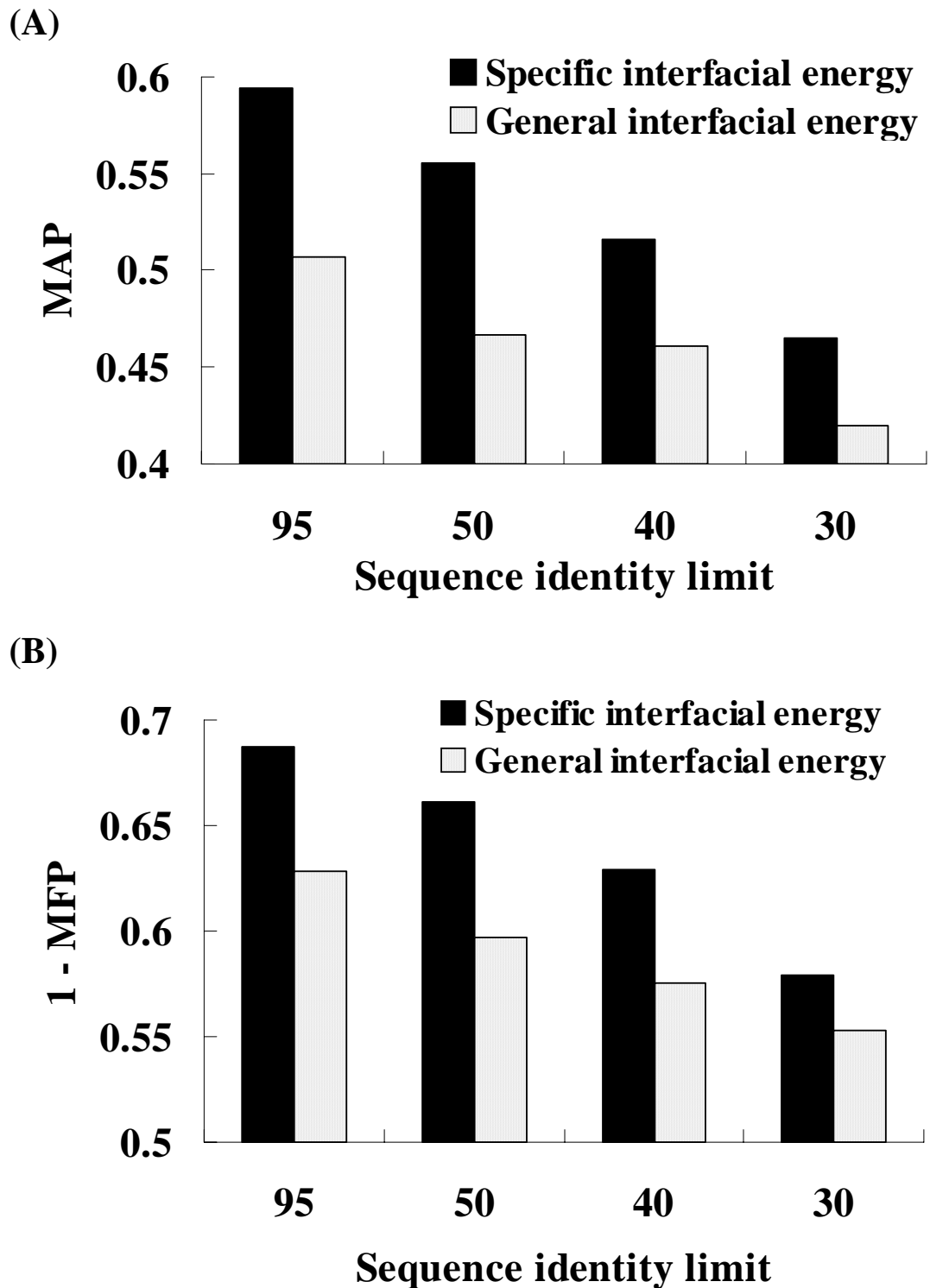


Figure 17. The mean average positions and mean false positive rate of 101 queries. The unannotated candidates are removed. Sequence identity limit means that if one protein of candidate one with sequence identity $>$ sequence identity limit, the candidate is removed. (A) Result of MAP in sequence identity limit with 95%, 50%, 40% and 30%, respectively. (B) Result of MFP in sequence identity limit with 95%, 50%, 40% and 30%, respectively.

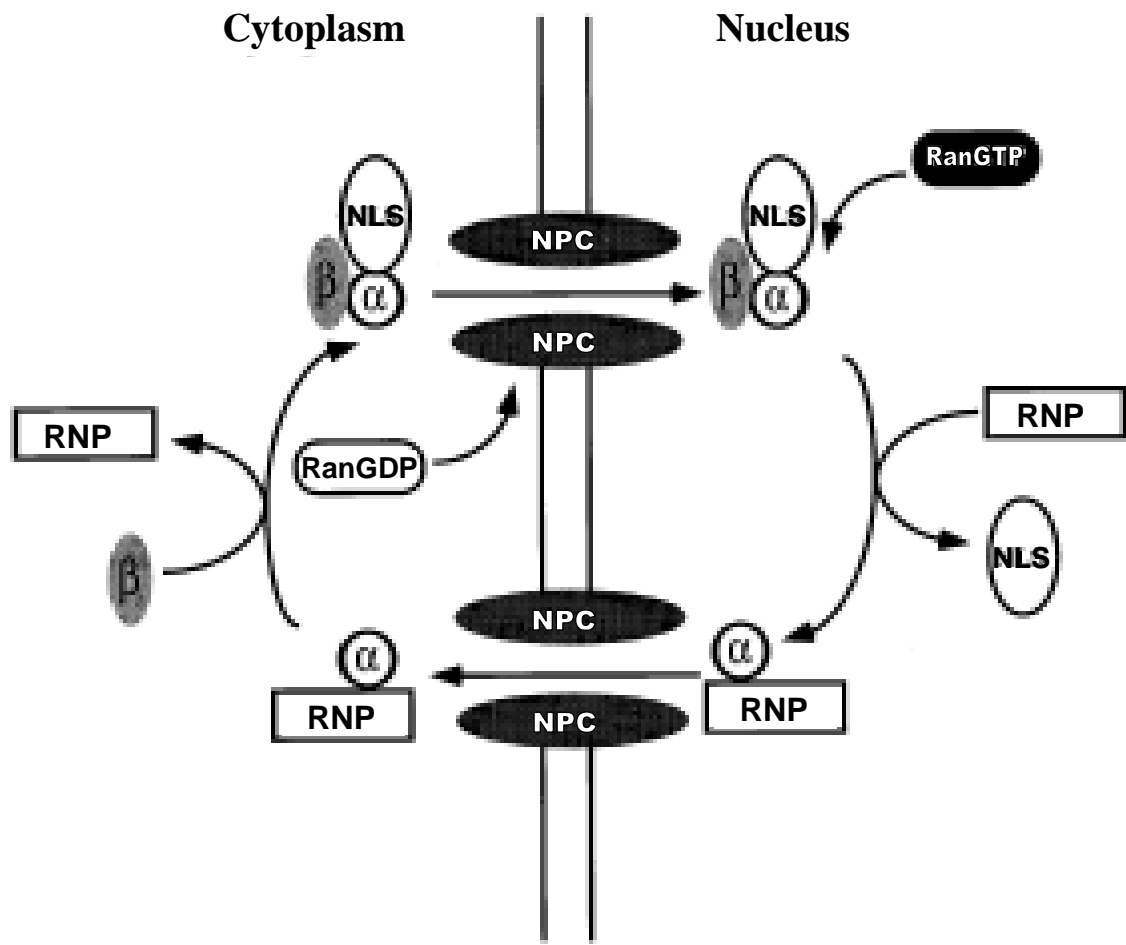
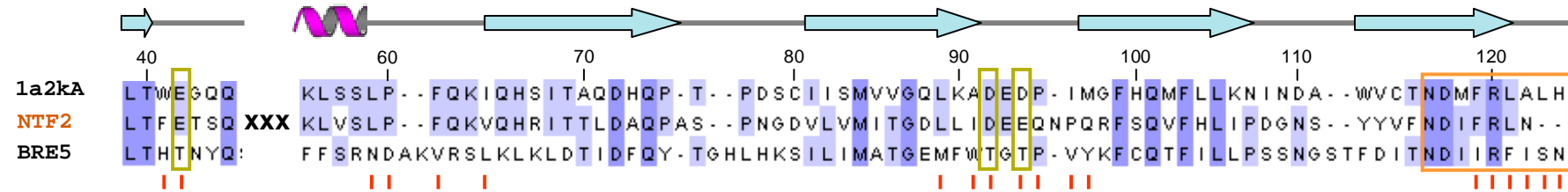


Figure 18. Model for cycling transport factors proposed by Koepp and Silver (51). The mechanism of nuclear transport factors cycle sees text for detail.

(A)



(B)

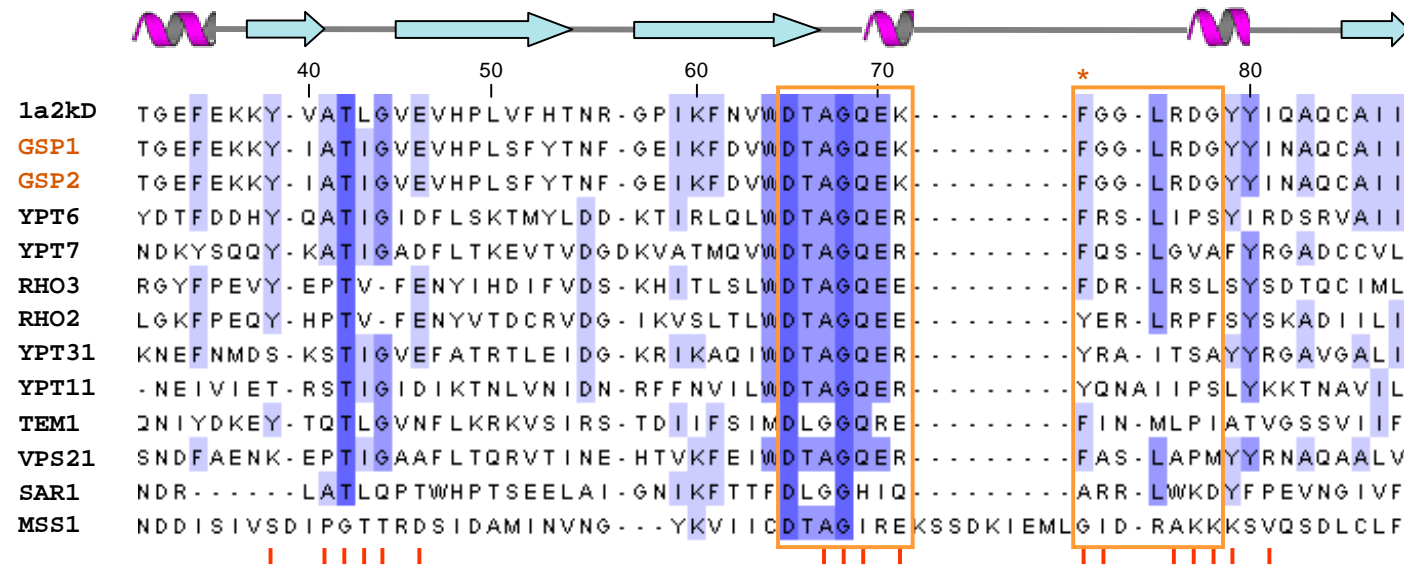


Figure 19. The multiple sequence alignment result of the 14 candidates to their corresponding template proteins of 1a2kAD. (A) Alignment result of 1a2k A chain. Three important negative residues mark in the yellow box. The C terminal hydrophobic peptide is also an important interactive site (orange box). The red bars in the bottom are the contact positions in 1a2k A chain. (B) Alignment result of 1a2k D chain. The switch II loops mark in orange box. The important aromatic residue Phe72 is mark by orange star. The red bars in the bottom are the contact positions in 1a2k D chain.

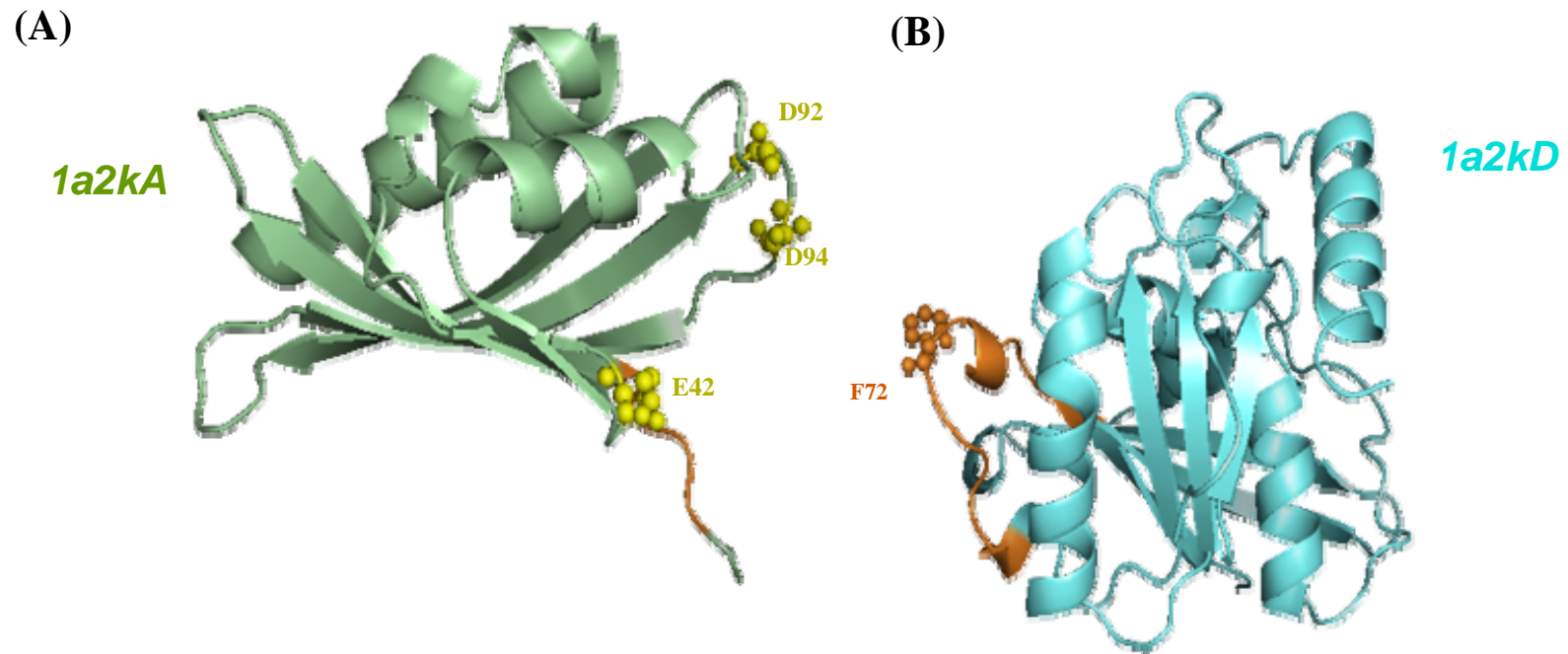


Figure 20. 3D-structure of 1a2kAD. (A) A chain of 1a2k. The three important negatively charged residues, Glu42, Asp92 and Asp94, are colored by yellow. The C terminal peptide is colored by orange. (B) D chain of 1a2k. The switch II loops is colored by orange.

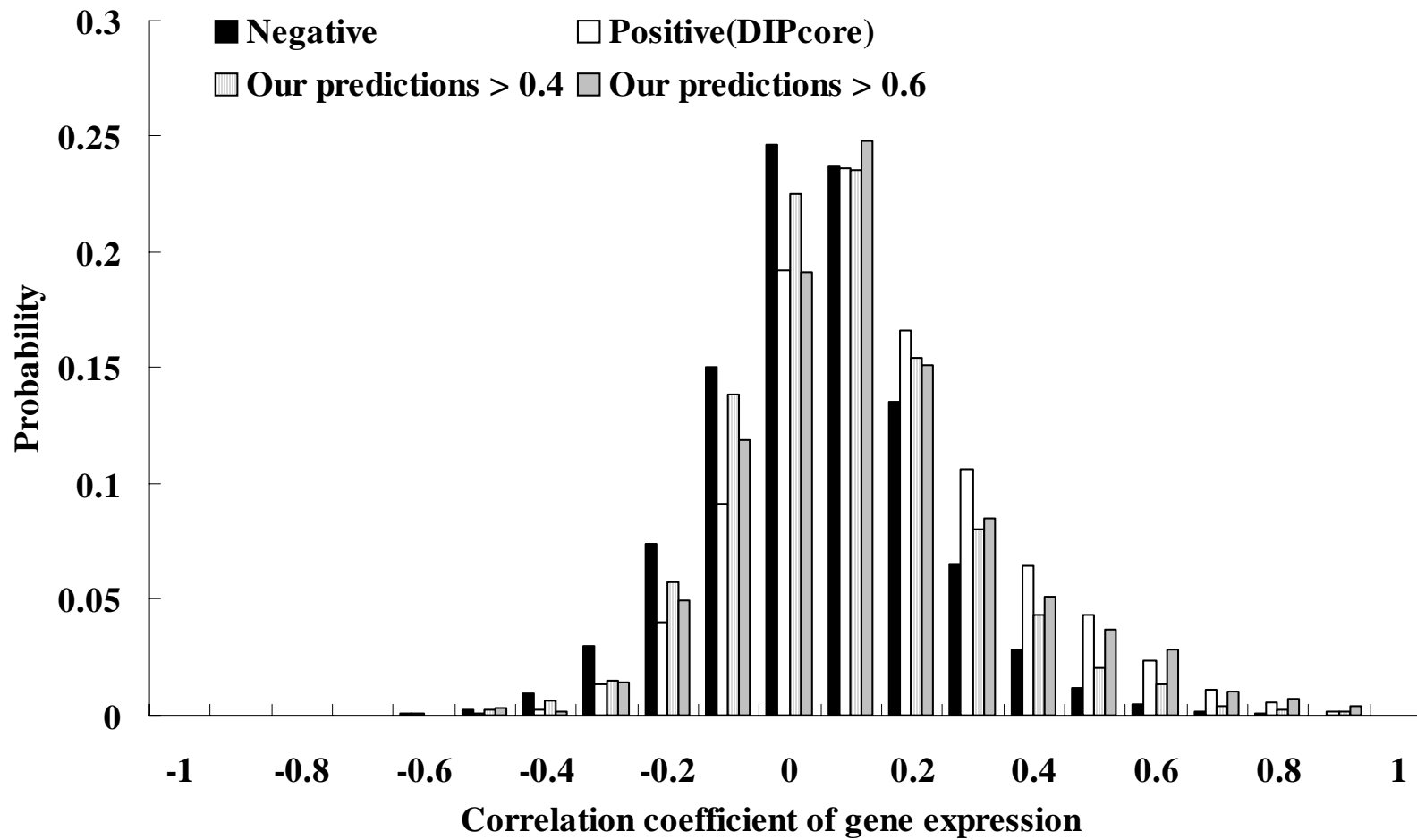


Figure 21. Distributions of the correlation coefficients of gene expression profiles for four interacting protein sets: our predicted protein pairs with thresholds 0.4 (band) and 0.6 (gray), the DIP core set (white), and the non-interacting protein pairs (black). The correlations of our predicted protein pairs are much higher than the one of non-interacting protein pairs.

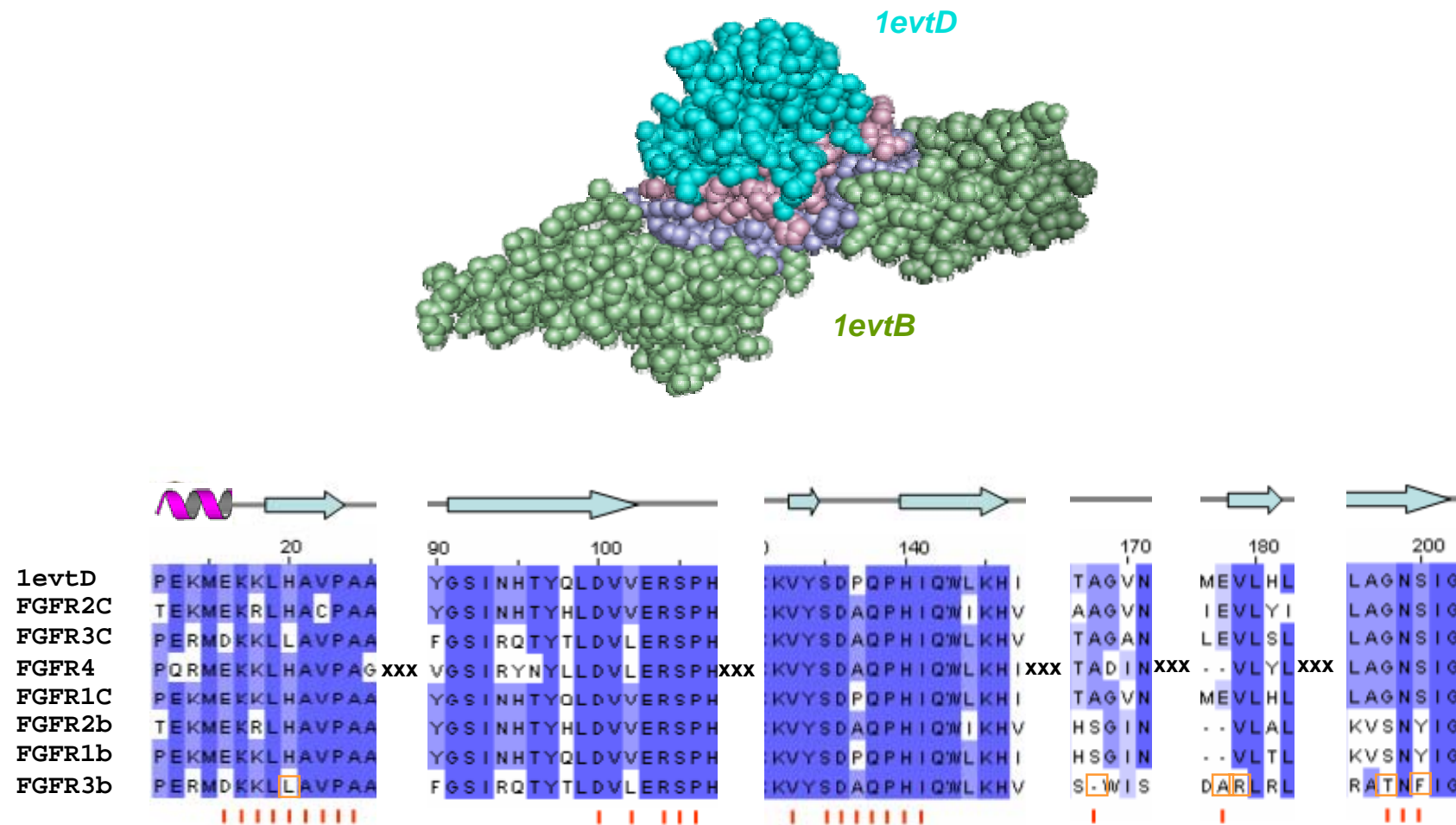


Figure 22. The 3D-structure of 1evtBD and multiple sequence alignment of seven homologous FGF receptors. (A) The 3D-structure of 1evtBD. (B) The multiple sequence alignment of seven homologous FGF receptors to chain D of 1evt. The red bars in the bottom are the contact positions in 1evt D chain.

REFERENCES

1. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399-403.
2. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Science of the USA*, **98**, 4569-4574.
3. Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E. *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727-1736.
4. Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T. *et al.* (2004) A map of the interactome network of the metazoan *C. elegans*. *Science*, **303**, 540-543.
5. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83-86.
6. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F. and Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449-453.
7. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Science of the USA*, **96**, 4285-4288.
8. Aloy, P. and Russell, R.B. (2002) Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Science of the USA*, **99**, 5896-5901.
9. Lu, L., Arakaki, A.K., Lu, H. and Skolnick, J. (2003) Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. *Genome Research*, **13**, 1146-1154.
10. Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.D., Bertin, N., Chung, S., Vidal, M. and Gerstein, M. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Research*, **14**, 1107-1118.
11. Wojcik, J. and Schachter, V. (2001) Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, **17**, S296-S305.
12. Tamames, J., Casari, G., Ouzounis, C. and Valencia, A. (1997) Conserved clusters of functionally related genes in two bacterial genomes. *Journal of Molecular Evolution*, **44**, 66-73.
13. Pazos, F. and Valencia, A. (2002) In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, **47**, 219-227.
14. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. and Eisenberg, D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, **32**, D449-D451.
15. Alfarano, C., Andrade, C.E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobeckho, B., Boutilier, K., Burgess, E. *et al.* (2005) The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Research*, **33**, D418-D424.
16. Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V. *et al.* (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Research*, **32**, D41-D44.
17. von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A. and Bork, P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, **33**,

D433-D437.

18. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, **33**, D501-D504.
19. Pawson, T. and Nash, P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**, 445-452.
20. Sprinzak, E. and Margalit, H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology*, **311**, 681-692.
21. Deng, M., Mehta, S., Sun, F. and Chen, T. (2002) Inferring domain-domain interactions from protein-protein interactions. *Genome Research*, **12**, 1540-1548.
22. Lu, L., Lu, H. and Skolnick, J. (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, **49**, 350-364.
23. Hetzer, M., Meyer, H.H., Walther, T.C., Bilbao-Cortes, D., Warren, G. and Mattaj, I.W. (2001) Distinct AAA-ATPase p97 complexes function in discrete steps of nuclear assembly. *Nature Cell Biololgy*, **3**, 1086-1091.
24. Uchiyama, K., Jokitalo, E., Kano, F., Murata, M., Zhang, X., Canas, B., Newman, R., Rabouille, C., Pappin, D., Freemont, P. *et al.* (2002) VCIP135, a novel essential factor for p97/p47-mediated membrane fusion, is required for Golgi and ER assembly in vivo. *Journal of Cell Biology*, **159**, 855-866.
25. Gadad, O., Strauss, D., Braspenning, J., Hoepfner, D., Petfalski, E., Philippsen, P., Tollervey, D. and Hurt, E. (2001) A nuclear AAA-type ATPase (Rix7p) is required for biogenesis and nuclear export of 60S ribosomal subunits. *The EMBO Journal*, **20**, 3695-3704.
26. Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N. and Vidal, M. (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*, **287**, 116-122.
27. Deshpande, N., Address, K.J., Bluhm, W.F., Merino-Ott, J.C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Research*, **33**, D233-D237.
28. Keskin, O., Ma, B. and Nussinov, R. (2005) Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. *Journal of Molecular Biology*, **345**, 1281-1294.
29. Bogan, A.A. and Thorn, K.S. (1998) Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology*, **280**, 1-9.
30. Ptitsyn, O.B. (1998) Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? *Journal of Molecular Biology*, **278**, 655-666.
31. Elcock, A.H. and McCammon, J.A. (2001) Identification of protein oligomerization states by analysis of interface conservation. *Proceedings of the National Academy of Science of the USA*, **98**, 2990-2994.
32. Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, **32**, D226-229.
33. Henrick, K. and Thornton, J.M. (1998) PQS: a protein quaternary structure file server. *Trends in Biochemical Sciences*, **23**, 358-361.
34. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389-3402.
35. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623-627.

36. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180-183.
37. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141-147.
38. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, **9**, 3273-3297.
39. Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109-126.
40. Hirschman, J.E., Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hong, E.L., Livstone, M.S., Nash, R. *et al.* (2006) Genome Snapshot: a new resource at the *Saccharomyces Genome Database* (SGD) presenting an overview of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Research*, **34**, D442-D445.
41. Kumar, A., Agarwal, S., Heyman, J.A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y. *et al.* (2002) Subcellular localization of the yeast proteome. *Genes and Develop*, **16**, 707-719.
42. Ge, H., Liu, Z., Church, G.M. and Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genetics*, **29**, 482-486.
43. Abagyan, R.A. and Batalov, S. (1997) Do aligned sequences share the same fold? *Journal of Molecular Biology*, **273**.
44. Kabsch, W. and Sander, C. (1984) On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proceedings of the National Academy of Science of the USA*, **81**, 1075-1078.
45. Tatusov, R.L., Altschul, S.F. and Koonin, E.V. (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proceedings of the National Academy of Science of the USA*, **91**, 12091-12095.
46. Henikoff, J.G. and Henikoff, S. (1996) Using substitution probabilities to improve position-specific scoring matrices. *Computer applications in the biosciences*, **12**, 135-143.
47. Lu, H., Lu, L. and Skolnick, J. (2003) Development of unified statistical potentials describing protein-protein interactions. *Biophysical Journal*, **84**, 1895-1901.
48. Saha, R.P., Bahadur, R.P. and Chakrabarti, P. (2005) Interresidue contacts in proteins and protein-protein interfaces and their use in characterizing the homodimeric interface. *Journal of proteome research*, **4**, 1600-1609.
49. Shindyalov, I.N. and Bourne, P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, **11**, 739-747.
50. Schneider, R. and Sander, C. (1996) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Research*, **24**, 201-205.
51. Stewart, M., Kent, H.M. and McCoy, A.J. (1998) Structural basis for molecular recognition between nuclear transport factor 2 (NTF2) and the GDP-bound form of the Ras-family GTPase Ran. *Journal of Molecular Biology*, **277**, 635-646.
52. Koepp, D.M. and Silver, P.A. (1996) A GTPase controlling nuclear trafficking: running the right way or walking RANdomly? *Cell*, **87**, 1-4.
53. Gorlich, D., Pante, N., Kutay, U., Aebi, U. and Bischoff, F.R. (1996) Identification of different roles for RanGDP and RanGTP in nuclear protein import. *The EMBO journal*, **15**, 5584-5594.
54. Wong, D.H., Corbett, A.H., Kent, H.M., Stewart, M. and Silver, P.A. (1997) Interaction

between the small GTPase Ran/Gsp1p and Ntf2p is required for nuclear transport. *Molecular and Cellular Biology*, **17**, 3755-3767.

55. Clarkson, W.D., Corbett, A.H., Paschal, B.M., Kent, H.M., McCoy, A.J., Gerace, L., Silver, P.A. and Stewart, M. (1997) Nuclear protein import is decreased by engineered mutants of nuclear transport factor 2 (NTF2) that do not bind GDP-Ran. *Journal of Molecular Biology*, **272**, 716-730.
56. Grigoriev, A. (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Research*, **29**, 3513-3519.
57. Todd, A.E., Marsden, R.L., Thornton, J.M. and Orengo, C.A. (2005) Progress of structural genomics initiatives: an analysis of solved target structures. *Journal of Cell Biology*, **348**, 1235-1260.
58. Burgess, W.H. and Maciag, T. (1989) The heparin-binding (fibroblast) growth factor family of proteins. *Annual review of biochemistry*, **58**, 575-606.
59. Basilico, C. and Moscatelli, D. (1992) The FGF family of growth factors and oncogenes. *Advances in cancer research*, **59**, 115-165.
60. Plotnikov, A.N., Hubbard, S.R., Schlessinger, J. and Mohammadi, M. (2000) Crystal structures of two FGF-FGFR complexes reveal the determinants of ligand-receptor specificity. *Cell*, **101**, 413-424.
61. Ornitz, D.M., Xu, J., Colvin, J.S., McEwen, D.G., MacArthur, C.A., Coulier, F., Gao, G. and Goldfarb, M. (1996) Receptor specificity of the fibroblast growth factor family. *The Journal of biological chemistry*, **271**, 15292-15297.
62. Clackson, T. and Wells, J.A. (1995) A hot spot of binding energy in a hormone-receptor interface. *Science*, **267**, 383-386.
63. Thorn, K.S. and Bogan, A.A. (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, **17**, 284-285.

