國立交通大學

生物資訊所

碩士論文

利用蛋白質結構中的功能區域交互作用推測
蛋白質交互作用

Inferring Protein-protein Interactions from Structural
Domain-domain Interactions

研究生：陳宏助

指導教授：楊進木　博士

中華民國九十四年七月

利用蛋白質結構中的功能區域交互作用推測蛋白質交互作用

Inferring Protein-protein Interactions from Structural Domain-domain Interactions

研究生：陳宏助 Student：Hung-Chu Chen

指導教授：楊進木 Advisor：Jinn-Moon Yang

國立交通大學

生物資訊所

碩士論文

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 2005

Hsinchu, Taiwan, Republic of China

中華民國九十四年七月

利用蛋白質結構中的功能區域交互作用推測蛋白質交互作用

學生：陳宏勛　　　　　　　　　　　　　指導教授：楊進木

國立交通大學生物資訊所碩士班

摘　　　要

功能區域-功能區域之間的交互關係對於研究、預測以及註解蛋白質-蛋白質交互作用是很有幫助的。然而目前大規模透過實驗證實的的功能區域-功能區域交互作用還無法產生。在本論文中我們發展了一套新的方法從已知三級結構的蛋白質聚合物中，計算粹取功能區域-功能區域交互作用以及利用其來預測蛋白質-蛋白質交互作用。我們總共得到 1008 對的功能區域-功能區域交互作用，我們認為還有眾多的蛋白質-蛋白質交互作用是透過這些已知的功能區域-功能區域交互作用所產生。在不同的分數條件下，利用這些功能區域-功能區域交互作用，我們的方法總共預測了超過 101483 對的蛋白質-蛋白質交互作用。

我們建構了一個叫做「DAPID」的蛋白質-蛋白質交互作用資料庫，DAPID 收錄了結構上的功能區域-功能區域交互作用跟我們利用這些功能區域所預測出來的蛋白質-蛋白質交互作用以及目前 DIP 所有的蛋白質資料。DAPID 包含了從 1008 對的功能區域-功能區域交互作用所預測出來的 101483 對的蛋白質-蛋白質交互作用（72% ），1241 對在 PDB 資料庫中已知的蛋白質-蛋白質交互作用（0.8% ）以及 DIP 資料庫中所紀錄的 38131 對的蛋白質-蛋白質交互作用（27% ）。DAPID 主要包含了 8 種物種的蛋白質-蛋白質交互作用，包括 *Homo sapiens*、*Mus musculus*、*Rattus norvegicus*、*Drosophila melanogaster*、*Caenorhabditis elegans*、*Saccharomyces cerevisiae*、*Helicobacter pylori* 以及 *Escherichia coli*。我們的結果顯示我們的計分函式的值與 TP/FP 比值跟基因表現的相關係數都呈現高度正相關。將我們的計分函式的分數門檻設定在 0.5 時，所預測出來的蛋白質-蛋白質配對在基因表現相關係數上與 Jansen 等所產生的非交互作用蛋白質配對或 DIP 中的 *S. cerevisiae* 蛋白質做基因表現的相關係數的比較，我們預測的蛋白質-蛋白質配對在基因表現相關係數上都較為顯著。

Inferring Protein-protein Interactions from Structural Domain-domain Interactions

Student: Hung-Chu Chen
Advisor: Jinn-Moon Yang

Institute of Bioinformatics
National Chiao Tung University

ABSTRACT

Domain-domain interactions can be useful for validating, annotating, and predicting protein-protein interactions. Currently, the large-scale experimentally determined domain-domain interactions do not exist. In this thesis, we have developed an approach to computationally derive protein-protein interactions and domain-domain interactions from 3D protein complexes. We obtained 1008 interacting domain-domain from Protein Data Bank (PDB) and considered many protein pairs may also interact by the same interacting domain pairs. Our method predicted over 101483 protein-protein interactions based on interacting domain pairs and different thresholds of our new scoring function.

We have developed a domain-annotated protein-protein interaction database, termed DAPID, based on these inferred protein-protein interactions and experimental database, Database of Interacting Protein (DIP). The DAPID includes 101483 protein-protein interactions (72%) derived from 1008 structural domain pairs from Protein Data Bank (PDB), 1241 interactions (0.8%) directly obtained from 3D complexes in PDB, and 38131 interactions (27%) summarized from Database of Interacting Protein (DIP). The DAPID has eight common animal models, including *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Helicobacter pylori* and, *Escherichia coli*. Experimental results show that the gene-expression profiles (and the ratio of true to false positives (TP/FP) are highly correlated to the values of our scoring function. At the same time, our predicting protein-interaction pairs, whose scores greater than 0.5 have higher probability to co-express than the ones of non-interacting set (defined in Jansen *et al.* , Science 2003) or DIP set in *S. cerevisiae*.

致　　謝

在這兩年來的研究所生涯中，首先衷心感謝指導教授—楊進木老師。在整個研究過程中，老師花費了很多心力及時間，悉心指導我的論文與研究。老師對於學術研究的細心、嚴謹、堅持與執著，將是我未來持續學習的目標。

這本論文能夠順利完成，要感謝實驗室所有同學的幫忙。特別感謝強哥、一原，強哥敏銳的研究天份以及撰寫程式的功力，讓我獲益良多，讓我的研究得以順利進行。一原是長腳的函式庫、我只要描述我的 input 格式跟我要的 output 格式，他就能給我一個函式幫我搞定。還要感謝跟我一起住的室友，一起討論程式的朋友，一起襪三的戰友，一起吃飯的飯友。

另外要感謝我的家人，默默地在背後給我支持、鼓勵與動力，讓我可在求學過程中無所顧忌，全力衝刺。在我心情最低落的時候，給予我無盡的鼓勵、安慰與溫暖。在我情緒最愉快的時刻，陪我一起出遊、聊天，抒解壓力。沒有他們的陪伴，我很難走到現在。

要感謝的人太多太多了，感謝所有幫助過我的人。

宏助

夏'05

CONTENTS

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Motivations and Purposes

The cellular machinery is a complex dynamic system with a multitude of biomolecular interactions. Simply knowing the existence of genes and proteins does not tell us much about the biological processes in which they participate. The interaction between proteins is one of the most important features of protein functions. A comprehensive description of protein-protein interactions is necessary to understand the genetic program of life. Behind protein-protein interactions there are protein domains interacting physically with one another to perform the necessary functions. Therefore, understanding protein interactions at the domain level gives a global view of the protein interaction network, and possibly of protein functions.

Domains are structural subunits of proteins that can be thought of as ‘building blocks’ that are conserved during evolution. Proteins can consist of a single domain, or more frequently as a combination of several domains: in prokaryotes about one-third of all proteins are single domain while in eukaryotes the fraction is about 20% [1]. Nature is able to construct a vast array of different proteins by combining domains, and the present-day variety of domains is believed to have evolved from a relatively small number of ancestral gene sequences [2]. Domains, which are related to each other by descent from a common ancestor are said to be homologous, can be grouped together to form a superfamily.

Lots of computational methods have been developed to predict protein-protein interactions. Because proteins interact with one another through their specific domains, predicting domain-domain interactions on a global scale from the entire protein interaction data set make it possible to predict previously unknown protein-protein interactions from their domains. Protein sequence databases such as Swiss-prot [3] is becoming increasingly large and unmanageable, primarily as a result of the growing number of genome sequencing projects. However, many of the newly added proteins are new members of existing protein families. Typically, between 40% and 65% of the proteins found by genomic sequencing show significant sequence similarity to proteins with known function [4; 5] and usually a large fraction of them show similarity with each other [5; 6]. Figure 1 shows the number of proteins discovered each year and the number of presently known nonredundant domains found in these proteins. Although the number of proteins discovered grows at increasingly higher rates, the number of domains found appears to be asymptotically reaching a limit.

Experimentally determined structures of protein complexes are deposited in the Protein Data Bank (PDB) [7]. The PDB is growing rapidly, in part due to the recent structural genomics effort. The PDB currently holds approximately 30000 structures. Each entry contains on average 2.2 protein chains, and each chain contains on average 2.1 domains [8]. Domains are considered the basal unit of protein structure, function and evolution [9]. These units fold independently, often mediate a specific biological function, and combine modularly to form larger proteins. Several approaches to the definition of domain boundaries in proteins have been developed based on sequence and structure [10]. The Pfam [11] is the most commonly used sequence-based domain definition and classification system.

Domain-domain interactions can be useful for validating, annotating, and even predicting protein-protein interactions. However, large-scale experimentally determined domain-domain interactions are not available [12]. In this thesis, we have developed an approach to computationally derive protein-protein interactions and domain-domain interactions from 3D protein complexes. Using evolutionarily conserved domains defined in a protein-domain database called Pfam [11], to identify interacting domains that are consistent with the observed protein-protein interactions. Because every protein can be characterized by either a distinct domain or a combination of domains, understanding domain interactions is crucial to understanding the nature and extent of biomolecular interactions. Our study identified structural binding site and map it to sequence-based domain-domain interactions solely on the basis of the information of 3D protein-protein interactions. Because proteins interact with one another through their specific domains, identify structural domain-domain interactions on a global scale from the entire protein interaction data set make it possible to predict previously unknown protein-protein interactions from their domains. Thus, domain interactions extend the functional significance of proteins and present a global view of the protein-protein interaction network within a cell responsible for carrying out various biological and cellular functions.

1.2 Related Works

Computational methods have been developed to predict protein-protein interactions. Those approaches include the Rosetta stone/gene fusion method [13; 14], the phylogenetic profile method [15] and the method combining multiple sources of data [16]. Other computational methods to predict protein-protein interaction have been presented on the basis of different principles, including the interaction domain pair profile method [17; 18] and the support vector machine learning method [19]. Gomez *et al.* [20] developed probabilistic models for protein-protein interactions. Sprinzak and Margalit [21] analyzed over-represented sequence-signature pairs among protein-protein interactions.

The yeast two-hybrid technique is a high throughput approach that has been used to screen the yeast genome for pairwise interactions by two different groups [22; 23]. The advantage of screening the entire genome (or at least a large part of it) for interactions is offset by a high rate of experimental error, and it is vital that the false positive rate is included in any analysis of the data. The potential benefits and drawbacks of such high throughput techniques have generated intense interest: see Legrain *et al.* [24] and von-Mering *et al.* [25] for reviews of this area, and Sprinzak *et al.* [26] for estimates of the false positive rate [2]. Given a list of interactions obtained from such experiments, absence of a protein pair from the list does not necessarily mean that the two proteins do not interact. It could be that interaction of the pair in question was not tested in the experimental set-up. This false negative probability has to be accounted for when making statistical inference based on absence from the list of interactions.

We notice that, large scale experimentally determined domain-domain interaction data are not available. In the past research, usually infer domain-domain interactions from protein-protein interactions (e.g. yeast-two-hybrid experiments). However, it is known that the yeast two-hybrid assay is not accurate in determining protein-protein interactions, and the interaction data certainly contain many false positive and false negative errors [27; 28; 29]. To infer domain-domain interactions from these large scale experimentally determined protein-protein interactions might not reflect the real domain-domain interactions. The other shortcoming of past approaches is that they did not explicitly know which domain pair is the interacting domains.

1.2.1 Probabilistic Domain-domain Interaction Models

Ng *et al.* [12] collected data from three sources: (i) the experimentally derived protein interaction data from DIP [30]; (ii) the intermolecular relationship data from protein complexes; and (iii) the computationally predicted data from Rosetta Stone sequences. Then they inferred putative domain–domain interactions based on the collected data through the development of InterDom, a database of interacting domains [31]. However, the accuracy of the data inferred from domain–domain interaction is not apparent. The approach of InterDom is a standard probabilistic domain-domain interaction model. In the large number of interacting protein data, domain pairs which with higher occurrence frequency will denote as high interacting potential domain pairs. This approach might bias to the common domains in protein composition.

Deng *et al.* [32] proposed a probabilistic prediction model for inferring domain interactions from protein interaction data. The maximum-likelihood estimation technique is mainly used in their method. It is known that the yeast two-hybrid assay is not accurate in determining protein-protein interactions, and the available interaction data certainly contain many false positive and false negative errors. Taking into account these errors, they apply the maximum-likelihood approach to estimate the probability of domain-domain interactions. They apply the maximum-likelihood approach to estimate the probability of domain-domain interactions and recursively to derive domain-domain interaction probabilities from the combined experimental data. The Pfam database is used to extract domain information and the MIPS [33] database is used to test their model. Finally, they compare the gene expression profile correlation coefficients of their predictions with those of random protein pairs and MIPS database.

Recently, predictions of protein interactions have been performed in the context of domain-domain interactions using experimentally identified interacting protein pairs. These approaches based on statistic theories might be weak on biological meanings.

1.2.2 Interologs

Walhout *et al.* [34] introduced the concept of “Interolog”: orthologous pairs of interacting proteins in different organisms. Yu *et al.* [35] apply this concept into a concrete interaction prediction approach, the generalized interolog mapping method. Interolog map the transfer of interaction annotation from one organism to another using comparative genomics. Yu *et al.* think that it is now possible to examine the degree to which protein-protein and protein-DNA interactions are transferred between different organisms as a function of the underlying sequence similarities of the interacting proteins. They comprehensively assessed the transferability of protein-protein and protein-DNA interactions by analyzing the relationships between sequence similarity and interaction conservation. Their approach based on interolog to explore the protein-protein interactions in another organism. We thought that the interolog method does not take the domain information into account. If it replenishes the domain information and biological meaning, it will be more reliable and robust.

1.3 Thesis Overview

In recent years, the PDB database has experienced rapid growth. To maximize the utility of the high resolution protein-protein interaction data stored in the PDB database, we developed a novel domain-based protein-protein interaction predict approach and build a database include structural domain-domain interactions and protein-protein interactions derived from structural domain pairs.

In chapter 2, we have prepared structural interaction protein pairs from PDB database. We divided structural binding site from structural contacted proteins in PDB database and mapped it to domain-domain interactions. Then, we inferred protein-protein interactions from structural domain-domain interactions. After that, we developed a computational approach to measure the similarity between the structural protein (template) and the others which we inferred (candidates). Our approach not only takes sequence similarity into account but also the biological function similarity. Finally, we also compared the gene expression profile correlation coefficients of our predictions with those of random protein pairs and DIP database, and our predictions have a higher mean correlation coefficient.

In chapter 3, we show the 101483 protein-protein interactions we inferred (Table 1) and the 1008 pairs of domain-domain interactions from PDB database (Table 2). In PDB database we obtain 36465 protein-protein interactions of known 3D structure making of 1008 type of domain-domain interactions. Based on these domain-domain interactions, we inferred over one million protein-protein interactions. By the way, we applied our scoring function to reduce the false positives. In addition, we demonstrated the scoring weight of each scoring term to reduce false positive rate. Finally, we applied our approach on eight common organism models and inferred 101483 protein-protein interactions which not stored in DIP. Based on the gene-expression profiles and the ratio of true to false positives (TP/FP), we determined the threshold of our scoring function as 0.5. Experimental results show that the gene-expression profiles and the ratio of true to false positives (TP/FP) are highly correlated to values of our scoring function.

Chapter 4 presented some conclusions and future works. Our major contribution is to develop a novel approach to infer protein-protein interactions from domain-domain interactions. We will add more features to enhance our scoring function and use the evolutionary approach to demonstrate the scoring weight of our scoring function. Furthermore, we will visualize our achievements as a web service.

Chapter 2 Materials and Methods

The core idea of this thesis is illustrated in Figure 2. In this thesis, we analyzed the structural protein-protein binding sites from PDB database to identify the domain-domain interactions. Then we inferred protein-protein interactions from the structural domain-domain interactions we obtained. Our approach included two parts: **Identify structural domain-domain interactions** and **Inferring protein-protein interactions from structural domain-domain interactions**. Two types of input data were used: protein structures and domain definitions. Structures were obtained from the PDB database. Domain definitions for the PDB structures were obtained from the Pfam classification system. First, inter-atomic distances were calculated for all structures using a specified distance cutoff. We analyzed the domain architectures of protein complex binding sites from the PDB database (Figure 3A). We mapped the structural binding site into the Pfam domains to identify the binding domains (Figure 3B). We have developed an approach and a new scoring function to yield many protein-protein interaction candidates (over one million) (Figure 4). Based on the gene-expression profiles and the ratio of true to false positives (TP/FP), we adapt the thresholds of our new scoring to reduce false positive rate. Experimental results show that the gene-expression profiles and the ratio of true to false positives (TP/FP) are highly correlated to values of our scoring function.

2.1 Preparing Dataset from PDB Database

To understand how protein domains interact at the molecular level, we need to know which residues, and their constituent atoms, in each protein are interacting. This kind of data is available in the PDB database of protein structures [7] where multiple domains are present in a single structure. We used the data of PDB database published at 22nd April 2005. It holds

29835 entries. We parsed the coordinate data of each C_β atom in PDB format text file of each entry to identify the protein-protein interaction binging sites (Figure 3A). To identify the protein-protein binding site, we calculated the C_β - C_β distance (GLY C_α) of each residue belongs to different chain. We introduce the rule we used: the 5-8 rule. It must five or more residues contact within 8Å. The binding site information between each chain were computed and stored.

2.2 Mapping Structural Binding Site to Domain-domain Interactions

In order to identify the domain architectures of protein-protein binding sites, we mapped the structural protein-protein binding sites to the Pfam defined domains (Figure 3B). We followed the two steps below:

2.2.1 Transform 4-Letter PDB and Chain Code to Swiss-prot Accession Number

We downloaded the “uniprot_sprot.dat.gz” (version 45) from:

ftp://us.expasy.org/databases/swiss-prot/release/

The Universal Protein Resource (UniProt/Swiss-prot) [36] provides the scientific community with a single, centralized, authoritative resource for protein sequences and functional information. The database cross-references in Swiss-prot are used as pointers to information related to entries and found in data collections other than Swiss-Prot database. For example, if the X-ray crystallographic atomic coordinates of a sequence are stored in the PDB database there will be one database cross-reference pointing to each of the corresponding entries in PDB database. After uncompressed the uniprot_sprot.dat.gz, we got a huge text file. It recorded the relationship between Swiss-prot accession number and PDB chain code. We transformed each PDB chain code to Swiss-prot accession number.

Unfortunately, not every entry in PDB has corresponding entry in Swiss-prot. If the PDB chain code could not find related Swiss-prot accession number, we employed BLAST [37] to identify the Swiss-prot accession number.

We download:

1) Standalone BLAST 2.2.11 from:

ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/

2) Swiss-prot fasta format BLAST database (version 45) from:

ftp://us.expasy.org/databases/swiss-prot/release/

Use program: “formatdb” to create our own database for BLAST searching. We execute BLAST by program “blastall”. Blastall may be used to perform all five flavors of blast comparison. A typical use of blastall would be to perform a “blastp” search (protein vs. protein) of a query file called INPUT would be:

blastall -p blastp -d DATABASE -i INPUT -o OUTPUT -M BLOSUM62 -G 8 -E 2 -F F

The output will be placed into the result OUTPUT and the search is performed against the “DATABASE” database. Other blastall options showed above “-M BlOSUM62” which is default scoring matrix, “-G 8 –E 2” which means that open gap penalty is 8 and extend one is 2, and “-F F” is to tell blastall do not filter query sequence. In this thesis, DATABASE we use the swissprot.fasta and default setting of BLAST.

We took sequence data of those PDB entries could not find corresponding entries in Swiss-prot as INPUT. BLAST will help us to identify Swiss-prot accession number of these PDB entries. We took the rank one Swiss-prot accession number of BLAST output as our query sequence.

2.2.2 Identify Domain Architectures of Structural Binding Site

We download the "swisspfam.gz" from:

ftp://ftp.sanger.ac.uk/pub/databases/Pfam/

Pfam is a large collection of protein families and domains. Pfam is represented by two multiple sequence alignments and two profile-Hidden Markov Models (profile-HMMs). In swisspfam, it well defined the domain boundary on Swiss-prot protein sequences. It is great help to us to identify the binding domains. We mapped the Swiss-prot/PDB in residue level. The residue number will indicate which domain it is localized. Usually, the same gene product stored in PDB and Swiss-prot will identical in sequence data. But the residue number might shift or the N' and C' terminal residues in PDB might be lost. Therefore, the residue number record in Swiss-prot and PDB might not the same. We must modify the PDB residue numbers to match Swiss-prot residue numbers. In order to map the Swiss-prot/PDB in residue level, we download "pdb_chain_uniprot.lst" from:

ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts/text/

In pdb_chain_uniprot.lst, it contains a summary of the PDB to Swiss-prot residue level mapping. We conform to the data in pdb_chain_uniprot.lst to modify the PDB residue numbers to match Swiss-prot residue numbers.

The number of residues in protein-protein binding site and located in Pfam defined domains must more than five. After the process above, we accurate known which residues provide the binding force between two protein chains and which domain they located.

2.3 Inferring Protein-protein Interactions from Domain-domain Interactions

2.3.1 Generate Protein-protein Interaction Candidates

Because of domain-domain interactions will be a good indicator to infer protein-protein interactions and domains are structural subunits of proteins that can be thought of as “building blocks” that are conserved during evolution. We can infer protein-protein interactions stride across different organisms via domain-domain interactions. For instance, we observed a domain-domain interaction in *Saccharomyces cerevisiae*; we can infer protein-protein interactions in other organisms (e.g. *Homo sapiens*) with similar domain-domain pair composition. We generated the protein-protein interaction candidates according to the domain composition of proteins; the domain is defined as Pfam domain (Figure 4). We employed the “swisspfam” to help us to identify the domain composition in proteins. We inferred over one million protein-protein interactions via domain-domain interactions which we obtained. Each structural domain pair will generate thousands of protein-protein interaction candidates. We applied our approach on eight common organism models, including *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Helicobacter pylori* and, *Escherichia coli*.

2.3.2 Scoring Function

To generate the protein-protein interaction candidates without any criteria will contain lots of false positives in candidates. In our approach, we inferred protein-protein interactions according to domain composition in proteins. Because of lots of proteins might contain the same sequence-based domain with various biological functions. We developed a scoring function to measure the similarity between the proteins we inferred (candidates) and the original proteins we identified domain-domain interactions (templates). Our scoring function is based on biological annotation, sequence similarity and the degree of bind site conservation.

Our scoring function is given as

$$S_{total} = W_k \times S_k + W_e \times S_e + W_p \times S_p + W_b \times S_b \qquad (1)$$

where W_k, W_e, W_p, and W_b are the weights for S_k, S_e, S_p, and S_b. The S_k is the score to measure biological function similarity between proteins. S_e and S_p are the scores of measuring the sequence similarity between two proteins. S_b is the score to measure the degree of binding site conservation.

We developed a novel approach to measure biological function similarity between proteins. Swiss-prot is a curated protein sequence database which strives to provide a high level of annotation. We notice the annotation in Swiss-prot will provide as a good indicator to identify the biological function similarity between proteins. We developed a scoring function to measure biological function similarity between proteins. It is similar to information retrieval. We download “uniprot_sprot.dat.gz” from:

ftp://us.expasy.org/databases/swiss-prot/release/

The uniprot_sprot.dat is a high level of annotation such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc. We focus on the keywords annotation (Figure 5) of each protein. Table 3 shows the total keywords in Swiss-prot database. We consult the well-known information retrieval skill to develop our scoring function. The keywords annotated of each protein should have different importance. According to each keyword frequency, we transformed the frequency to score by TF/IDF (Term Frequency × Inverse Document Frequency) approach. Term Frequency of a word is the number of its occurrence in a corpus. In our study, the TF value will be 1, because the keyword will not appear twice in a protein annotation. Inverse Document Frequency of a word is the number of document where the word occurs at least once. The IDF of each keyword is given as

$$IDF_i = V_i = \log_2^{s/fi} \tag{2}$$

where IDF_i as score of keyword i (i=1 to n, n as total types of keywords in Swiss-prot), V_i as vector of keyword i, s as total number of protein stored in Swiss-prot (in this thesis s=163235), fi as the frequency of keyword i in Swiss-prot database (we remove the keyword $fi < 10$ and $fi > 10000$). After we transform the frequency of each keyword to score, we notice that some keywords might more important than others. We download “spkw2go” from:

http://www.geneontology.org/external2go/spkw2go

The spkw2go means that Swiss-prot keywords mapping to Gene Ontology (GO)[38]. Some keywords will repeat in this data. We consider that these repeat keywords might have more significant importance (Table 2). In our scoring function, it will calculate these keywords twice or treble. In other hand, we remove the keywords IDF of which keywords with huge frequency. Because of the keyword appear in each protein will not have significance to distinguish each other. Each keyword have its own score, a protein might contain more than one keywords. How to measure the similarity between proteins via keywords annotations? We apply the “Vector Space Model” to do this. The vector space model is given as

$$S_k(Px, Py) = \frac{\sum_{i=1}^{n}(V_i x \times V_i y)}{\sqrt{\left(\sum_{i=1}^{n} V_i x^2\right) \times \left(\sum_{i=1}^{n} V_i y^2\right)}} \tag{3}$$

where Px, Py are protein x and protein y, n as total unique keyword in Swiss-prot database (n dimensional vectors, in this thesis n=949), $V_i x$, $V_i y$ as the vector strength of protein x, y in ith dimension. In this scoring function, we extracted the proteins corresponding to the template protein in keywords annotation. After the S_k were calculated, we normalize the score to Z-score and scale to 0~1. The normalized Z-score is used for measuring the keywords score separation between template and candidates.

We employed PSI-BLAST [39] to measure the sequence similarity between proteins. In session 2.2.1 we had download the standalone BLAST and fasta format sequence data. The command to perform PSI-BLAST:

```
blastpgp –d DATABASE –i INPUT –o OUTPUT F F –G 8 –E 2 –j 3 –t F –h 5
```

Program “blastpgp” take a protein query and perform PSI-BLAST search to create a position specific matrix using a protein database. Some of arguments used in PSI-BLAST are the same as BLAST. There are different options between BLAST and PSI-BLAST, such as “-j 3” which is maximum number of rounds, “-t F” which means that program do not use composition based statistics, and “-h 5” that is the e-value threshold for including sequences in the score matrix model. The e-value threshold default is 0.001. However in order to obtain correct result and best performance, we change the value from 0.001 to 5 for PSI-BLAST. The top part of output of PSI-BLAST for each round distinguishes the sequences into: sequence found previously and used in the score model, and sequences not used in the score model. The output currently includes lots of diagnostics requested by users in NCBI. To skip quickly from the output of one round to the next, search for the string “producing”, that is part of the header for each round and likely does not appear elsewhere in output. PSI-BLAST “converges” and stops if all sequences found at round i+1 below the e-value threshold were already in the model at the beginning of the round.

We took the sequence of the template protein as PSI-BLAST input and change the e-value threshold to 5; the maximum number of rounds as 3. The S_e, S_p and S_b are base on result of PSI-BLAST result.

The S_e is given as

$$S_e = -\log^E \tag{4}$$

where E means the e-value of the candidate protein. We transform e-value of PSI-BLAST output to our sequence similarity score. Then, we normalize the score to Z-score and scale to 0~1.

The S_p is given as

$$S_p = pos + iden \tag{5}$$

where *pos* is the positive percentage of PSI-BLAST sequence alignment result; *iden* is the identical percentage of PSI-BLAST sequence alignment result. Then, we normalize the score to Z-score and scale to 0~1.

The S_b is given as

$$S_b = \frac{b_{candadate}}{b_{template}} \tag{6}$$

where $b_{candidate}$ means that the binding site conservation percentage of PSI-BLAST sequence alignment result (candidate); $b_{template}$ is the binding site score of query protein (template).

The $b_{candidate}$ is given as

$$b_{candidates} = \sum_{i=0}^{n} S_i \tag{7}$$

where n as the length of sequence alignment of binding site, S_i means that the substitution score of sequence alignment result at position i. The $b_{template}$ is completely imitating the same equation. We use the “BLOSUM62” substitution matrix to calculate the binding site score for each PSI-BLAST sequence alignment result. Then, we transform the score to percentage. Then, we normalize the score to Z-score and scale to 0~1.

Our purpose is to find out proteins which were composed of specific domain and similar to template protein in biological functions. We sum the four scoring terms of each protein candidate. The total score will show that how similar between template and candidates. In general, the template protein will be the highest score in our scoring function. We define a threshold to determine how the candidate similar to template protein. Our threshold is defined as the percentage between candidates and template protein.

2.4 Verification

2.4.1 TP/FP Ratio

In order to prove our strategy is reasonable, we prepare two standard datasets, one is the positive dataset which contains 14779 *Saccharomyces cerevisiae* interacting protein pairs in DIP [30], the other is the negative dataset which contains 2599785 non-interacting protein pairs in *Saccharomyces cerevisiae* [40]. The TP/FP ratio is defined that our predicting protein-pairs overlapping with the positive set divides overlapping with the negative set [40]. We calculated the TP/FP ratio in different thresholds and different weights of each scoring term to observe the relation between TP/FP ratio and thresholds.

2.4.2 Genes Expression Profiles

Recently, it was noted that genes with similar expression profiles are likely to encode interacting proteins [41; 42]. We studied the distribution of correlation coefficients for protein pairs with predicted interaction probability greater than a certain threshold. We used the gene expression data from Hughes *et al.* [43]. At the same time, our predicting protein-interaction pairs, whose scores greater than 0.5 have higher probability to co-express than the ones of non-interacting set (defined in Jansen *et al.* , Science 2003 [40]) or DIP set in *S. cerevisiae*.

Chapter 3 Results and Discussions

3.1 Inferred Protein-protein Interactions from Domain-domain Interactions

Currently, in our structural domain-domain interaction database includes information on 36465 protein chains of known 3D structure making a total of 1008 type of domain-domain interactions. Of these, we grouped these interactions into 1008 types according to the Pfam domain mediating them (Table 2). We used the domain-domain interactions to predict protein-protein interactions and assess the prediction accuracies at the protein level. After calculated the score of each protein and template, we applied these domain-domain interactions to infer protein-protein interactions on eight common organism models, including *Homo sapiens, Mus musculus, Rattus norvegicus, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae, Helicobacter pylori and, Escherichia coli*. In these organisms, we inferred 53669 protein-protein interactions in *Homo sapiens*; 39689 protein-protein interactions in *Mus musculus*; 4461 protein-protein interactions in *Rattus norvegicus*; 857 protein-protein interactions in *Drosophila melanogaster*; 941 protein-protein interactions in *Caenorhabditis elegans*; 1130 protein-protein interactions in *Saccharomyces cerevisiae*; 603 protein-protein interactions in *Escherichia coli*; 133 protein-protein interactions in *Helicobacter pylori*. In total, we inferred 101483 protein-protein interactions (Table 1). It is a great quantity more than the data stored in DIP database.

3.2 Determine the Threshold

In this thesis, our scoring function contains four scoring terms and each has its own weight. We tested different weight combination of scoring terms (Figure 6). We considered that each scoring term of our scoring function has certain effect. We thought that the function similarity, sequence similarity and binding site conservation are equally important. In our opinion, W_e and W_p are both extended from PSI-BLAST to measure the sequence similarity. In order to take each scoring term into account, we determine the W_k as 1, W_e as 0.5, W_p as 0.5 and W_b as 1.

The benchmark we used to verify our predictions is TP/FP ratio. The TP/FP ratio is defined that our predicting protein-pairs overlapping with the positive set divides overlapping with the negative set. The positive dataset which contains 14779 *Saccharomyces cerevisiae* interacting protein pairs in DIP [30], the other is the negative dataset which contains 2599785 non-interacting protein pairs in *Saccharomyces cerevisiae* [40]. Experimental results show that the TP/FP ratio is highly correlated to values of our scoring function (Figure 7). Because of Jansen *et al.* indicated that the TP/FP ratio greater than 1 the predictions will be more reliable [40]. The Figure 7 shows that the scores of our scoring function are highly correlated to the TP/FP ratio. We defined a threshold at TP/FP ratio as 1. When TP/FP ratio as 1 the threshold is 0.5. The threshold as 0.5 means that the score of candidate protein divides the score of template must greater than 0.5.

Figure 8 shows the number of inferred protein-protein interactions (candidates) at different TP/FP ratio. In Figure 7 and Figure 8, we found that with the severer threshold the numbers of candidates were going down and the TP/FP ratios were on the rising. The TP/FP ratio will reflect the quality of our predictions. The TP/FP ratio is regarded as the accuracy of our predictions. At higher accuracy, we will get the high quality predictions, but the number of inferred protein-protein interactions will be decreased. We determine the threshold at 0.5. It is a compromise between quality and quantity. In Figure 7 and Figure 8, the TP/FP ratio is the most important information. Because of the TP/FP ratio is defined as our predicting protein-pairs overlapping with the positive set divides overlapping with the negative set. When the number of our predicting protein-pairs overlapping with the positive set is too low, the TP/FP ratio might lose its statistic meaning. And it brought out the perturbations in Figure 7 and Figure 8 when positive-overlap was insufficient.

3.3 Correlation Coefficients of Gene Expression

Genes with similar expression profiles are likely to encode interacting proteins. We study the distribution of correlation coefficients for protein pairs with predicted interaction probability greater than a certain threshold. Figure 9 shows that the gene-expression profiles are highly correlated to values of our threshold. We compared the gene expression profile correlation coefficients of our predictions with those of random protein pairs and DIP database, and our predictions have a higher mean correlation coefficient.

3.4 Examples

We sought an example from the literature to illustrate the operation and accuracy of the method. Some of the most intensively studied interactions are those between fibroblast growth factors (FGFs) and receptors. FGF signaling pathways are intricate and are intertwined with insulin-like growth factor, transforming growth factor-beta, bone morphogenetic protein, and vertebrate homologs of Drosophila wingless activated pathways. FGFs are major regulators of embryonic development: They influence the formation of the primary body axis, neural axis, limbs, and other structures. The activities of FGFs depend on their coordination of fundamental cellular functions, such as survival, replication, differentiation, adhesion, and motility, through effects on gene expression and the cytoskeleton. FGFs play key roles in morphogenesis, development, angiogenesis, and wound healing. There are more than 20 human FGFs that bind to one or more of 7 FGF receptors (FGFR1c, -1b, -2c, -2b, -3c, -3b, and -4; c and b denote isoforms IIIc & IIIb formed by alternative splicing [44]). For example, PDB ID 1DJS is a protein complex (FGFR2 complex with FGF1), the chain A of 1DJS

(Swiss-Prot accession number: P21802) and chain B (Swiss-Prot accession number: P05230) are interacting with each other. In DIP database, chain A is 3788N and chain B is 3787N. We analyzed the domain architecture of the interface between them. We discovered that the interaction between these two chains can be reduced to immunoglobulin domain (Pfam ID: PF00047) interact with fibroblast growth factor domain (Pfam ID: PF00167). First, we analyzed the other proteins stored in DIP which interacted with 3788N. We found the other proteins interact with 3788N all constituted with fibroblast growth factor domain. We tried to seek more other proteins which were composed of fibroblast growth factor domain. We found 22 human proteins were composed of fibroblast growth factor domain and some of them have crystal structure (1II4, 1NUN) to confirm they indeed interact with FGF2 (P21802).

3.5 Web Service

We developed a web-base database named "DAPID" to present our result. DAPID has been setup to a web service as shown in Figure 10. Users can input a Swiss-prot accession number as a query. The DAPID will return the interacting partners of the query protein. And we will show the detail information about this pair of proteins. The website of DAPID is http://gemdock.life.nctu.edu.tw/dapid/.

Chapter 4 Conclusions and Future Works

4.1 Summary

We developed a novel approach to infer protein-protein interactions from domain-domain interactions. Domain-domain interaction information can be useful for predicting protein interactions. Proteins with the same domain compositions and biological function might have the same type of domain-domain interactions. Without other criterion, domain-based protein-protein interaction prediction will obtain lots of false positives. We developed a scoring function based on biological function similarity and sequence similarity to reduce the false positive rate. We applied our approach on eight common organism models, including *Homo sapiens, Mus musculus, Rattus norvegicus, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae, Helicobacter pylori and, Escherichia coli*. In these organisms, we inferred 53669 protein-protein interactions in *Homo sapiens*; 39689 protein-protein interactions in *Mus musculus*; 4461 protein-protein interactions in *Rattus norvegicus*; 857 protein-protein interactions in *Drosophila melanogaster*; 941 protein-protein interactions in *Caenorhabditis elegans*; 1130 protein-protein interactions in *Saccharomyces cerevisiae*; 603 protein-protein interactions in *Escherichia coli*; 133 protein-protein interactions in *Helicobacter pylori*. In total, we inferred 101483 protein-protein interactions. We employed TP/FP ratio to help us determine the threshold cutoff of our scoring function. And then we compared the gene expression profile correlation coefficients of our predictions with those of random protein pairs and DIP database, and our predictions have a higher mean correlation coefficient. Finally, we build a domain-annotated protein-protein interaction database named “DAPID”, including structural domain pairs and inferred protein-protein interactions.

4.2 Major Contributions and Future Works

Here, we have developed an approach to infer protein-protein interactions from domain-domain interactions. The most important contribution of this thesis is the domain-domain interactions we obtained are real interacting domains. It is quiet different to probabilistic domain-domain interaction models. Our domain-domain interactions are more significant and reliable than probabilistic domain-domain interaction models. According to these domain-domain interactions, we predicted 101483 protein pairs on eight common organism models. It is a great quantity more than the data stored in DIP or MIPS database.

Nowadays, we mapped the protein-protein binding site to Pfam domains to identify the domain-domain interactions. However, we considered that the Pfam defined domains are based on HMM-profile and multiple sequence alignment, it might lose structural information. Our method depends on the domain classification database to infer proteins which were composed of specific domains. The quality of domain classification database will great influence our predictions. In the future, we hope to infer proteins which were composed of specific domains including structural information. We will employ other domain classification systems which include structural information to improve our method.

Our scoring function depends on high quality, curated annotations and sequence similarity. We employed the annotation information in Swiss-prot database and took the “keywords” to measure the protein function similarity. We only took the “keywords” information to construct our scoring function. The keywords might be too rough to classify protein functions. We will combine other annotation in Swiss-prot database (e.g. descriptions, functions and EC numbers) to improve our scoring function.

Sometimes, protein-protein interactions might unnecessary belongs to the same organism (e.g. enzyme-inhibitor; virus-host proteins). Our method will ignore these kinds of protein-protein interactions. We will study the domains which will interact across organisms (e.g. inhibitors and virus proteins). We will modify our approach to recover these kinds of protein-protein interactions.

Another difficulty is that it is often impossible to distinguish between direct physical interactions and functional associations that may not involve direct atomic contacts between macromolecules. According to our approach, we focused on identify structural domain-domain interactions from crystal structures. These kinds of protein-protein interactions are all most the direct physical interactions. Therefore, our method might not be able to predict the transient protein-protein interactions. In the future works, we will identify other more domain-domain interactions from other data source (e.g. DIP and MIPS) to predict transient protein-protein interactions.

Table 1. Protein-protein interactions stored in different database. In total, we inferred 101483 protein-protein interactions on eight common organism models.

Species	Our predictions	DIP	PDB
Homo sapiens	53669	1783	482
Mus musculus	39689	312	326
Rattus norvegicus	4461	129	78
Drosophila melanogaster	857	12019	6
Caenorhabditis elegans	941	3878	1
Saccharomyces cerevisiae	1130	18274	259
Escherichia coli	603	1602	88
Helicobacter pylori	133	134	1
Total	**101483**	38131	1241

Table 2. 1008 Domain-domain interactions we inferred from structural protein-protein interactions.

PF00001-PF00001; PF00001-PF02978; PF00004-PF00004; PF00004-PF00705; PF00004-PF02617; PF00004-PF07499;
PF00005-PF00005; PF00005-PF01032; PF00006-PF00006; PF00006-PF00231; PF00006-PF00306; PF00006-PF04568;
PF00007-PF00007; PF00007-PF00236; PF00008-PF00008; PF00008-PF00089; PF00009-PF00009; PF00009-PF00889;
PF00009-PF03143; PF00009-PF03764; PF00010-PF01056; PF00011-PF00011; PF00012-PF00012; PF00012-PF01025;
PF00012-PF02179; PF00013-PF00013; PF00014-PF00014; PF00014-PF00051; PF00014-PF00089; PF00014-PF00431;
PF00015-PF00015; PF00016-PF00016; PF00016-PF00101; PF00017-PF00017; PF00017-PF00018; PF00018-PF00018;
PF00018-PF00469; PF00018-PF03114; PF00018-PF05038; PF00019-PF00019; PF00020-PF00020; PF00020-PF00229;
PF00022-PF00022; PF00022-PF00063; PF00022-PF00235; PF00022-PF00626; PF00022-PF03372; PF00022-PF04045;
PF00022-PF04062; PF00022-PF04699; PF00022-PF05856; PF00023-PF00023; PF00023-PF00069; PF00023-PF00554;
PF00023-PF01833; PF00024-PF00024; PF00025-PF00025; PF00025-PF01369; PF00025-PF01465; PF00025-PF05351;
PF00026-PF00026; PF00026-PF06394; PF00026-PF07966; PF00027-PF00027; PF00028-PF00028; PF00028-PF00514;
PF00030-PF00030; PF00032-PF00355; PF00032-PF01333; PF00032-PF02167; PF00032-PF02529; PF00032-PF02939;
PF00033-PF00033; PF00033-PF00355; PF00033-PF02921; PF00034-PF00034; PF00034-PF00124; PF00034-PF00141;
PF00034-PF02167; PF00035-PF00035; PF00036-PF00036; PF00036-PF00063; PF00036-PF00069; PF00036-PF00612;
PF00036-PF00992; PF00036-PF02063; PF00036-PF07679; PF00037-PF02427; PF00038-PF00038; PF00040-PF00040;
PF00041-PF00041; PF00041-PF00103; PF00041-PF00489; PF00041-PF00758; PF00041-PF00812; PF00041-PF03039;
PF00041-PF07686; PF00042-PF00042; PF00043-PF00043; PF00044-PF00044; PF00045-PF00965; PF00047-PF00047;
PF00047-PF00516; PF00047-PF07654; PF00048-PF00048; PF00050-PF00050; PF00050-PF00082; PF00050-PF00089;
PF00051-PF00051; PF00051-PF00079; PF00051-PF00594; PF00051-PF00713; PF00051-PF02821; PF00051-PF03974;
PF00051-PF07648; PF00056-PF02866; PF00057-PF00073; PF00059-PF00059; PF00059-PF00089; PF00059-PF00092;
PF00059-PF00129; PF00059-PF01826; PF00061-PF00061; PF00061-PF00576; PF00062-PF00062; PF00062-PF00969;
PF00062-PF00993; PF00062-PF02709; PF00062-PF07654; PF00062-PF07686; PF00063-PF00063; PF00063-PF02736;
PF00064-PF00064; PF00064-PF07654; PF00064-PF07686; PF00067-PF00067; PF00067-PF00258; PF00068-PF00068;
PF00069-PF00069; PF00069-PF00134; PF00069-PF00254; PF00069-PF00688; PF00069-PF01111; PF00069-PF01214;
PF00069-PF02827; PF00069-PF02984; PF00069-PF03261; PF00069-PF05350; PF00069-PF05706; PF00070-PF00070;
PF00070-PF00085; PF00071-PF00071; PF00071-PF00102; PF00071-PF00169; PF00071-PF00415; PF00071-PF00454;
PF00071-PF00595; PF00071-PF00621; PF00071-PF00786; PF00071-PF00996; PF00071-PF01833; PF00071-PF02115;
PF00071-PF02185; PF00071-PF02196; PF00071-PF03545; PF00071-PF06456; PF00071-PF07487; PF00072-PF00072;
PF00072-PF00158; PF00072-PF01627; PF00072-PF04344; PF00072-PF07194; PF00073-PF00073; PF00073-PF00084;
PF00073-PF02226; PF00073-PF07654; PF00073-PF07686; PF00074-PF00074; PF00074-PF07686; PF00075-PF00075;
PF00076-PF00076; PF00077-PF00077; PF00077-PF00078; PF00077-PF00540; PF00077-PF06815; PF00078-PF00078;
PF00078-PF06815; PF00079-PF00079; PF00079-PF00089; PF00079-PF01033; PF00080-PF00080; PF00081-PF00081;
PF00082-PF00082; PF00082-PF00280; PF00082-PF00720; PF00082-PF01483; PF00082-PF02428; PF00082-PF05922;
PF00084-PF00084; PF00085-PF00085; PF00085-PF00476; PF00085-PF03372; PF00085-PF07686; PF00086-PF00112;
PF00087-PF00087; PF00087-PF00135; PF00087-PF02931; PF00088-PF00088; PF00089-PF00089; PF00089-PF00228;

PF00089-PF00280; PF00089-PF00299; PF00089-PF01403; PF00089-PF01826; PF00089-PF02177; PF00089-PF02821;
PF00089-PF02822; PF00089-PF03974; PF00089-PF05375; PF00089-PF07648; PF00092-PF00092; PF00092-PF03921;
PF00092-PF07654; PF00094-PF07686; PF00101-PF00101; PF00101-PF02788; PF00102-PF00102; PF00104-PF00104;
PF00105-PF00105; PF00106-PF00106; PF00107-PF00107; PF00108-PF00108; PF00109-PF00109; PF00109-PF02801;
PF00111-PF00111; PF00111-PF00175; PF00112-PF00112; PF00113-PF00113; PF00114-PF00114; PF00115-PF00116;
PF00115-PF00510; PF00115-PF02046; PF00115-PF05392; PF00115-PF06481; PF00115-PF07835; PF00116-PF00116;
PF00116-PF00510; PF00116-PF02297; PF00116-PF02936; PF00116-PF05392; PF00116-PF07686; PF00116-PF07835;
PF00117-PF00117; PF00117-PF00425; PF00117-PF00977; PF00118-PF00118; PF00118-PF00166; PF00119-PF00137;
PF00121-PF00121; PF00122-PF00122; PF00124-PF00124; PF00124-PF01405; PF00124-PF02276; PF00124-PF02419;
PF00125-PF00125; PF00127-PF00127; PF00127-PF02975; PF00127-PF06433; PF00128-PF00128; PF00128-PF00138;
PF00128-PF00197; PF00128-PF07686; PF00129-PF00129; PF00129-PF00798; PF00129-PF00906; PF00129-PF01454;
PF00129-PF04253; PF00129-PF07654; PF00129-PF07686; PF00134-PF02234; PF00135-PF00135; PF00136-PF00969;
PF00136-PF00993; PF00136-PF03104; PF00137-PF00137; PF00138-PF00138; PF00138-PF00139; PF00139-PF00139;
PF00141-PF00141; PF00142-PF00142; PF00142-PF00148; PF00143-PF00143; PF00144-PF00144; PF00144-PF07467;
PF00147-PF00147; PF00149-PF00149; PF00150-PF00150; PF00151-PF00151; PF00152-PF00152; PF00155-PF00155;
PF00156-PF00156; PF00157-PF00157; PF00158-PF00158; PF00160-PF00160; PF00160-PF00540; PF00161-PF00161;
PF00161-PF00652; PF00162-PF00162; PF00163-PF03719; PF00164-PF00575; PF00166-PF00166; PF00167-PF00167;
PF00167-PF07679; PF00169-PF00169; PF00169-PF07714; PF00170-PF00170; PF00170-PF01833; PF00171-PF00171;
PF00173-PF00173; PF00175-PF00175; PF00177-PF00380; PF00178-PF00178; PF00179-PF00179; PF00179-PF00632;
PF00179-PF07834; PF00180-PF00180; PF00183-PF00183; PF00183-PF03234; PF00184-PF00184; PF00185-PF00185;
PF00185-PF02729; PF00185-PF02748; PF00186-PF00186; PF00186-PF00303; PF00187-PF00187; PF00190-PF00190;
PF00191-PF00191; PF00194-PF00194; PF00196-PF00196; PF00197-PF00197; PF00198-PF00198; PF00200-PF00200;
PF00202-PF00202; PF00203-PF00416; PF00204-PF00204; PF00205-PF00205; PF00208-PF00208; PF00210-PF00210;
PF00211-PF00211; PF00211-PF00503; PF00215-PF00215; PF00216-PF00216; PF00217-PF00217; PF00218-PF00218;
PF00221-PF00221; PF00223-PF00223; PF00223-PF01241; PF00223-PF02427; PF00223-PF02507; PF00223-PF02531;
PF00223-PF02605; PF00223-PF07465; PF00224-PF00224; PF00225-PF00225; PF00227-PF00227; PF00227-PF02251;
PF00227-PF02252; PF00227-PF07724; PF00229-PF00229; PF00229-PF07654; PF00231-PF00401; PF00231-PF02823;
PF00231-PF04627; PF00232-PF00232; PF00233-PF00233; PF00234-PF00234; PF00235-PF00235; PF00236-PF00236;
PF00238-PF00297; PF00238-PF01246; PF00239-PF00239; PF00240-PF00240; PF00240-PF02809; PF00240-PF02845;
PF00241-PF00241; PF00243-PF00243; PF00244-PF00244; PF00244-PF00583; PF00245-PF00245; PF00246-PF02244;
PF00248-PF00248; PF00248-PF02214; PF00251-PF00251; PF00253-PF00318; PF00253-PF00338; PF00253-PF00417;
PF00254-PF00254; PF00258-PF00258; PF00261-PF00261; PF00264-PF00264; PF00266-PF00266; PF00267-PF00267;
PF00268-PF00268; PF00270-PF00270; PF00273-PF00273; PF00273-PF01468; PF00274-PF00274; PF00275-PF00275;
PF00276-PF00831; PF00276-PF00832; PF00278-PF00278; PF00280-PF00280; PF00282-PF00282; PF00285-PF00285;
PF00288-PF00288; PF00290-PF00290; PF00290-PF00291; PF00291-PF00291; PF00292-PF00292; PF00293-PF00293;
PF00294-PF00294; PF00295-PF00295; PF00296-PF00296; PF00300-PF00300; PF00301-PF00301; PF00302-PF00302;
PF00302-PF02898; PF00303-PF00303; PF00306-PF00306; PF00306-PF04568; PF00307-PF00307; PF00310-PF00310;
PF00311-PF00311; PF00312-PF01250; PF00313-PF00313; PF00314-PF00314; PF00316-PF00316; PF00317-PF00462;

PF00318-PF00318; PF00318-PF00333; PF00318-PF00338; PF00318-PF00410; PF00325-PF00325; PF00326-PF00326; PF00328-PF00328; PF00330-PF00330; PF00331-PF00331; PF00331-PF00553; PF00331-PF00704; PF00332-PF00332; PF00334-PF00334; PF00337-PF00337; PF00338-PF07650; PF00339-PF02752; PF00340-PF00340; PF00341-PF00341; PF00341-PF07654; PF00343-PF00343; PF00345-PF02753; PF00348-PF00348; PF00350-PF00350; PF00352-PF00352; PF00352-PF02751; PF00352-PF07741; PF00354-PF00354; PF00355-PF07686; PF00359-PF00359; PF00365-PF00365; PF00365-PF02286; PF00368-PF00368; PF00370-PF00370; PF00373-PF00373; PF00374-PF01058; PF00377-PF07654; PF00381-PF00381; PF00381-PF00391; PF00381-PF07475; PF00381-PF07686; PF00383-PF00383; PF00386-PF00386; PF00387-PF00388; PF00390-PF00390; PF00394-PF00394; PF00400-PF00400; PF00400-PF00503; PF00400-PF00631; PF00400-PF02114; PF00400-PF05856; PF00401-PF04627; PF00403-PF00403; PF00405-PF00405; PF00405-PF00969; PF00405-PF00993; PF00405-PF04253; PF00405-PF07686; PF00406-PF00406; PF00407-PF00407; PF00407-PF07654; PF00410-PF00410; PF00410-PF03719; PF00413-PF00413; PF00413-PF00965; PF00415-PF00415; PF00419-PF00419; PF00419-PF02753; PF00421-PF01737; PF00421-PF02419; PF00421-PF02533; PF00421-PF06514; PF00423-PF00423; PF00425-PF00425; PF00426-PF00426; PF00431-PF00431; PF00432-PF00432; PF00435-PF00435; PF00436-PF00436; PF00437-PF00437; PF00438-PF00438; PF00441-PF00441; PF00443-PF00443; PF00445-PF00445; PF00447-PF00447; PF00448-PF00448; PF00448-PF02881; PF00450-PF00450; PF00457-PF00457; PF00457-PF00704; PF00459-PF00459; PF00462-PF00462; PF00463-PF00463; PF00464-PF00464; PF00465-PF00465; PF00469-PF00469; PF00475-PF00475; PF00476-PF00476; PF00476-PF07686; PF00478-PF00478; PF00483-PF00483; PF00484-PF00484; PF00485-PF00485; PF00489-PF00489; PF00490-PF00490; PF00491-PF00491; PF00497-PF00497; PF00501-PF00501; PF00502-PF00502; PF00502-PF01383; PF00502-PF02972; PF00503-PF00503; PF00503-PF00615; PF00503-PF04868; PF00504-PF00504; PF00508-PF00508; PF00508-PF00519; PF00509-PF00509; PF00509-PF00969; PF00509-PF00993; PF00509-PF07654; PF00509-PF07686; PF00510-PF01215; PF00510-PF02046; PF00510-PF02238; PF00510-PF02297; PF00510-PF07835; PF00511-PF00511; PF00512-PF00512; PF00514-PF00514; PF00514-PF03066; PF00514-PF06384; PF00515-PF00515; PF00516-PF00516; PF00516-PF00517; PF00517-PF00517; PF00519-PF00519; PF00520-PF00520; PF00523-PF00523; PF00529-PF00529; PF00531-PF00531; PF00532-PF00532; PF00533-PF00870; PF00533-PF06632; PF00537-PF00537; PF00538-PF00538; PF00540-PF00540; PF00540-PF00969; PF00540-PF00993; PF00541-PF00541; PF00541-PF07686; PF00542-PF00542; PF00543-PF00543; PF00545-PF00545; PF00545-PF01337; PF00548-PF00680; PF00549-PF00549; PF00551-PF00551; PF00554-PF00554; PF00556-PF00556; PF00557-PF00557; PF00560-PF00560; PF00561-PF00561; PF00562-PF01193; PF00565-PF00565; PF00568-PF00568; PF00571-PF00571; PF00574-PF00574; PF00576-PF00576; PF00578-PF00578; PF00579-PF00579; PF00580-PF00580; PF00582-PF00582; PF00583-PF00583; PF00587-PF00587; PF00587-PF03129; PF00588-PF00588; PF00589-PF00589; PF00591-PF00591; PF00594-PF00594; PF00594-PF00713; PF00594-PF01108; PF00594-PF03973; PF00595-PF00595; PF00595-PF00667; PF00596-PF00596; PF00599-PF00599; PF00600-PF00600; PF00601-PF00601; PF00607-PF07686; PF00612-PF01031; PF00615-PF05924; PF00619-PF00619; PF00620-PF00620; PF00623-PF04539; PF00624-PF00624; PF00626-PF00626; PF00632-PF00632; PF00640-PF00640; PF00645-PF00645; PF00646-PF00646; PF00646-PF01466; PF00648-PF00648; PF00650-PF00650; PF00650-PF01105; PF00651-PF00651; PF00653-PF00653; PF00654-PF00654; PF00654-PF07686; PF00656-PF00656; PF00657-PF00657; PF00659-PF00659; PF00668-PF00668; PF00672-PF01036; PF00673-PF00673; PF00676-PF00676; PF00676-PF02780; PF00677-PF00677; PF00680-PF00680; PF00680-PF00910; PF00681-PF00681; PF00682-PF02396; PF00685-PF00685; PF00686-PF00686; PF00686-PF01373; PF00686-PF01833; PF00688-PF00688; PF00688-PF01064; PF00688-PF05806;

PF00689-PF00689; PF00692-PF00692; PF00693-PF00693; PF00696-PF00696; PF00697-PF00697; PF00701-PF00701; PF00702-PF00702; PF00704-PF00704; PF00705-PF02747; PF00708-PF00708; PF00709-PF00709; PF00710-PF00710; PF00712-PF00817; PF00714-PF00714; PF00714-PF07140; PF00715-PF00715; PF00716-PF00716; PF00717-PF00717; PF00718-PF00718; PF00718-PF00761; PF00719-PF00719; PF00722-PF00722; PF00723-PF00723; PF00724-PF00724; PF00725-PF00725; PF00726-PF07654; PF00728-PF00728; PF00729-PF00729; PF00730-PF00730; PF00731-PF00731; PF00733-PF00733; PF00737-PF06596; PF00753-PF00753; PF00754-PF07654; PF00756-PF00756; PF00757-PF00757; PF00757-PF01030; PF00758-PF00758; PF00758-PF07654; PF00758-PF07686; PF00759-PF00759; PF00764-PF00764; PF00766-PF01012; PF00769-PF00769; PF00770-PF02993; PF00772-PF00772; PF00774-PF00774; PF00775-PF00775; PF00787-PF00787; PF00790-PF00790; PF00793-PF00793; PF00797-PF00797; PF00808-PF00808; PF00809-PF00809; PF00816-PF00816; PF00817-PF00817; PF00820-PF07654; PF00825-PF00825; PF00827-PF01248; PF00834-PF00834; PF00840-PF00840; PF00851-PF01577; PF00853-PF00853; PF00853-PF02312; PF00856-PF00856; PF00857-PF00857; PF00861-PF01157; PF00869-PF00869; PF00869-PF01004; PF00870-PF00870; PF00871-PF00871; PF00878-PF00878; PF00879-PF00879; PF00881-PF00881; PF00882-PF00882; PF00882-PF01477; PF00883-PF00883; PF00884-PF00884; PF00885-PF00885; PF00887-PF00887; PF00889-PF00889; PF00889-PF03143; PF00890-PF00890; PF00896-PF00896; PF00897-PF00897; PF00897-PF01700; PF00903-PF00903; PF00905-PF00905; PF00906-PF00906; PF00907-PF00907; PF00908-PF00908; PF00913-PF00913; PF00917-PF00917; PF00918-PF00918; PF00921-PF00921; PF00923-PF00923; PF00926-PF00926; PF00927-PF00927; PF00928-PF01217; PF00928-PF01602; PF00930-PF00930; PF00930-PF00962; PF00940-PF00940; PF00941-PF01799; PF00943-PF01563; PF00944-PF00944; PF00957-PF00957; PF00957-PF01742; PF00957-PF05739; PF00957-PF05835; PF00959-PF00959; PF00960-PF00960; PF00963-PF00963; PF00963-PF02018; PF00964-PF00964; PF00969-PF00969; PF00969-PF00993; PF00969-PF01669; PF00969-PF02072; PF00969-PF02876; PF00969-PF07654; PF00969-PF07686; PF00975-PF00975; PF00975-PF05932; PF00977-PF00977; PF00979-PF00979; PF00979-PF05993; PF00982-PF00982; PF00993-PF00993; PF00993-PF01123; PF00993-PF01669; PF00993-PF02072; PF00993-PF02876; PF00993-PF07654; PF00993-PF07686; PF00994-PF00994; PF00995-PF00995; PF00995-PF05739; PF00996-PF01239; PF01003-PF01003; PF01004-PF01004; PF01007-PF01007; PF01008-PF01008; PF01012-PF02771; PF01014-PF01014; PF01022-PF01022; PF01023-PF01023; PF01025-PF01025; PF01029-PF01029; PF01032-PF01032; PF01036-PF01036; PF01039-PF01039; PF01041-PF01041; PF01042-PF01042; PF01044-PF01044; PF01047-PF01047; PF01048-PF01048; PF01051-PF01051; PF01052-PF01052; PF01053-PF01053; PF01057-PF01057; PF01063-PF01063; PF01068-PF01068; PF01070-PF01070; PF01071-PF01071; PF01074-PF07748; PF01081-PF01081; PF01083-PF01083; PF01094-PF01094; PF01095-PF01095; PF01096-PF04567; PF01099-PF01099; PF01102-PF01102; PF01105-PF01105; PF01108-PF01108; PF01108-PF07654; PF01109-PF01109; PF01110-PF01110; PF01111-PF01111; PF01112-PF01112; PF01115-PF01115; PF01115-PF01267; PF01116-PF01116; PF01117-PF01117; PF01118-PF02774; PF01119-PF01119; PF01121-PF01121; PF01123-PF01123; PF01123-PF02876; PF01126-PF01126; PF01128-PF01128; PF01129-PF01129; PF01131-PF01131; PF01135-PF01135; PF01137-PF01137; PF01138-PF01138; PF01140-PF01140; PF01142-PF01142; PF01144-PF01144; PF01149-PF01149; PF01152-PF01152; PF01154-PF01154; PF01156-PF01156; PF01161-PF01161; PF01166-PF01166; PF01174-PF01174; PF01175-PF01175; PF01177-PF01177; PF01179-PF01179; PF01180-PF01180; PF01182-PF01182; PF01187-PF01187; PF01188-PF01188; PF01191-PF04998; PF01192-PF04560; PF01193-PF01193; PF01193-PF01194; PF01193-PF04983; PF01196-PF01196; PF01202-PF01202; PF01212-PF01212; PF01213-PF01213; PF01214-PF01214; PF01217-PF01602; PF01218-PF01218; PF01220-PF01220; PF01221-PF01221; PF01223-PF01223;

PF01225-PF01225; PF01227-PF01227; PF01229-PF01229; PF01230-PF01230; PF01233-PF01233; PF01238-PF01238; PF01239-PF01239; PF01243-PF01243; PF01248-PF01547; PF01255-PF01255; PF01259-PF01259; PF01262-PF01262; PF01262-PF02233; PF01263-PF01263; PF01264-PF01264; PF01266-PF01266; PF01274-PF01274; PF01278-PF01278; PF01279-PF01279; PF01289-PF01289; PF01293-PF01293; PF01297-PF01297; PF01303-PF01303; PF01306-PF01306; PF01314-PF01314; PF01315-PF01315; PF01315-PF02738; PF01316-PF01316; PF01318-PF01318; PF01320-PF01320; PF01320-PF01844; PF01322-PF01322; PF01323-PF01323; PF01325-PF04023; PF01327-PF01327; PF01329-PF01329; PF01329-PF04814; PF01330-PF01330; PF01333-PF02529; PF01333-PF03742; PF01333-PF05115; PF01336-PF01336; PF01337-PF01337; PF01339-PF01339; PF01340-PF01340; PF01341-PF01341; PF01342-PF01342; PF01346-PF01346; PF01351-PF01351; PF01353-PF01353; PF01355-PF01355; PF01361-PF01361; PF01363-PF01363; PF01367-PF01367; PF01368-PF01368; PF01369-PF01369; PF01370-PF01370; PF01371-PF01371; PF01373-PF01373; PF01374-PF01374; PF01375-PF01375; PF01375-PF01376; PF01375-PF06453; PF01376-PF01376; PF01380-PF01380; PF01382-PF01382; PF01391-PF01391; PF01392-PF01392; PF01393-PF01393; PF01394-PF01394; PF01395-PF01395; PF01401-PF01401; PF01402-PF01402; PF01403-PF01403; PF01404-PF01404; PF01405-PF02419; PF01406-PF01406; PF01411-PF01411; PF01419-PF01419; PF01421-PF01421; PF01423-PF01423; PF01425-PF01425; PF01426-PF01426; PF01427-PF01427; PF01441-PF01441; PF01442-PF01442; PF01447-PF01447; PF01450-PF01450; PF01451-PF01451; PF01452-PF01452; PF01453-PF01453; PF01458-PF01458; PF01465-PF01465; PF01467-PF01467; PF01470-PF01470; PF01473-PF01473; PF01477-PF01477; PF01477-PF02740; PF01481-PF01481; PF01483-PF01483; PF01488-PF01488; PF01497-PF01497; PF01501-PF01501; PF01515-PF01515; PF01517-PF01517; PF01521-PF01521; PF01532-PF01532; PF01536-PF01536; PF01537-PF01537; PF01539-PF01539; PF01539-PF01542; PF01539-PF01560; PF01542-PF01542; PF01542-PF01543; PF01543-PF01543; PF01543-PF07654; PF01546-PF01546; PF01546-PF07687; PF01547-PF01547; PF01555-PF01555; PF01557-PF01557; PF01560-PF01560; PF01564-PF01564; PF01568-PF03892; PF01568-PF05738; PF01569-PF01569; PF01577-PF01577; PF01582-PF01582; PF01583-PF01583; PF01590-PF01590; PF01591-PF01591; PF01593-PF01593; PF01597-PF01597; PF01601-PF01601; PF01608-PF01608; PF01611-PF01611; PF01613-PF01613; PF01614-PF01614; PF01625-PF01625; PF01627-PF01627; PF01633-PF01633; PF01634-PF01634; PF01636-PF01636; PF01637-PF01637; PF01641-PF01641; PF01642-PF01642; PF01652-PF01652; PF01652-PF05456; PF01653-PF01653; PF01658-PF01658; PF01663-PF01663; PF01664-PF01664; PF01665-PF01665; PF01670-PF01670; PF01700-PF01700; PF01702-PF01702; PF01704-PF01704; PF01706-PF01706; PF01712-PF01712; PF01725-PF01725; PF01729-PF01729; PF01729-PF02749; PF01740-PF01740; PF01740-PF02518; PF01741-PF01741; PF01742-PF01742; PF01747-PF01747; PF01750-PF01750; PF01751-PF01751; PF01756-PF01756; PF01761-PF01761; PF01764-PF01764; PF01780-PF03947; PF01784-PF01784; PF01791-PF01791; PF01799-PF03450; PF01804-PF01804; PF01807-PF01807; PF01812-PF01812; PF01814-PF01814; PF01815-PF01815; PF01817-PF01817; PF01818-PF01818; PF01819-PF01819; PF01829-PF01829; PF01833-PF01833; PF01835-PF01835; PF01842-PF01842; PF01844-PF01844; PF01845-PF01845; PF01857-PF01858; PF01858-PF01858; PF01862-PF01862; PF01869-PF01869; PF01870-PF01870; PF01880-PF01880; PF01884-PF01884; PF01892-PF01892; PF01896-PF01896; PF01910-PF01910; PF01920-PF02996; PF01921-PF01921; PF01948-PF01948; PF01965-PF01965; PF01974-PF01974; PF01978-PF01978; PF01979-PF01979; PF01992-PF01992; PF01993-PF01993; PF01997-PF01997; PF02011-PF02011; PF02017-PF02017; PF02018-PF02018; PF02025-PF02025; PF02035-PF02035; PF02041-PF02041; PF02056-PF02056; PF02074-PF02074; PF02075-PF02075; PF02080-PF02080; PF02081-PF02081; PF02085-PF02085; PF02091-PF02091; PF02098-PF02098; PF02110-PF02110; PF02115-PF02115; PF02121-PF02121; PF02126-PF02126;

PF02129-PF02129; PF02136-PF02136; PF02138-PF02138; PF02142-PF02142; PF02146-PF02146; PF02151-PF02151; PF02152-PF02152; PF02156-PF02156; PF02157-PF02157; PF02159-PF02159; PF02167-PF02167; PF02167-PF02320; PF02167-PF05365; PF02172-PF02173; PF02181-PF02181; PF02189-PF07654; PF02197-PF02197; PF02198-PF02198; PF02210-PF02210; PF02211-PF02211; PF02211-PF02979; PF02214-PF02214; PF02216-PF02216; PF02219-PF02219; PF02221-PF02221; PF02223-PF02223; PF02226-PF02226; PF02230-PF02230; PF02231-PF02811; PF02233-PF02233; PF02239-PF02239; PF02240-PF02249; PF02240-PF02745; PF02241-PF02241; PF02241-PF02249; PF02244-PF02244; PF02249-PF02249; PF02250-PF02250; PF02251-PF02251; PF02251-PF02252; PF02252-PF02252; PF02253-PF02253; PF02254-PF02254; PF02254-PF07885; PF02255-PF02255; PF02258-PF02258; PF02261-PF02261; PF02264-PF02264; PF02267-PF02267; PF02270-PF02270; PF02270-PF05793; PF02271-PF02939; PF02273-PF02273; PF02274-PF02274; PF02275-PF02275; PF02282-PF02282; PF02283-PF02283; PF02284-PF02790; PF02284-PF02936; PF02284-PF02937; PF02285-PF02935; PF02285-PF02936; PF02286-PF02286; PF02291-PF02969; PF02297-PF02297; PF02298-PF02298; PF02300-PF02313; PF02301-PF05557; PF02303-PF02303; PF02304-PF02305; PF02305-PF02925; PF02305-PF04726; PF02306-PF02925; PF02310-PF06368; PF02313-PF02313; PF02319-PF02319; PF02321-PF02321; PF02329-PF02329; PF02331-PF02331; PF02332-PF02332; PF02334-PF02334; PF02335-PF02335; PF02348-PF02348; PF02350-PF02350; PF02353-PF02353; PF02359-PF02933; PF02365-PF02365; PF02378-PF02378; PF02391-PF02391; PF02396-PF07836; PF02404-PF02404; PF02419-PF05151; PF02423-PF02423; PF02427-PF02507; PF02428-PF02428; PF02429-PF02429; PF02431-PF02431; PF02437-PF02437; PF02441-PF02441; PF02452-PF02452; PF02463-PF02463; PF02502-PF02502; PF02504-PF02504; PF02511-PF02511; PF02518-PF02518; PF02525-PF02525; PF02529-PF03742; PF02531-PF02605; PF02542-PF02542; PF02545-PF02545; PF02548-PF02548; PF02556-PF02556; PF02556-PF07517; PF02560-PF02560; PF02561-PF02561; PF02566-PF02566; PF02567-PF02567; PF02569-PF02569; PF02570-PF02570; PF02574-PF02574; PF02575-PF02575; PF02577-PF02577; PF02578-PF02578; PF02580-PF02580; PF02581-PF02581; PF02585-PF02585; PF02590-PF02590; PF02595-PF02595; PF02597-PF05237; PF02602-PF02602; PF02614-PF02614; PF02615-PF02615; PF02617-PF02861; PF02627-PF02627; PF02657-PF02657; PF02664-PF02664; PF02668-PF02668; PF02670-PF02670; PF02676-PF02676; PF02700-PF02700; PF02709-PF02709; PF02710-PF02710; PF02710-PF03996; PF02716-PF02716; PF02730-PF02730; PF02734-PF02734; PF02737-PF02737; PF02738-PF02738; PF02738-PF03450; PF02741-PF02741; PF02742-PF02742; PF02744-PF02744; PF02746-PF02746; PF02747-PF02747; PF02750-PF02750; PF02751-PF03153; PF02752-PF02752; PF02753-PF02753; PF02762-PF02762; PF02765-PF02765; PF02768-PF02768; PF02768-PF06144; PF02773-PF02773; PF02774-PF02774; PF02776-PF02776; PF02777-PF02777; PF02780-PF02780; PF02781-PF02781; PF02784-PF02784; PF02785-PF02785; PF02787-PF02787; PF02792-PF02792; PF02797-PF02797; PF02799-PF02799; PF02800-PF02800; PF02801-PF02801; PF02803-PF02803; PF02806-PF02806; PF02812-PF02812; PF02816-PF02816; PF02817-PF02852; PF02821-PF02821; PF02823-PF04627; PF02826-PF02826; PF02833-PF02833; PF02836-PF02836; PF02839-PF02839; PF02840-PF02840; PF02842-PF02842; PF02844-PF02844; PF02845-PF02845; PF02852-PF02852; PF02854-PF03467; PF02861-PF02861; PF02863-PF02863; PF02866-PF02866; PF02867-PF02867; PF02876-PF02876; PF02876-PF07654; PF02876-PF07686; PF02879-PF02879; PF02880-PF02880; PF02882-PF02882; PF02883-PF02883; PF02887-PF02887; PF02894-PF02894; PF02898-PF02898; PF02900-PF02900; PF02901-PF02901; PF02903-PF02903; PF02906-PF02906; PF02909-PF02909; PF02911-PF02911; PF02915-PF02915; PF02917-PF02918; PF02921-PF02939; PF02921-PF05193; PF02921-PF05365; PF02923-PF02923; PF02924-PF02924; PF02925-PF02925; PF02930-PF02930; PF02931-PF02931; PF02932-PF02932; PF02936-PF02937; PF02936-PF05392; PF02940-PF02940; PF02941-PF02943;

PF02947-PF02947; PF02954-PF02954; PF02970-PF02970; PF02975-PF06433; PF02982-PF02982; PF02983-PF02983; PF02985-PF02985; PF02991-PF02991; PF03009-PF03009; PF03033-PF03033; PF03051-PF03051; PF03061-PF03061; PF03062-PF03062; PF03063-PF03063; PF03066-PF03066; PF03070-PF03070; PF03084-PF06016; PF03091-PF03091; PF03098-PF03098; PF03099-PF03099; PF03122-PF03122; PF03123-PF03123; PF03129-PF03129; PF03143-PF03143; PF03143-PF03144; PF03144-PF03144; PF03145-PF03145; PF03166-PF03166; PF03167-PF03167; PF03205-PF03205; PF03220-PF03220; PF03256-PF03256; PF03259-PF03259; PF03275-PF03275; PF03328-PF03328; PF03331-PF03331; PF03349-PF03349; PF03358-PF03358; PF03372-PF03372; PF03388-PF03388; PF03392-PF03392; PF03404-PF03404; PF03405-PF03405; PF03411-PF03411; PF03414-PF03414; PF03435-PF03435; PF03441-PF03441; PF03453-PF03453; PF03459-PF03459; PF03466-PF03466; PF03475-PF03475; PF03513-PF03513; PF03515-PF03515; PF03519-PF03519; PF03550-PF03550; PF03561-PF03561; PF03572-PF03572; PF03576-PF03576; PF03623-PF03623; PF03632-PF03632; PF03641-PF03641; PF03652-PF03652; PF03720-PF03720; PF03725-PF03725; PF03727-PF03727; PF03730-PF03730; PF03730-PF03731; PF03737-PF03737; PF03740-PF03740; PF03767-PF03767; PF03775-PF05209; PF03776-PF03776; PF03786-PF03786; PF03795-PF03795; PF03830-PF03830; PF03846-PF03846; PF03861-PF03861; PF03869-PF03869; PF03870-PF04998; PF03892-PF03892; PF03892-PF04879; PF03902-PF03902; PF03919-PF03919; PF03925-PF03925; PF03936-PF03936; PF03938-PF03938; PF03949-PF03949; PF03965-PF03965; PF03971-PF03971; PF03972-PF03972; PF03974-PF03974; PF03992-PF03992; PF03996-PF03996; PF04014-PF04014; PF04014-PF07686; PF04030-PF04030; PF04043-PF04043; PF04045-PF05856; PF04072-PF04072; PF04074-PF04074; PF04092-PF04092; PF04095-PF04095; PF04101-PF04101; PF04127-PF04127; PF04166-PF04166; PF04218-PF04218; PF04227-PF04227; PF04231-PF04231; PF04234-PF04234; PF04253-PF04253; PF04309-PF04309; PF04345-PF04345; PF04349-PF04349; PF04379-PF04379; PF04386-PF04386; PF04423-PF04423; PF04451-PF04451; PF04461-PF04461; PF04485-PF04485; PF04509-PF04509; PF04514-PF04514; PF04539-PF04539; PF04539-PF04545; PF04539-PF04560; PF04542-PF04542; PF04548-PF04548; PF04568-PF04568; PF04612-PF04612; PF04616-PF04616; PF04619-PF04619; PF04650-PF04650; PF04692-PF04692; PF04699-PF05856; PF04711-PF04711; PF04728-PF04728; PF04811-PF04811; PF04814-PF04814; PF04851-PF04851; PF04908-PF04908; PF04928-PF04928; PF04951-PF04951; PF04960-PF04960; PF04961-PF04961; PF05008-PF05739; PF05025-PF05025; PF05067-PF05067; PF05076-PF05076; PF05138-PF05138; PF05164-PF05164; PF05168-PF05168; PF05173-PF05173; PF05184-PF05184; PF05193-PF05193; PF05194-PF05194; PF05199-PF05199; PF05202-PF05202; PF05221-PF05221; PF05224-PF05224; PF05239-PF05239; PF05247-PF05247; PF05312-PF05312; PF05352-PF05352; PF05357-PF05357; PF05362-PF05362; PF05367-PF05367; PF05368-PF05368; PF05408-PF05408; PF05470-PF05470; PF05557-PF05557; PF05652-PF05652; PF05697-PF05697; PF05738-PF05738; PF05739-PF05739; PF05772-PF05772; PF05790-PF05790; PF05866-PF05866; PF05932-PF05932; PF05993-PF05993; PF06026-PF06026; PF06083-PF06083; PF06087-PF06087; PF06134-PF06134; PF06184-PF06184; PF06330-PF06330; PF06368-PF06368; PF06369-PF06369; PF06397-PF06397; PF06399-PF06399; PF06406-PF06406; PF06411-PF06411; PF06418-PF06418; PF06426-PF06426; PF06431-PF06431; PF06433-PF06433; PF06438-PF06438; PF06453-PF06453; PF06456-PF06456; PF06457-PF06457; PF06470-PF06470; PF06560-PF06560; PF06613-PF06613; PF06628-PF06628; PF06632-PF06632; PF06652-PF06652; PF06689-PF06689; PF06766-PF06766; PF06815-PF06815; PF06815-PF06817; PF06817-PF06817; PF06954-PF06954; PF06968-PF06968; PF07016-PF07016; PF07124-PF07124; PF07140-PF07140; PF07140-PF07654; PF07361-PF07361; PF07412-PF07412; PF07447-PF07447; PF07471-PF07471; PF07474-PF07686; PF07475-PF07475; PF07476-PF07476; PF07478-PF07478; PF07516-PF07516; PF07645-PF07645; PF07651-PF07651; PF07654-PF07654; PF07654-PF07686;

PF07654-PF07885; PF07676-PF07676; PF07678-PF07678; PF07679-PF07679; PF07686-PF07686; PF07687-PF07687; PF07688-PF07688; PF07700-PF07700; PF07710-PF07710; PF07714-PF07714; PF07722-PF07722; PF07731-PF07731; PF07736-PF07736; PF07748-PF07748; PF07823-PF07823; PF07827-PF07827; PF07828-PF07828; PF07832-PF07832; PF07834-PF07834; PF07836-PF07836; PF07840-PF07840; PF07858-PF07858; PF07859-PF07859; PF07880-PF07880; PF07912-PF07912; PF07917-PF07917; PF07938-PF07938; PF07943-PF07943;

Table 3. The keywords list (887 keywords) we used in our methods. The keyword under line means that it is repeated in spkw2go.

2Fe-2S; 3D-structure; 3Fe-4S; 4Fe-4S; Acetoin biosynthesis; Acetoin catabolism; Acetylation; Acetylcholine receptor inhibitor; Actin capping; Actin-binding; Activator; Acute phase; Acyltransferase; ADP-ribosylation; AIDS; Albinism; Alginate biosynthesis; Alkaloid metabolism; Alkylphosphonate uptake; Allergen; Allosteric enzyme; Alpha-amylase inhibitor; Alport syndrome; Alternative initiation; Alternative promoter usage; Alternative splicing; Alzheimer's disease; Amidation; Amino-acid biosynthesis; Amino-acid transport; Aminoacyl-tRNA synthetase; Aminopeptidase; Aminotransferase; Amphibian defense peptide; Amyloid; Amyloplast; Amyotrophic lateral sclerosis; Angiogenesis; Anhidrotic ectodermal dysplasia; Anion exchange; ANK repeat; Annexin; Antenna complex; Antibiotic; Antibiotic biosynthesis; Antibiotic resistance; Antifreeze protein; Antigen; Anti-oncogene; Antioxidant; Antiport; Antiviral; Apoplast; Apoptosis; Arabinose catabolism; Arginine biosynthesis; Arginine metabolism; Aromatic amino acid biosynthesis; Aromatic hydrocarbons catabolism; Arsenical resistance; Ascorbate biosynthesis; Asparagine biosynthesis; Aspartic protease inhibitor; Aspartyl esterase; Aspartyl protease; Atherosclerosis; ATP synthesis; ATP-binding; Autocatalytic cleavage; Autoimmune encephalomyelitis; Autoimmune uveitis; Autoinducer synthesis; Autophagy; Auxin biosynthesis; Bacterial capsule; Bacteriochlorophyll; Bacteriochlorophyll biosynthesis; Bacteriocin; Bacteriocin biosynthesis; Bacteriocin immunity; Bacteriocin transport; Bacteriolytic enzyme; Bait region; Bardet-Biedl syndrome; Bartter syndrome; Basement membrane; B-cell activation; Behavior; Bence-Jones protein; Bernard Soulier syndrome; Bile acid catabolism; Bile pigment; Biological rhythms; Biomineralization; Biotin; Biotin biosynthesis; Blood coagulation; Blood group antigen; Bombesin family; Bradykinin; Branched-chain amino acid biosynthesis; Branched-chain amino acid catabolism; Bromination; Bromodomain; Cadmium; Cadmium resistance; Calcium; Calcium channel; Calcium channel inhibitor; Calcium transport; Calcium/phospholipid-binding; Calcium-binding; Calmodulin-binding; Calvin cycle; cAMP; cAMP biosynthesis; cAMP-binding; Capsid assembly; Carbohydrate metabolism; Carbon dioxide fixation; Carboxypeptidase; Cardiomyopathy; Cardiotoxin; Carnitine biosynthesis; Carotenoid biosynthesis; Cataract; Catecholamine biosynthesis; Catecholamine metabolism; CBS domain; Cell adhesion; Cell cycle; Cell division; Cell shape; Cell wall; Cellulose biosynthesis; Cellulose degradation; Centromere; CF(0); CF(1); cGMP; cGMP biosynthesis; cGMP-binding; Chaperone; Charcot-Marie-Tooth disease; Chemotaxis; Chitin degradation; Chitin-binding; Chloride; Chloride channel; Chloride channel inhibitor; Chlorophyll; Chlorophyll biosynthesis; Chlorophyll catabolism; Chloroplast; Chlorosome; Cholesterol biosynthesis; Cholesterol metabolism; Chondrogenesis; Chorion; Chromate resistance; Chromatin regulator; Chromophore; Chromosomal protein; Chromosomal translocation; Chromosome partition; Chronic granulomatous disease; Chylomicron; Citrate utilization; Citrullination; Cleavage on pair of basic residues; Cnidocyst; Coat protein; Coated pits; Cobalamin biosynthesis; Cobalt; Cobalt transport; Cockayne's syndrome; Coenzyme A biosynthesis; Coenzyme M biosynthesis; Coiled coil; Collagen; Collagen degradation; Competence; Complement alternate pathway; Complement pathway; Complete proteome; Cone-rod dystrophy; Congenital disorder of glycosylation; Conidiation; Conjugation; Copper; Copper transport; Copulatory plug; Core protein; Covalent protein-DNA linkage; Covalent protein-RNA linkage; Crown gall tumor; CTQ; Cuticle; Cyanelle; Cyclin; Cycloheximide resistance; Cyclosporin; Cysteine biosynthesis;

Cystinuria; Cytadherence; Cytochrome c-type biogenesis; Cytokine; Cytokinin biosynthesis; Cytolysis; Cytosine metabolism; Cytoskeleton; D-amino acid; Deafness; Decarboxylase; Defensin; Dehydrin; Dejerine-Sottas syndrome; Dental caries; Deoxyribonucleotide synthesis; Detoxification; Developmental protein; Diabetes insipidus; Diabetes mellitus; Diaminopimelate biosynthesis; Differentiation; Digestion; Dioxygenase; Dipeptidase; Direct protein sequencing; Disease mutation; DNA condensation; DNA damage; DNA excision; DNA integration; DNA invertase; DNA packaging; DNA priming; DNA recombination; DNA repair; DNA replication; DNA replication inhibitor; DNA synthesis; DNA-binding; DNA-directed DNA polymerase; DNA-directed RNA polymerase; Down's syndrome; Dwarfism; Dynein; Early protein; EGF-like domain; Ehlers-Danlos syndrome; Electron transport; Elliptocytosis; Elongation factor; Embryo; Endocytosis; Endonuclease; Endoplasmic reticulum; Endorphin; Enterobactin biosynthesis; Enterotoxin; Envelope protein; Epidermolysis bullosa; Epilepsy; ERV; Erythrocyte; Erythrocyte maturation; Ethylene biosynthesis; Excision nuclease; Exocytosis; Exonuclease; Exopolysaccharide synthesis; Exosome; Extracellular matrix; Eye lens protein; FAD; Fatty acid biosynthesis; Fatty acid metabolism; Feather; Fertilization; Fiber protein; Fibrinolysis; Fimbria; Flagellar rotation; Flagellum; Flavonoid biosynthesis; Flavoprotein; Flight; Flowering; FMN; Folate biosynthesis; Folate-binding; Formylation; Fruit ripening; Fruiting body; Fucose metabolism; Fungicide; Fusion protein; Galactitol metabolism; Galactose metabolism; Galectin; Gamma-carboxyglutamic acid; Gap junction; Gap protein; Gas vesicle; Gaseous exchange; Gastrulation; Gaucher disease; Genetically modified food; Germination; Glucagon family; Gluconate utilization; Gluconeogenesis; Glucose metabolism; Glutamate biosynthesis; Glutamine amidotransferase; Glutaricaciduria; Glutathione biosynthesis; Glutathionylation; Glycerol metabolism; Glycogen biosynthesis; Glycogen metabolism; Glycogen storage disease; Glycolate pathway; Glycolysis; Glycoprotein; Glycosidase; Glycosome; Glycosyltransferase; Glyoxylate bypass; Glyoxysome; GM2-gangliosidosis; GMP biosynthesis; Golgi stack; Gonadal differentiation; Gout; GPI-anchor; GPI-anchor biosynthesis; G-protein coupled receptor; Growth arrest; Growth factor; Growth factor binding; Growth regulation; GTPase activation; GTP-binding; Guanine-nucleotide releasing factor; HDL; Heat shock; Helicase; Hemagglutinin; Heme; Heme biosynthesis; Hemoglobin-binding; Hemolymph; Hemolymph clotting; Hemolysis; Hemophilia; Hemostasis; Heparan sulfate; Heparin-binding; Herbicide resistance; Hereditary hemolytic anemia; Hereditary multiple exostoses; Hereditary nonpolyposis colorectal cancer; Hereditary spastic paraplegia; Hermansky-Pudlak syndrome; Heterocyst; Hexon protein; Hexon-associated protein; Hibernation; Hirschsprung disease; Histidine biosynthesis; Histidine metabolism; Holoprosencephaly; Homeobox; Hormone; Hyaluronic acid; Hybridoma; Hydrogen ion transport; Hydrogen peroxide; Hydrogenosome; Hydrolase; Hydroxylation; Hyperlipidemia; Hypersensitive response; Hypotensive agent; Hypothalamus; Hypothetical protein; Hypusine; Hypusine biosynthesis; Ice nucleation; IgA-binding protein; IgE-binding protein; IgG-binding protein; Immune response; Immunoglobulin C region; Immunoglobulin domain; Immunoglobulin V region; Inflammatory response; Initiation factor; Inner membrane; Inositol biosynthesis; Insect immunity; Insulin family; Integrin; Interferon induction; Intermediate filament; Intron homing; Iodination; Ion transport; Ionic channel; Ionic channel inhibitor; Iron; Iron storage; Iron transport; Iron-sulfur; Isoleucine biosynthesis; Isomerase; Isoprene biosynthesis; Karyogamy; Kelch repeat; Keratin; Keratinization; Kinase; Kinetoplast; Kringle; Lactation; Lactose biosynthesis; Lactose metabolism; Laminin EGF-like domain; Lantibiotic; Late protein; LDL; Leader peptide; Leber congenital amaurosis; Leber hereditary optic neuropathy; Lectin; Leigh syndrome; Leucine biosynthesis; Leucine-rich repeat;

Leukotriene biosynthesis; Li-Fraumeni syndrome; Ligase; Light-harvesting polypeptide; Lignin biosynthesis; Lignin degradation; LIM domain; Lipid A biosynthesis; Lipid degradation; Lipid metabolism; Lipid synthesis; Lipid transport; Lipid-binding; Lipocalin; Lipopolysaccharide biosynthesis; Lipoprotein; Lipoyl; Lissencephaly; Lithium; Long QT syndrome; LTQ; Luminescence; Lyase; Lysine biosynthesis; Lysosome; Macrophage; Magnesium; Malaria; Maltose metabolism; Mandelate pathway; Manganese; Mannose-binding; Maple syrup urine disease; Mast cell degranulation; Matrix protein; Meiosis; Melanin biosynthesis; MELAS syndrome; Melatonin biosynthesis; Membrane; Membrane attack complex; Menaquinone biosynthesis; Mercuric resistance; Mercury; Merozoite; Metachromatic leukodystrophy; Metal-binding; Metalloenzyme inhibitor; Metalloprotease; Metalloprotease inhibitor; Metal-thiolate cluster; Methanogenesis; Methanol utilization; Methionine biosynthesis; Methotrexate resistance; Methylation; Methyltransferase; MHC I; MHC II; Microsome; Microtubule; Milk; Mineral balance; Mitochondrion; Mitogen; Mitosis; Mobility protein; Molybdenum; Molybdenum cofactor biosynthesis; Monoclonal antibody; Monooxygenase; Morphogen; Motor protein; mRNA capping; mRNA processing; mRNA splicing; mRNA transport; Mucopolysaccharidosis; Multifunctional enzyme; Multigene family; Muscle protein; Mutator protein; Myelin; Myogenesis; Myosin; Myristate; NAD; NADP; Neurodegeneration; Neurogenesis; Neuronal ceroid lipofuscinosis; Neuropeptide; Neurotoxin; Neurotransmitter; Neurotransmitter biosynthesis; Neurotransmitter degradation; Neurotransmitter transport; Nickel; Nitrate assimilation; Nitrogen fixation; Nodulation; Noncapsid protein; Nonsense-mediated mRNA decay; Nonstructural protein; Nuclear pore complex; Nuclear protein; Nuclease; Nucleocapsid; Nucleomorph; Nucleoprotein; Nucleosome core; Nucleotide biosynthesis; Nucleotide metabolism; Nucleotide-binding; Nucleotidyltransferase; Nylon degradation; Obesity; Oil body; Olfaction; Oncogene; One-carbon metabolism; Oogenesis; Opioid peptide; Organic radical; Osteogenesis; Outer membrane; Oxidation; Oxidoreductase; Oxygen transport; Paired box; Pair-rule protein; Palmitate; Pancreas; Pantothenate biosynthesis; Parkinsonism; Parkinson's disease; Pathogenesis-related protein; Pentaxin; Pentose shunt; Peptide transport; Peptidoglycan synthesis; Peptidoglycan-anchor; Periplasmic; Peroxidase; Peroxisome; PHA biosynthesis; Phage lysis protein; Phage maturation; Phage recognition; Phagocytosis; Pharmaceutical; PHB biosynthesis; Phenylalanine biosynthesis; Phenylalanine catabolism; Phenylketonuria; Phenylpropanoid metabolism; Pheromone; Pheromone response; Pheromone-binding; Phorbol-ester binding; Phosphate transport; Phospholipase A2 inhibitor; Phospholipid biosynthesis; Phospholipid degradation; Phosphopantetheine; Phosphorylation; Phosphotransferase system; Photoprotein; Photoreceptor; Photorespiration; Photosynthesis; Photosystem I; Photosystem II; Phycobilisome; Phytochrome; Phytochrome signaling pathway; Pigment; Pituitary; Placenta; Plant defense; Plant toxin; Plasma; Plasmid; Plasmid copy control; Plasmid partition; Plasminogen activation; Plastoquinone; Platelet; Polyamine biosynthesis; Polymorphism; Polyneuropathy; Polyprotein; Polysaccharide degradation; Polysaccharide transport; Porin; Porphyrin biosynthesis; Postsynaptic membrane; Postsynaptic neurotoxin; Potassium; Potassium channel; Potassium channel inhibitor; Potassium transport; PQQ; PQQ biosynthesis; Pregnancy; Prenylation; Prenyltransferase; Presynaptic neurotoxin; Primary microcephaly; Primosome; Prion; Proline biosynthesis; Proline metabolism; Prostaglandin biosynthesis; Prostaglandin metabolism; Protease; Protease inhibitor; Proteasome; Protein biosynthesis; Protein kinase inhibitor; Protein phosphatase; Protein phosphatase inhibitor; Protein splicing; Protein synthesis inhibitor; Protein transport; Proteoglycan; Prothrombin activator; Proto-oncogene; Pseudohermaphroditism; Purine biosynthesis; Purine metabolism; Purine salvage; Putrescine biosynthesis; Pyridine nucleotide biosynthesis; Pyridoxal phosphate; Pyridoxine biosynthesis; Pyrimidine biosynthesis; Pyrogen; Pyrokinin; Pyropoikilocytosis; Pyrrolidone carboxylic acid; Pyrrolysine; Pyruvate; Queuosine biosynthesis; Quinate metabolism; Quinone; Quorum sensing; Reaction center; Receptor; Redox-active center; Repeat; Repressor; Respiratory chain; Restriction system; Retinal protein; Retinitis pigmentosa; Retinol-binding; Rhamnose metabolism; Rhizomelic chondrodysplasia punctata; Riboflavin biosynthesis; Ribonucleoprotein; Ribosomal frameshift; Ribosomal protein; Ribosome biogenesis; RNA editing; RNA repair; RNA replication; RNA-binding; RNA-directed DNA polymerase; RNA-directed RNA polymerase; Rotamase; rRNA processing; rRNA-binding; Rubredoxin; Sarcoplasmic reticulum; Schiff base; SCID; Seed; Seed embryo; Seed storage protein; Segmentation polarity protein; Selectin; Selenium; Selenocysteine; Self-incompatibility; Seminal vesicle; Sensory transduction; Septation; Serine biosynthesis; Serine esterase; Serine protease; Serine protease homolog; Serine protease inhibitor; Serine/threonine-protein kinase; Serotonin biosynthesis; Serpin; Sexual differentiation; SH2 domain; SH3 domain; SH3-binding; Sialic acid; Sigma factor; Signal; Signal recognition particle; Signal transduction inhibitor; Signal-anchor; Signalosome; Silk; S-layer; S-nitrosylation; Sodium channel; Sodium channel inhibitor; Sodium transport; Sodium/potassium transport; SOS mutagenesis; SOS response; Sperm; Spermatogenesis; Spermidine biosynthesis; Sphingolipid metabolism; Spliceosome; Sporozoite; Sporulation; Starch biosynthesis; Stargardt disease; Steroid biosynthesis; Steroid metabolism; Steroid-binding; Steroidogenesis; Sterol biosynthesis; Stickler syndrome; Storage protein; Streptomycin biosynthesis; Structural protein; Submandibular gland; Sugar transport; Sulfate respiration; Sulfate transport; Sulfation; Superantigen; Surface film; Sushi; Symport; Synapse; Synaptosome; Systemic lupus erythematosus; Tachykinin; Taste-modifying protein; Taxol biosynthesis; T-cell; Teichoic acid biosynthesis; Tellurium resistance; Telomere; Terminal addition; Testis; Testosterone; Tetrahydrobiopterin biosynthesis; Thiamine biosynthesis; Thiamine catabolism; Thiamine pyrophosphate; Thick filament; Thioester bond; Thioether bond; Thiol protease; Thiol protease inhibitor; Thionin; Threonine biosynthesis; Threonine protease; Thrombophilia; Thylakoid; Thymus; Thyroid hormone; Thyroid hormones biosynthesis; Tight junction; Tissue remodeling; TonB box; Topoisomerase; Toxin; TPQ; TPR repeat; Trans-acting factor; Transcription; Transcription antitermination; Transcription regulation; Transcription termination; Transducer; Transferase; Transit peptide; Translation regulation; Translocation; Transmembrane; Transport; Transposable element; Transposition; Tricarboxylic acid cycle; Trimethoprim resistance; Triplet repeat expansion; tRNA processing; tRNA-binding; Trypanosomiasis; Tryptophan biosynthesis; Tryptophan catabolism; TTQ; Tumor antigen; Tungsten; Two-component regulatory system; Tyrosine biosynthesis; Tyrosine catabolism; Tyrosine-protein kinase; Ubiquinone; Ubiquinone biosynthesis; Ubl conjugation; Ubl conjugation pathway; Unfolded protein response; Urea cycle; Usher syndrome; Vanadium; Vasoactive; Vasoconstrictor; Vasodilator; Viral immunoevasion; Viral occlusion body; Virulence; Vision; Vitamin A; Vitamin B12; Vitamin C; Vitamin D; Vitamin K; VLDL; Voltage-gated channel; von Willebrand disease; Waardenburg syndrome; WD repeat; Whey; Whooping cough; Williams-Beuren syndrome; Wnt signaling pathway; Xeroderma pigmentosum; Xylan degradation; Xylose metabolism; Zellweger syndrome; Zinc; Zinc transport; Zinc-finger; Zymogen;

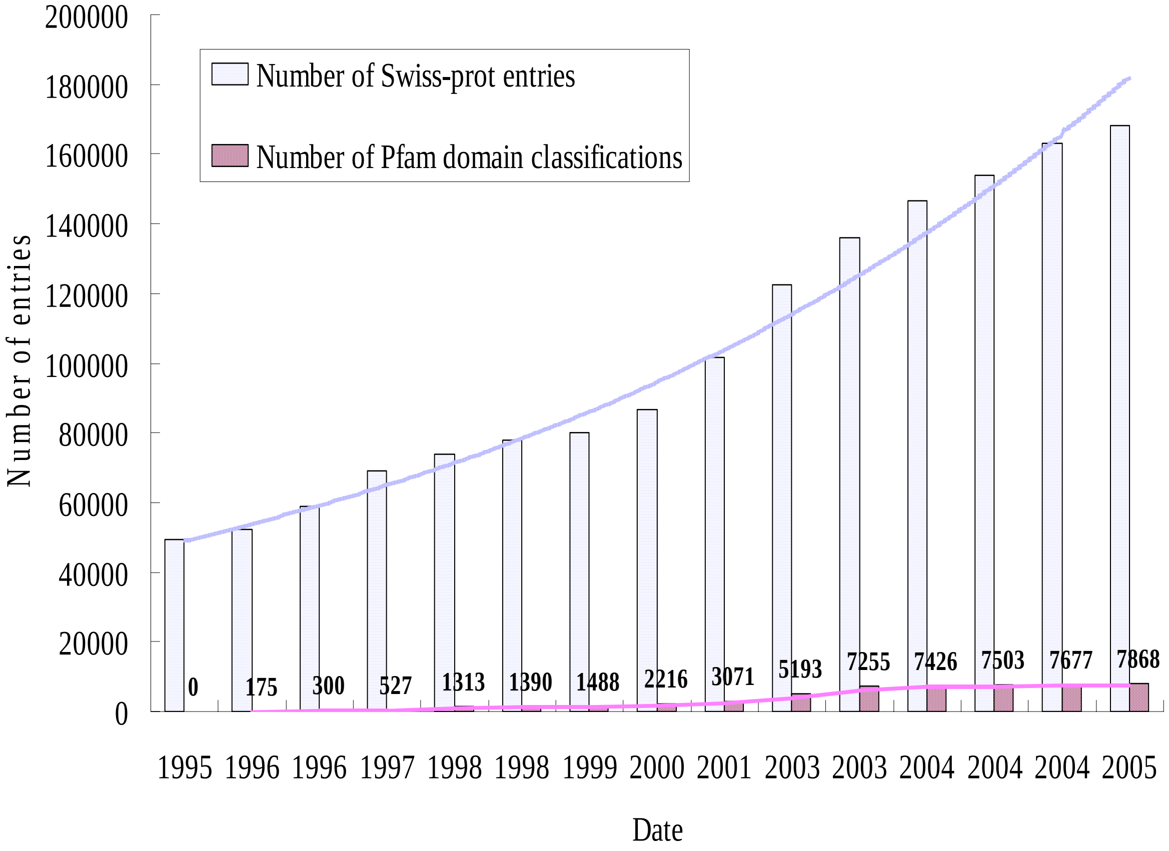


Figure 1. The growth in the number of proteins known in the Swiss-prot versus the growth in the number of unique domains in the Pfam. The number of proteins discovered each year and the number of presently known nonredundant domains found in these proteins. Although the number of proteins discovered grows at increasingly higher rates, the number of domains found appears to be asymptotically reaching a limit.

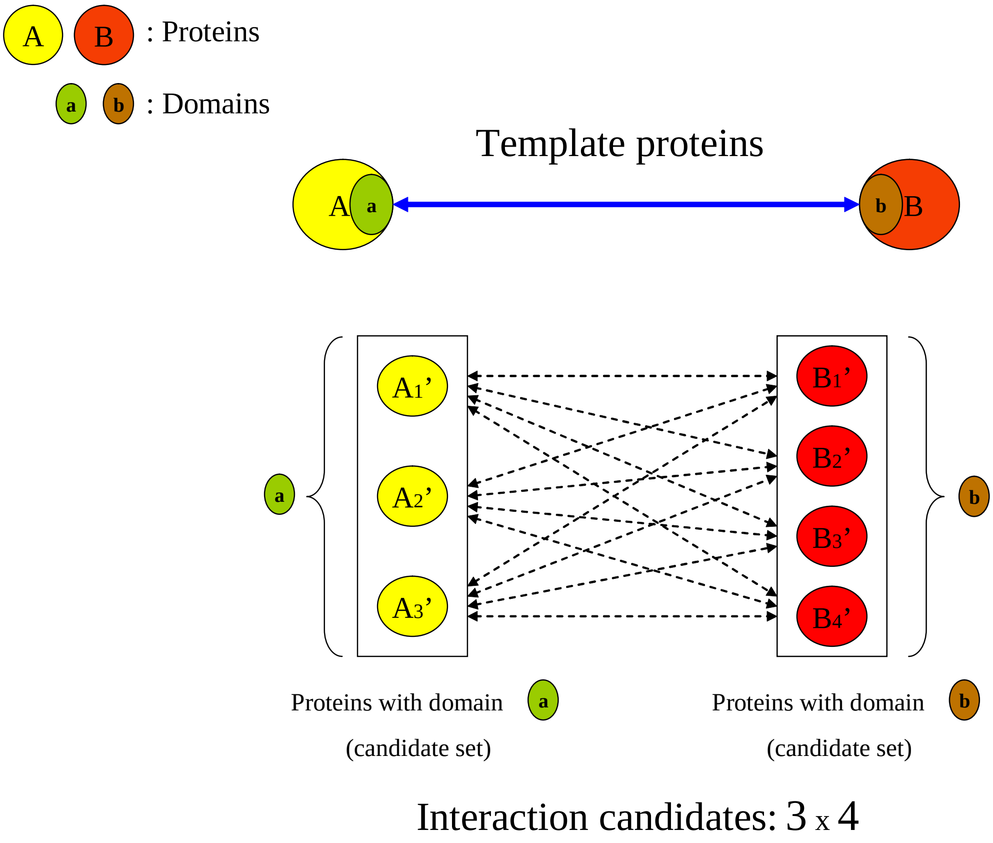


Figure 2. Core ideas of our method. We identified the domain-domain interactions from structural binding proteins (template), and inferred protein-protein interactions (candidates) from domain-domain interactions. The capital letters denote proteins and the lower cases denote domains. In this example, we obtained 12 candidate protein pairs.

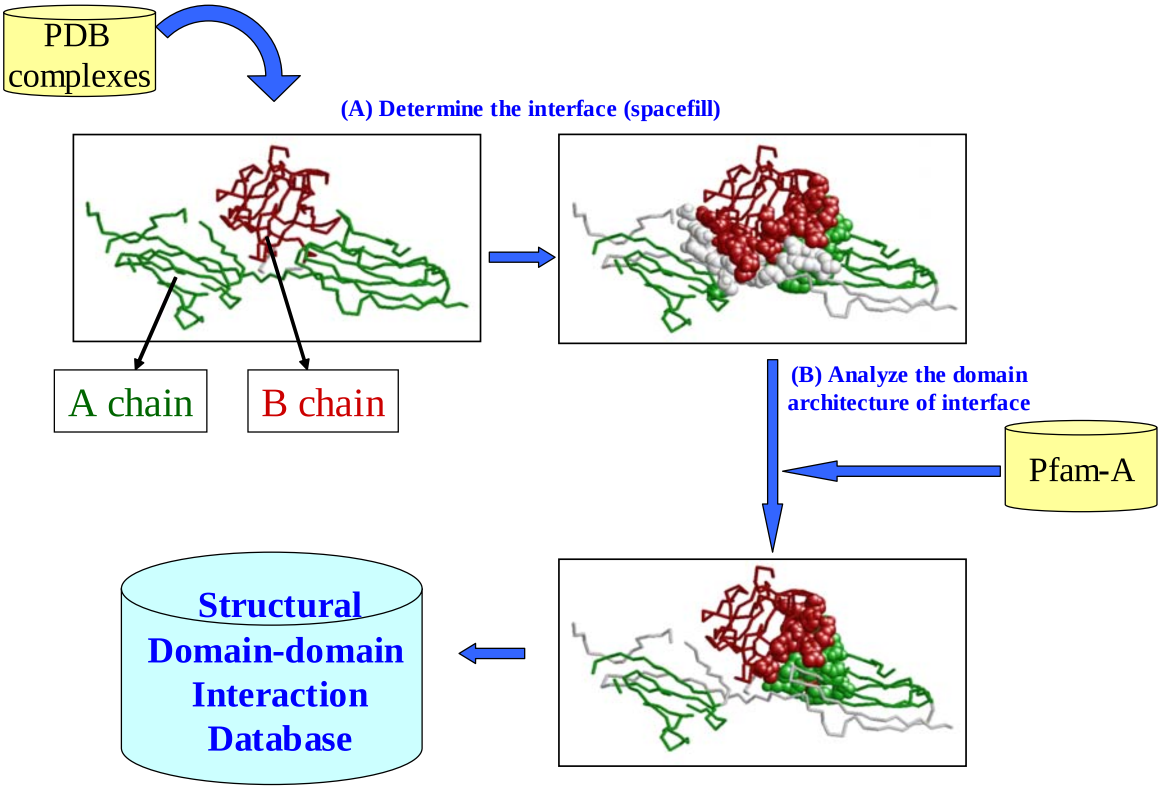

PDB complexes
(A) Determine the interface (spacefill)
A chain
B chain
(B) Analyze the domain architecture of interface
Pfam-A
Structural Domain-domain Interaction Database

Figure 3. The flowchart of extracting structural domain-domain interactions from PDB. We identified the structural protein-protein binding site and mapped to Pfam domain-domain interactions. (A) We used the 5-8 rule (5 or more residues C_β contacted within 8 Å) to identify the protein-protein binding site. (B) To identify the domain architecture of protein-protein binding site.

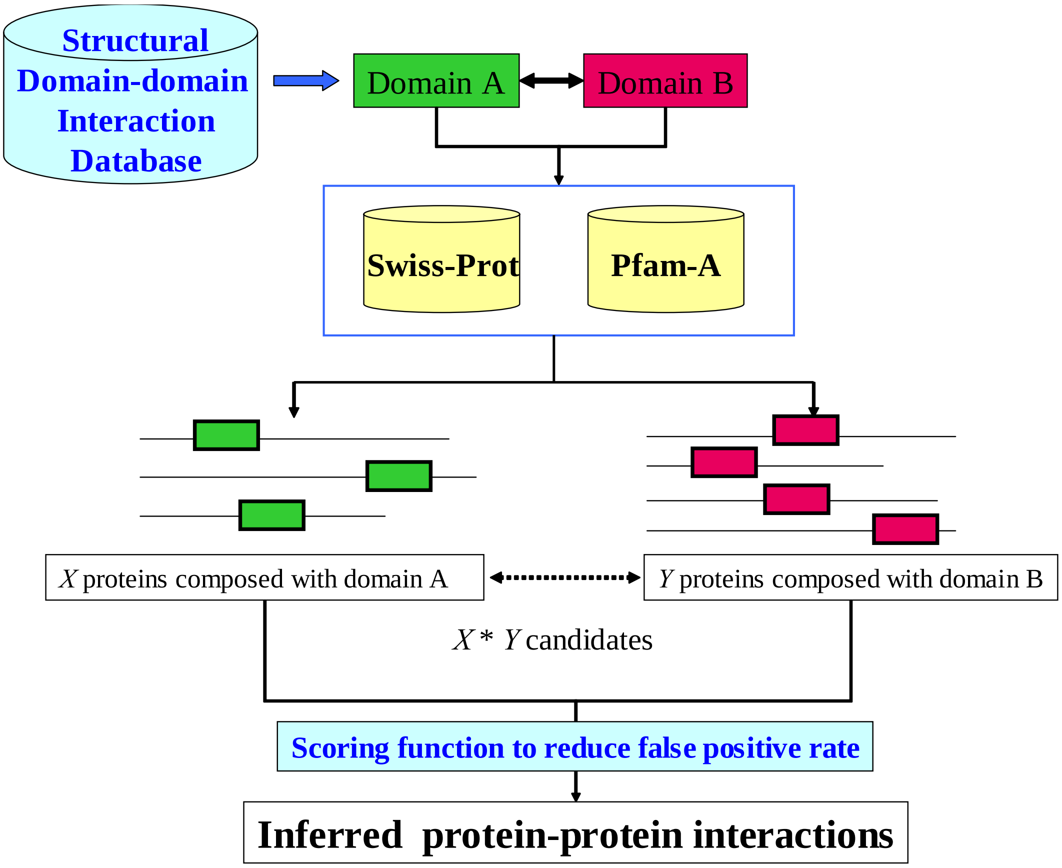
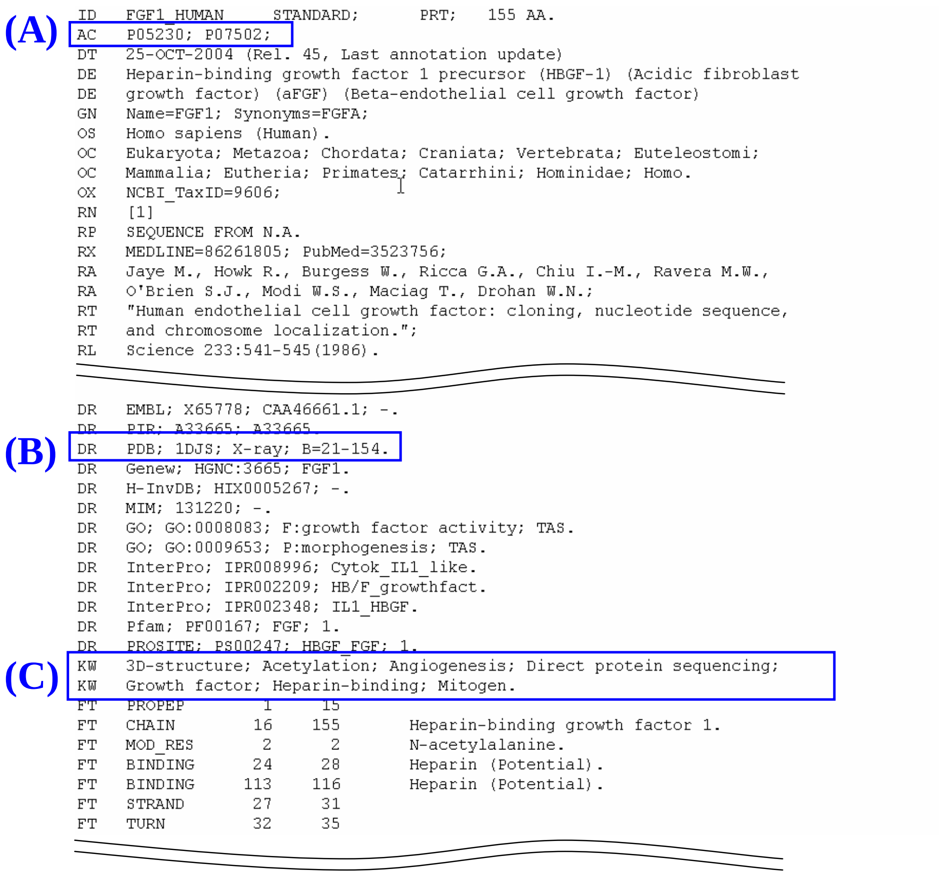

Structural Domain-domain Interaction Database
Domain A
Domain B
Swiss-Prot
Pfam-A
X proteins composed with domain A
Y proteins composed with domain B
X * Y candidates
Scoring function to reduce false positive rate
Inferred protein-protein interactions

Figure 4. Overview of mapping structural domain-domain interactions into protein-protein interactions. We inferred protein-protein interactions from structural domain-domain interactions and employed our scoring function to reduce the false positive rate.



```
ID   FGF1_HUMAN      STANDARD;      PRT;   155 AA.
AC   P05230; P07502;
DT   25-OCT-2004 (Rel. 45, Last annotation update)
DE   Heparin-binding growth factor 1 precursor (HBGF-1) (Acidic fibroblast
DE   growth factor) (aFGF) (Beta-endothelial cell growth factor)
GN   Name=FGF1; Synonyms=FGFA;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=86261805; PubMed=3523756;
RA   Jaye M., Howk R., Burgess W., Ricca G.A., Chiu I.-M., Ravera M.W.,
RA   O'Brien S.J., Modi W.S., Maciag T., Drohan W.N.;
RT   "Human endothelial cell growth factor: cloning, nucleotide sequence,
RT   and chromosome localization.";
RL   Science 233:541-545(1986).

DR   EMBL; X65778; CAA46661.1; -.
DR   PIR; A33665; A33665.
DR   PDB; 1DJS; X-ray; B=21-154.
DR   Genew; HGNC:3665; FGF1.
DR   H-InvDB; HIX0005267; -.
DR   MIM; 131220; -.
DR   GO; GO:0008083; F:growth factor activity; TAS.
DR   GO; GO:0009653; P:morphogenesis; TAS.
DR   InterPro; IPR008996; Cytok_IL1_like.
DR   InterPro; IPR002209; HB/F_growthfact.
DR   InterPro; IPR002348; IL1_HBGF.
DR   Pfam; PF00167; FGF; 1.
DR   PROSITE; PS00247; HBGF_FGF; 1.
KW   3D-structure; Acetylation; Angiogenesis; Direct protein sequencing;
KW   Growth factor; Heparin-binding; Mitogen.
FT   PROPEP        1     15
FT   CHAIN        16    155       Heparin-binding growth factor 1.
FT   MOD_RES       2      2       N-acetylalanine.
FT   BINDING      24     28       Heparin (Potential).
FT   BINDING     113    116       Heparin (Potential).
FT   STRAND       27     31
FT   TURN         32     35
```

Figure 5. An example of the protein annotation (FGF1 Human) in the Swiss-prot database. (A) AC line is the accession number of this protein. (B) DR lines are used as pointers to information related to entries and found in data collections other than Swiss-Prot including the PDB database. (C) KW lines provide the information that can be used to generate indexes of the protein based on functional, structural, or other categories. The keywords chosen for each entry serve as a subject reference for the protein.

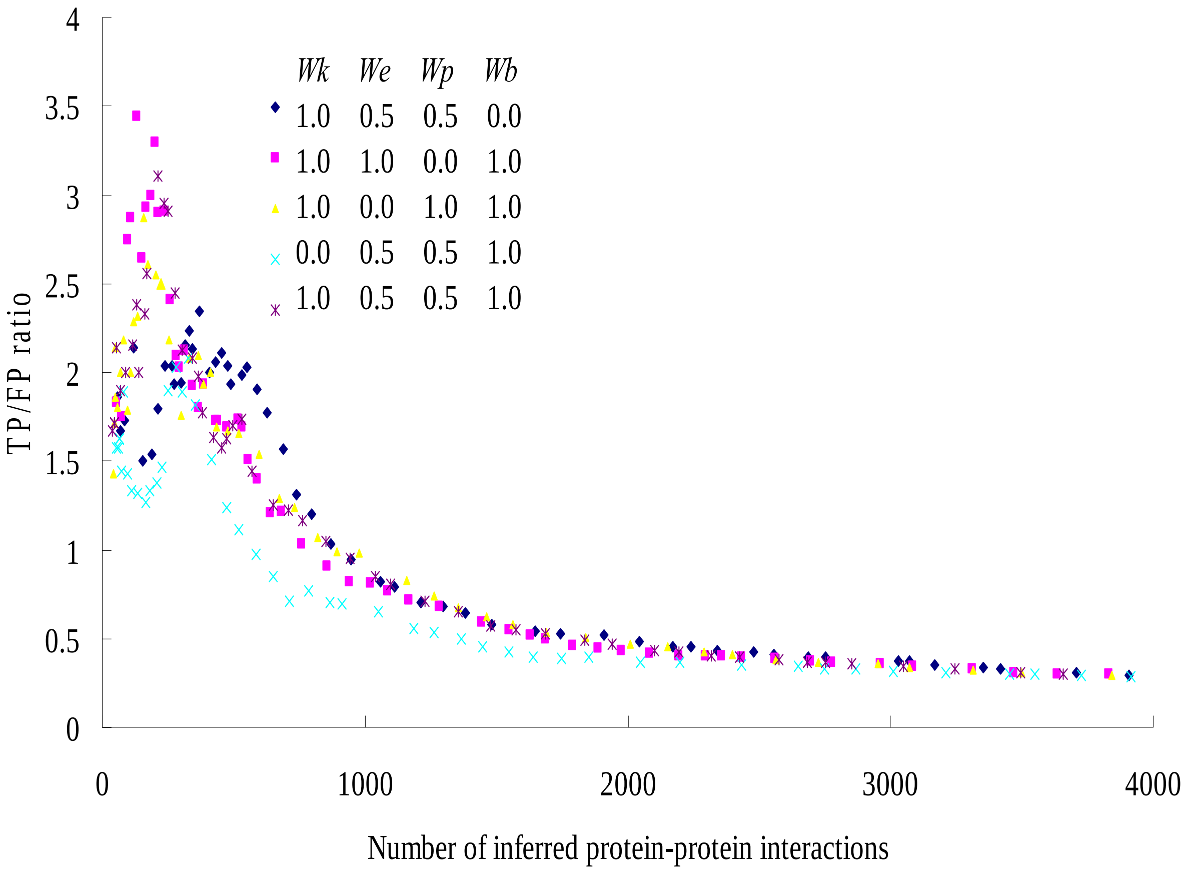


Figure 6. The relationship between the TP/FP ratios and the numbers of protein-protein candidates with different weight combinations of scoring terms. W_k , W_e , W_p, and W_b are defined in Equation 1.

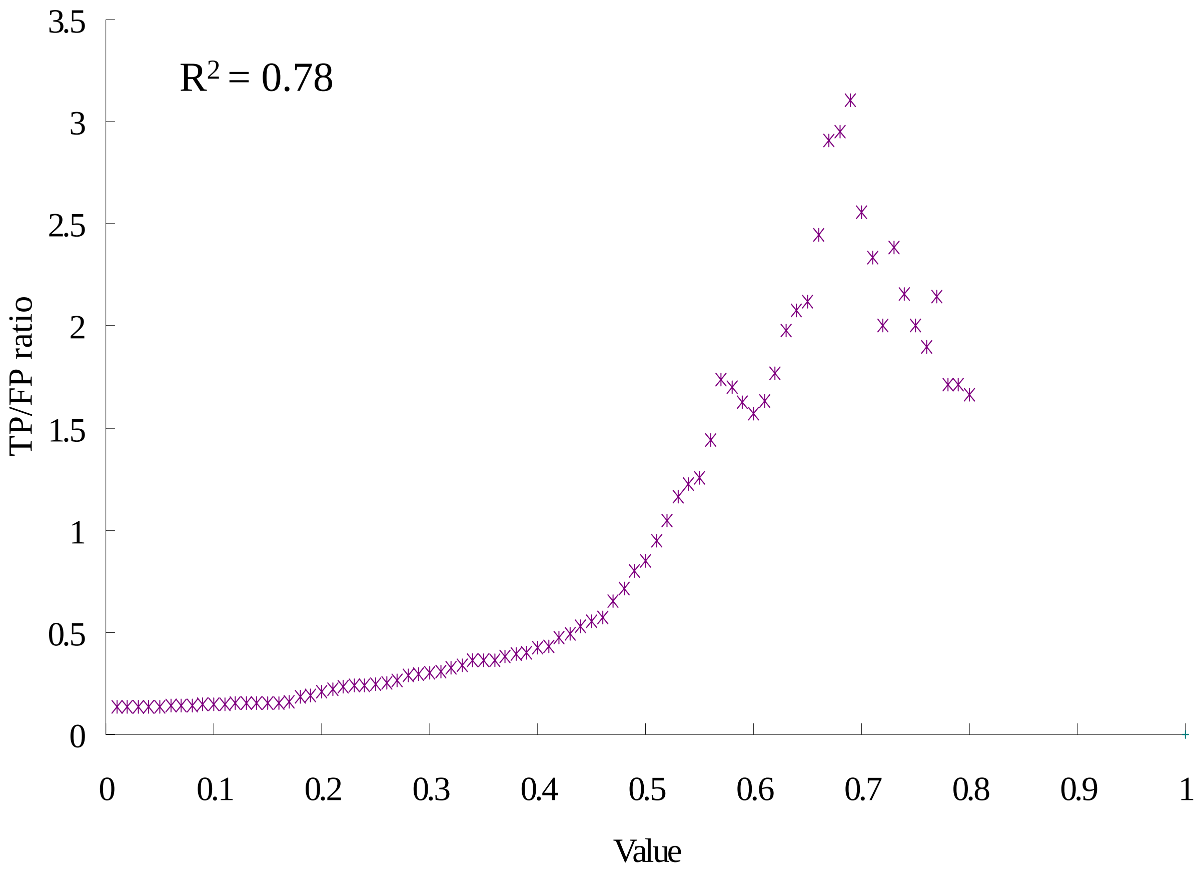


Figure 7. The relationship between the values of our scoring function and the TP/FP ratios. The R^2 means the correlation coefficient between the values of our scoring function and the TP/FP ratios.

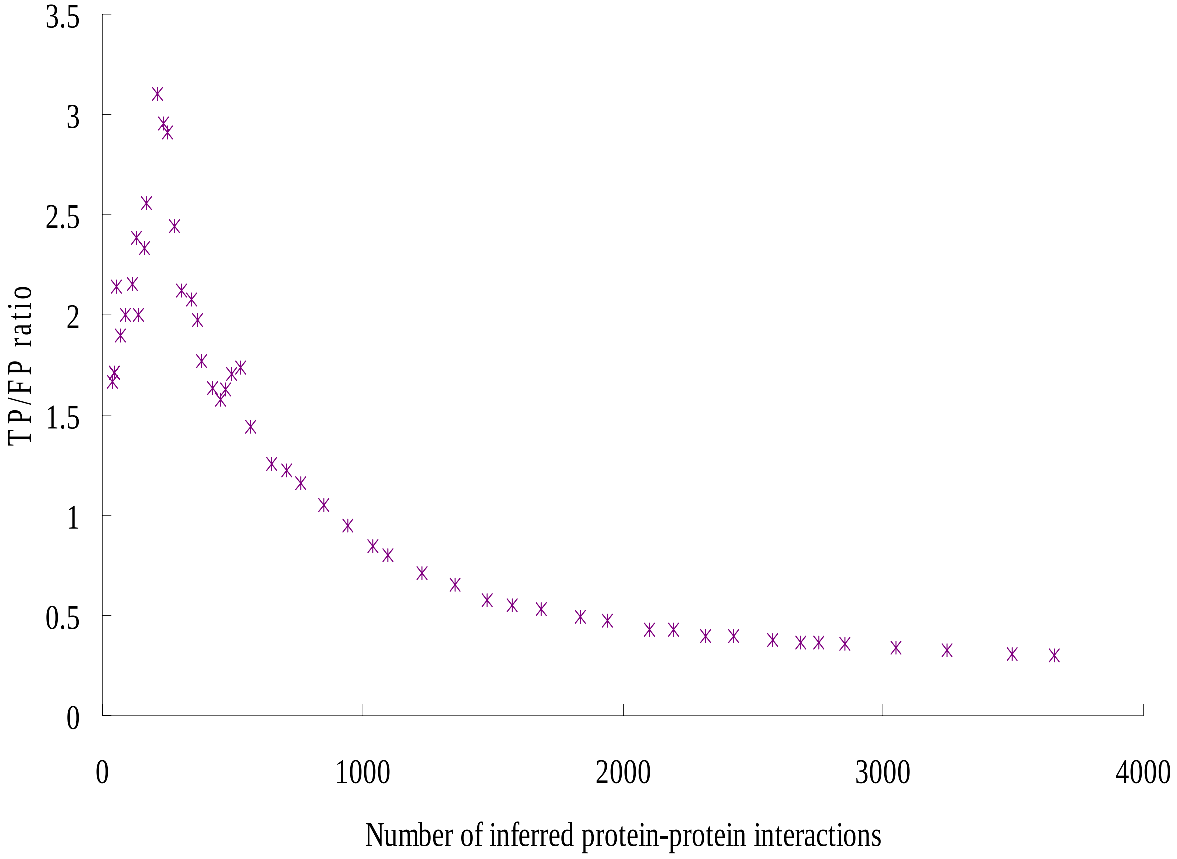


Figure 8. The number of inferred protein-protein interactions (candidates) at different TP/FP ratios.

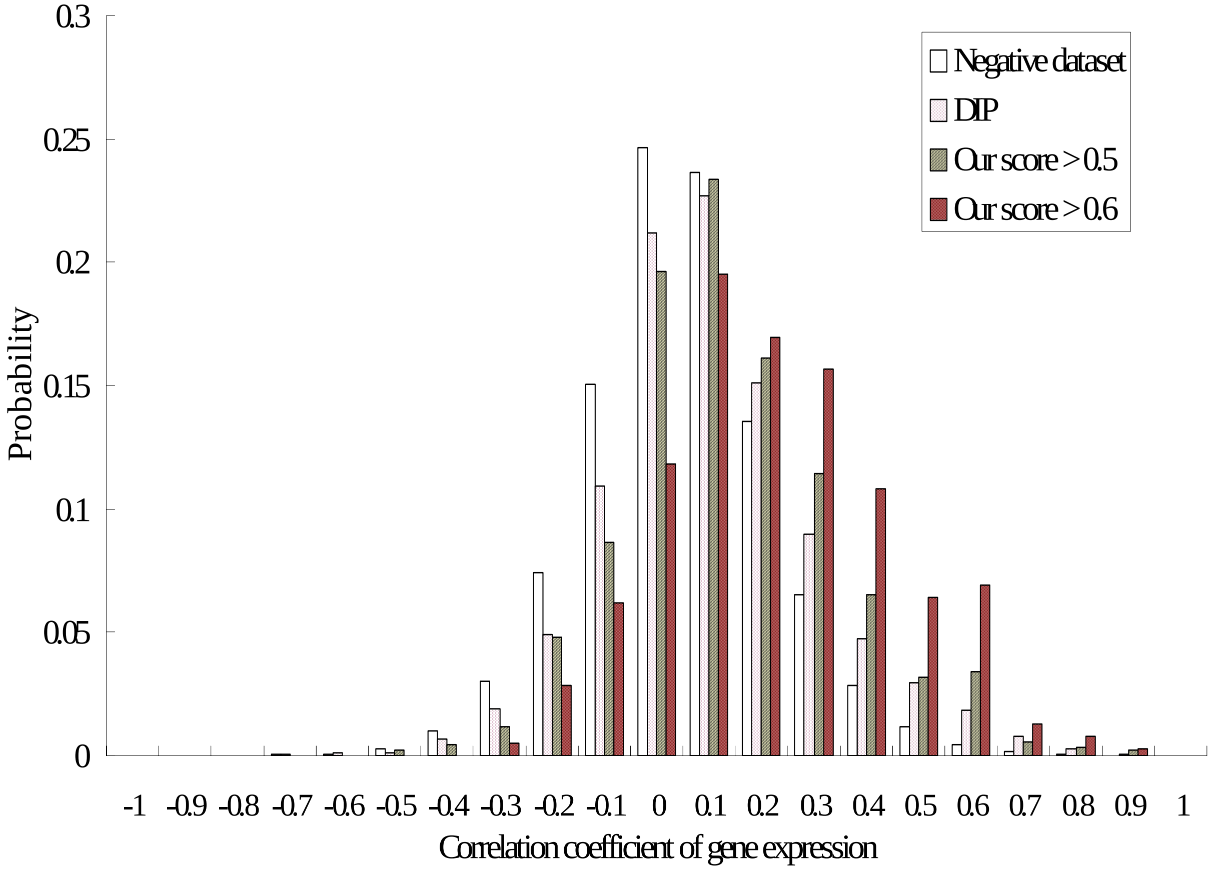


Figure 9. Distributions of the correlation coefficients of gene expression in different datasets. We compared the gene expression profile correlation coefficients of our predictions at different threshold with those of random protein pairs and DIP database, and our predictions have a higher mean correlation coefficient than the ones calculated from the negative data set (defined in Jansen *et al.* , Science 2003 [40]) and the DIP set in *S. cerevisiae*.

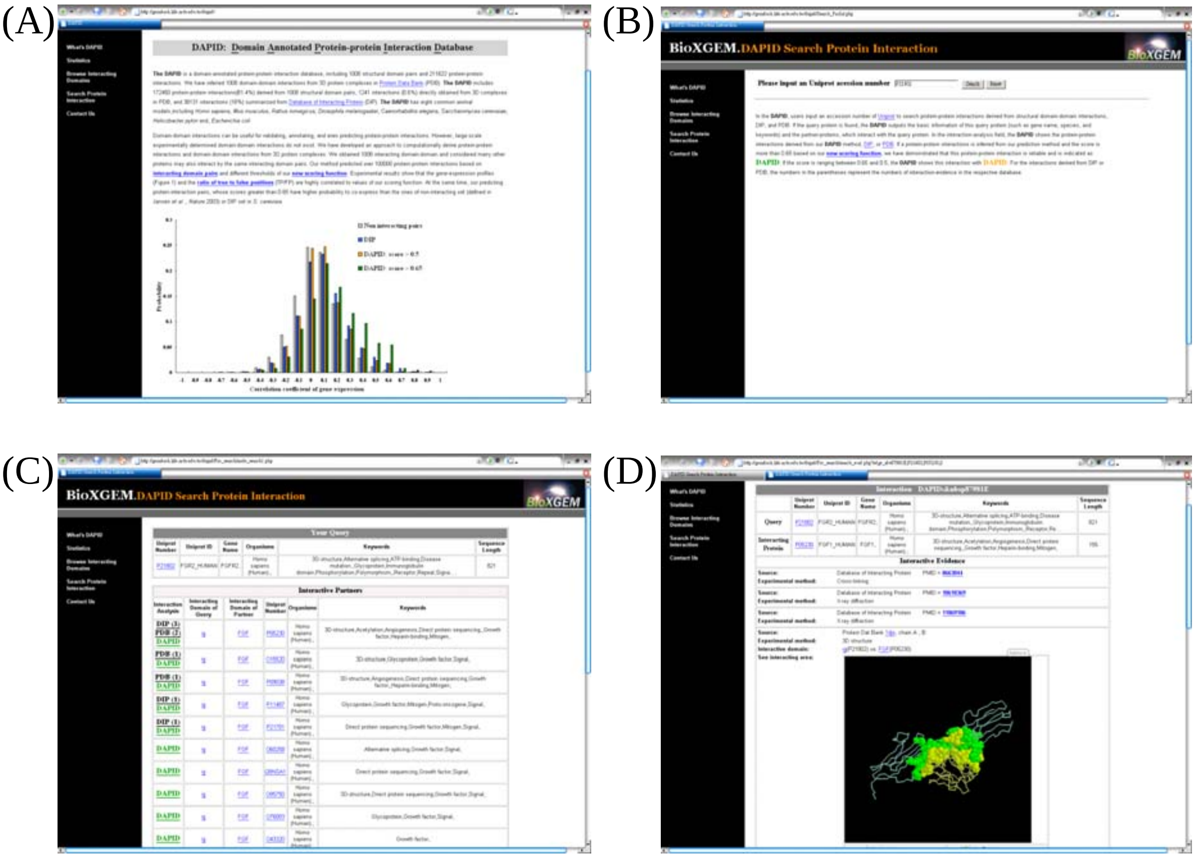

(A)
DAPID: Domain Annotated Protein-protein Interaction Database
What's DAPID
Statistics
Browse Interacting Domains
Search Protein Interaction
Contact Us
Non interacting pairs
DIP
DAPID: score > 0.5
DAPID: score > 0.65
Probability
Correlation coefficient of gene expression
(B)
BioXGEM.DAPID Search Protein Interaction
BioXGEM
Please input an Uniprot accession number
(C)
BioXGEM.DAPID Search Protein Interaction
Your Query
Interactive Partners
(D)
Interactive Evidence

Figure 10. The illustrations of DAPID web service. (A) The home page of DAPID. (B) Query interface of DAPID, users input a Swiss-prot accession number as query. (C) DAPID returns the interacting partners of query protein. (D) The detail information of the query protein and partner proteins.

References

1. Apic, G., Gough, J. & Teichmann, S. A. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology* 310, 311-325.
2. Nye, T. M. W., Berzuini, C., Gilks, W. R., Babu, M. M. & Teichmann, S. A. (2005). Statistical analysis of domains in interacting protein pairs. *Bioinformatics* 21, 993-1001.
3. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. & Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* 31, 365-370.
4. Casari, G., De Daruvar, A., Sander, C. & Schneider, R. (1996). Bioinformatics and the discovery of gene function. *Trends in genetics : TIG* 12, 244-245.
5. Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N. P., Hayes, W. S., Borodovsky, M., Rudd, K. E. & Koonin, E. V. (1996). Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with Escherichia coli. *Current biology : CB* 6, 279-291.
6. Brenner, S. E., Hubbard, T., Murzin, A. & Chothia, C. (1995). Gene duplications in H. influenzae. *Nature* 378, 140.
7. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research* 28, 235-242.
8. Davis, F. P. & Sali, A. (2005). PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics* 21, 1901-1907.
9. Ponting, C. P. & Russell, R. R. (2002). The natural history of protein domains. *Annual review of biophysics and biomolecular structure* 31, 45-71.
10. Veretnik, S., Bourne, P. E., Alexandrov, N. N. & Shindyalov, I. N. (2004). Toward consistent assignment of structural domains in proteins. *Journal of Molecular Biology* 339, 647-678.
11. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C. & Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Research* 32, 138-141.
12. Ng, S. K., Zhang, Z. & Tan, S. H. (2003). Integrative approach for computationally inferring protein domain interactions. *Bioinformatics* 19, 923-929.
13. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86-90.
14. Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D.

(1999). Detecting protein function and protein-protein interactions from genome sequences. *Science* 285, 751-753.

15. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America* 96, 4285-4288.
16. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature* 402, 83-86.
17. Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A. & Legrain, P. (2001). The protein-protein interaction map of Helicobacter pylori. *Nature* 409, 211-215.
18. Wojcik, J. & Schachter, V. (2001). Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics* 17, 296S-305.
19. Bock, J. R. & Gough, D. A. (2001). Predicting protein-protein interactions from primary structure. *Bioinformatics* 17, 455-460.
20. Gomez, S. M., Lo, S.-H. & Rzhetsky, A. (2001). Probabilistic Prediction of Unknown Metabolic and Signal-Transduction Networks. *Genetics* 159, 1291-1298.
21. Sprinzak, E. & Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology* 311, 681-692.
22. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America* 98, 4569-4574.
23. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. & Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 403, 623-627.
24. Legrain, P., Wojcik, J. & Gauthier, J. M. (2001). Protein-protein interaction maps: a lead towards cellular functions. *Trends in genetics : TIG* 17, 346-352.
25. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417, 399-403.
26. Sprinzak, E., Sattath, S. & Margalit, H. (2003). How reliable are experimental protein-protein interaction data? *Journal of Molecular Biology* 327, 919-923.
27. Legrain, P. & Selig, L. (2000). Genome-wide protein interaction maps using two-hybrid systems. *FEBS letters* 480, 32-36.
28. Hazbun, T. R. & Fields, S. (2001). Networking proteins in yeast. *Proceedings of the*

National Academy of Sciences of the United States of America 98, 4277-4278.

29. Mrowka, R., Patzak, A. & Herzel, H. (2001). Is There a Bias in Proteome Research? *Genome Research* 11, 1971-1973.

30. Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M. & Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Research* 28, 289-291.

31. Ng, S. K., Zhang, Z., Tan, S. H. & Lin, K. (2003). InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Research* 31, 251-254.

32. Deng, M., Mehta, S., Sun, F. & Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Research* 12, 1540-1548.

33. Mewes, H. W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., Warfsmann, J. & Ruepp, A. (2004). MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Research* 32, 41-44.

34. Walhout, A. J. M., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N. & Vidal, M. (2000). Protein Interaction Mapping in C. elegans Using Proteins Involved in Vulval Development. *Science* 287, 116-122.

35. Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J.-D. J., Bertin, N., Chung, S., Vidal, M. & Gerstein, M. (2004). Annotation Transfer Between Genomes: Protein-Protein Interologs and Protein-DNA Regulogs. *Genome Research* 14, 1107-1118.

36. Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. & Yeh, L. S. (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Research* 33, D154-159.

37. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215, 403-410.

38. Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T. & White, R. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* 32, D258-261.

39. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. &

Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25, 3389-3402.

40. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F. & Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302, 449-453.

41. Ge, H., Liu, Z., Church, G. M. & Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nature genetics* 29, 482-486.

42. 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.

43. 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., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. & Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell* 102, 109-126.

44. Szebenyi, G. & Fallon, J. F. (1999). Fibroblast growth factors as multifunctional signaling factors. *International review of cytology* 185, 45-106.