Computational Methods to Uncover the Protein-Protein Interactions

Fei-Hung Hung[1], Hung-Wen Chiu[1*]
[1]*Graduate Institute of Medical Informatics, Taipei Medical University, Taiwan*
[*]*Correspondence E-mail: hwchiu@tmu.edu.tw*

Abstract

Protein-protein interaction (PPI) is an important mechanism for the cell's life. Advances on PPI information acquisitions made the prediction of PPI possible. Some high throughput experimental methods have been applied to find novel PPIs. Computational methods have been proposed for inferring PPI on the basis of known information about protein function and sequences. Once the protein-protein interaction can be predicted successfully, it will helpful to design new medications. In this paper, we briefly introduce some computational methods to uncover PPI and illustrate their achievements. Furthermore, we designed and demonstrate a tool to retrieve useful information about PPI.

1. Introduction

Protein-protein interaction (PPI) provides the valuable information for a better understanding to cell life. Traditionally, PPI has been studied by chemical experiments [1-3]. The in-silico prediction of PPI is to know whether two proteins can interact or not by non-chemical experiment. The key point of PPI prediction is to find the rule about how two proteins interact based on known protein information. Any protein information can be used for the attempt -- the protein sequence, the function, the comparative relationship and many technical literatures for example. Depending on the work of wet lab, many useful information about PPI has been obtained and stored in databases for further research. Recently, a few high throughput approaches, e.g. yeast two-hybrid system [2,3], produces more than one half of total PPI information. A computer with higher speed process ability is essential in understanding PPI. Owning to the rapid accumulation of PPI knowledge, the rule about how two proteins interact has the chance to be extracted. Once the rule is uncovered the complex life system will be comprehended more clearly. The prediction of PPI also provides the new direction for the development of medication and treatment on molecular level. Pathogenic protein can be inhibited by its interacted protein. Moreover some researchers display PPI information using network graph to make protein-protein interrelationship more clearly. In this paper we presented some computational methods for PPI prediction and analyzed their achievements to demonstrate the feasibility of in-silico PPI prediction. Furthermore, we presented some our developed tools to prepare the data for predicting PPI.

2. PPI Related Databases

2.1. Protein Sequence and Knowledge Databases

The Protein Information Resource (PIR, http://pir.georgetown.edu/) is devoted to spreading protein annotation and standardization. Now we could use PIR to classify protein families, find protein Knowledge and search literatures about protein.

SWISS-PROT (http://ca.expasy.org/sprot/) is a curated protein sequence database which strives to provide a high level of annotation and clearly protein sequence. Because it collects only experimental verified existences, related researchers could use it to find non-redundant information.

National Center for Biotechnology Information (NCBI, http://www.ncbi.nih.gov/) was established in 1988 as a national resource for molecular biology information. NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. After collection data form many different databases, researchers may discover that their data forms are not alike. At this moment, researchers can apply information in NCBI to transform them, so as to continuous.

Protein Families Database (Pfam, http://www.sanger.ac.uk/Software/Pfam/) is a large database covering many common protein domains and families. Investigators could look at multiple alignments of protein, view protein domain architectures and get known protein structures here. For application, researchers could use this database to check the function of one protein and apply it to find similarly functional regions. In addition, Pfam will show the visual graph of proteins the users searched.

2.2. PPI Databases

In addition to protein database, there are many databases for protein-protein interaction. Database of Interacting Proteins (DIP, http://dip.doe-mbi.ucla.edu/) database catalogs experimentally determined interactions between proteins. It combines information from a variety

of sources to create a single, consistent set of protein-protein interactions.

Biomolecular Interaction Network Database (BIND, http://www.bind.ca/Action) is a database collecting protein-protein interaction, cellular data and protein pathway. This web site describes PPI's characters and uses public literatures to show us.

The MIPS Mammalian Protein-Protein Interaction Database (MPPI, http://mips.gsf.de/proj/ppi/) is a collection of manually curated high-quality PPI data collected from the scientific literature by expert curators. They took great care to include only data from individually performed experiments since they usually provide the most reliable evidence for physical interactions.

2.3. Databases for PPI Prediction

These protein databases and PPI databases lead investigators to interest the prediction of PPI. Now this kinds of databases increase constantly. There is special for human PPI, besides biomolecular. Some team uses PPI to infer protein domain-domain interaction.

PreBIND (http://prebind.bind.ca/) is a tool of PPI prediction based on BIND and combines published papers. PreBIND contains biomolecular interaction. That is not merely protein-protein interactions. It would rather give a suitable interaction than a prediction. Returns from the PreBIND import links of related literatures to NCBI. When researchers link to its homepage, they could see the search frame. You could use protein name, NCBI accession number and PubMed ID to search. If some one wants to know a protein MSN2's interactions, he keys the protein name to this search frame and push submit down (Figure 1). The result page shows him the information of MSN2 above (Figure 2) and appears those interactions in published literatures below (Figure 3).

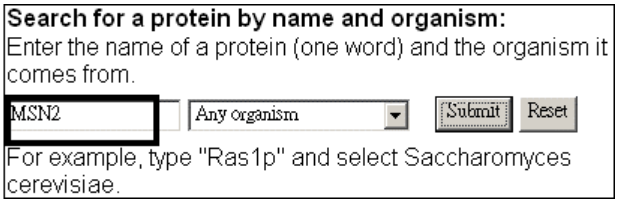

Figure 1. Picture of PreBIND's Search This is the main page of PreBIND. Type your key word in the search frame, then push " Submit" bottom to search.

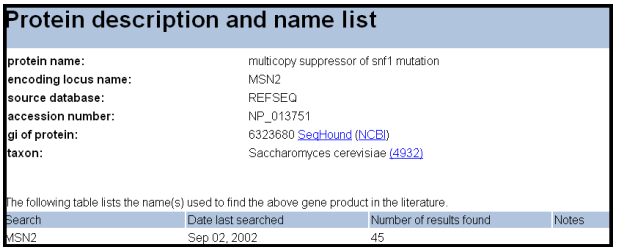

Figure 2. Upper the Result of PreBIND The frame is the basic information of this protein you key in.

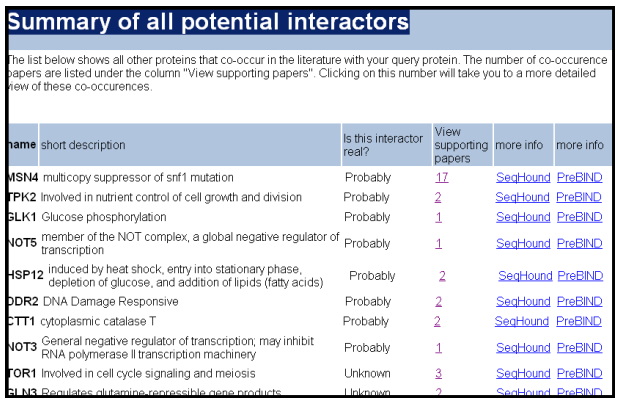

Figure 3. Under the Result of PreBIND It shows all interactions of literatures on the internet.

Human Protein Interaction Database (HPID, http://wilab.inha.ac.kr/hpid/) is a database for human protein-protein interactions. HPID applies statistics and computations to provide researchers finding interactions. You enter their homepage and see a frame for search. Here we could type IDs of many database including NCBI to process. Figure 4 shows the procedures of search.

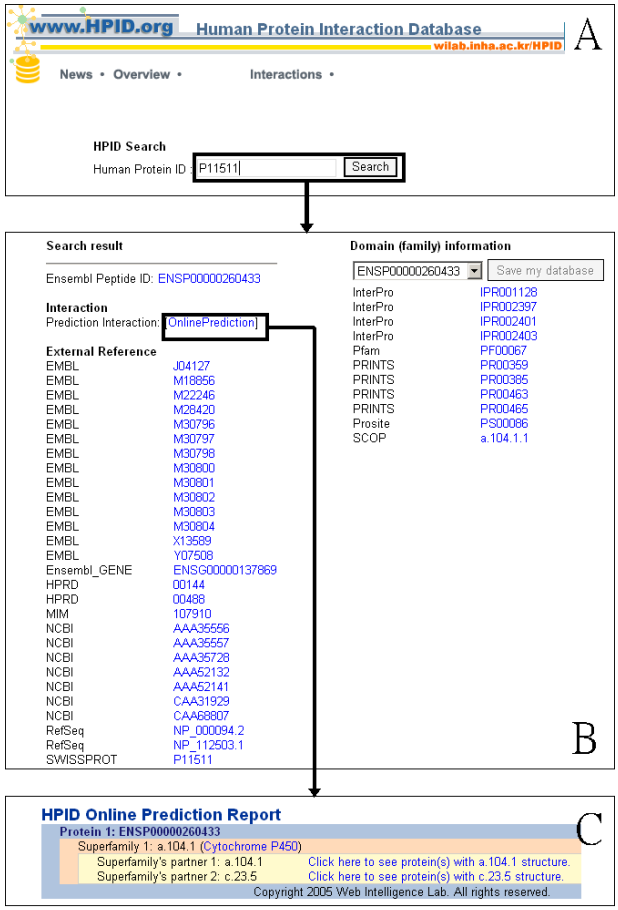

Figure 4. Procedures of using HPID (**A**) Researchers can type what they want in the "Human Protein ID" frame to search. (**B**) There is the result page of searching. Users can see its basic information and click the link "Online Prediction" to appear a new window to show the prediction of PPI. (**C**) The new window of PPI prediction.

Protein-Protein Interaction Database (PPID, http://www.anc.ed.ac.uk/mscs/PPID/) is a database covering human, mouse and rats. We can use association numbers, type of protein and their PPID number to search what you want. As the figure F, we chose the kind of input (Figure 5A) and key what you want to search (Figure 5B), then press the "Search". In an instant, it appears the result below (Figure 5C).

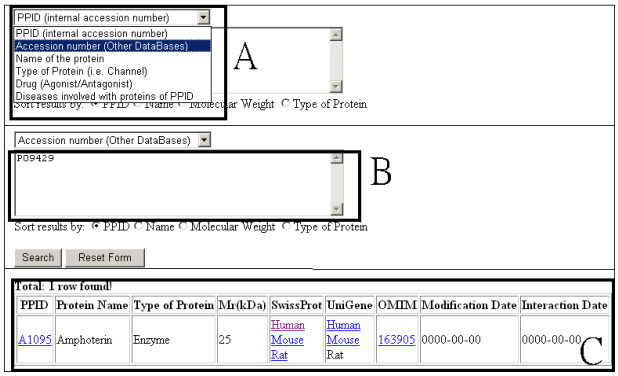

Figure 5. Return of PPID (**A**) Upper the input frame, researcher has several choices for input type. (**B**) Users can type the protein to search. (**C**) The result.

Database of Interacting Domain (InterDom, http://interdom.lit.org.sg/) is a specially database. They predict protein domain-domain interactions from PPIs. Though it is not for prediction of PPI, their approach is applied on prediction of PPI. InterDom is a database of *putative* interacting protein domains derived from multiple sources, ranging from domain fusions (Rosetta Stone), protein interactions (DIP and BIND), protein complexes, to scientific literatures. When investigators use, they should chose "Search" and it would give three selections. Researchers can use protein, domain and sequence to do their search (Figure 6A). We use protein P09429 to set an example. Here we should click the "by Protein" button, then type P09429 into the search frame (Figure 6B). When we push "Search", we get the answer. You could see their description, domain name and domain-domain interaction (Figure 6C).

By many investigators' striving, these related information tools have many achievements. Through communications of information, researchers can almost apply these database without misunderstanding (Table 1).

Table 1. Data Sources of Related Protein Databases

	PIR	SWISS-PROT	NCBI	Pfam	DIP	BIND
DIP	*	*	*			
BIND		*	*			
MPPI		*	*			
PreBIND						*
HPID					*	*
InterDom			*	*	*	*

Ψ It is appeared that data of PPI interflow by a wide margin. The consistence of data form means researchers are not depressed to share their research any more. This research domain carves out further.

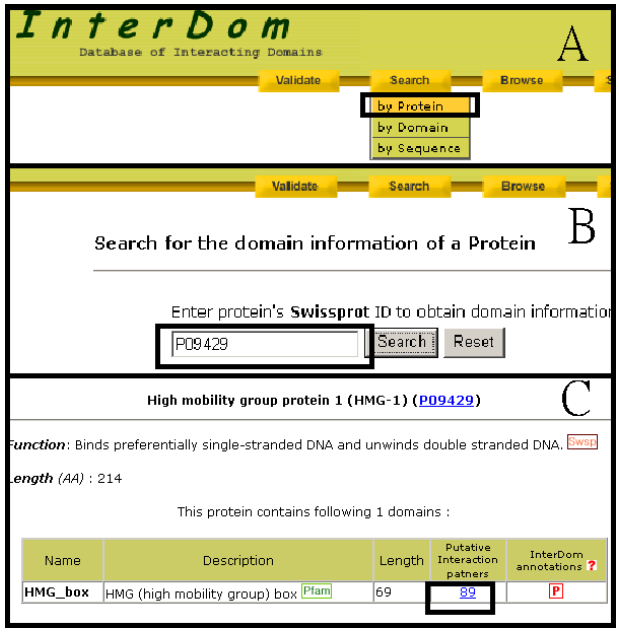

Figure 6. Use InterDom (**A**) At the main page, it shows three options for search. (**B**) Search by protein. (**C**) It is return for result. Under this page, it shows the information for the protein and interaction.

3. In-Silico Methods to Predict PPI

Because of the high throughput of protein-protein interaction, research workers will not be able to analyze these data by hand gradually. Thus researchers apply the computer and designed algorithms to analyze and then transform those data into useful information. That is the mainstream approach now.

3.1. Homology

Protein-protein interaction and DNA-protein interaction is the important sources for protein research. However these data are limited. It spent much manpower and material resources at many laboratory. The final result is only specific specie. How to apply known information into other organisms? One way is homology. Now investigators know protein X and protein Y interact in organism A. In organism B, there are protein X' that is ortholog to protein X and protein Y' that is ortholog to protein Y. In that case, researchers could infer there is a interaction between protein X' and protein Y'. Such relationship as above is named the homlology [4-6].

3.2. Rosetta Stone

Rosetta stone is a stele. Archaeologista translate the ancient Egypt hieroglyph by it. One of protein sequence alignment approach is named it [4]. If investigators want to know whether two protein interact or not, they search similar sequence fusing by the foregoing two in other organisms. These protein sequences in other organisms are called Rosetta stone protein. In case they find similar one or more, they think the prediction exists. Now this

concept is applied for many relationships, for instance, domain-domain fusions [6].

3.3. Data Mining

Data mining is a method to mine useful information from a large number of data. For example, some investigators used Maximum Likelihood Estimation (MLE) to predict [7].

Text mining and graph mining are both kinds of data mining. Text mining is to use published literatures, journals and the contents of many web sites to test their key words and contexts for protein-protein interactions. Finally, researchers store the resulted relationship for get PPI or domain-domain interaction information [6].

E. Segal *et al.* point out that many cellular pathway have two characteristics. One is their gene expression are similar. Second, products of these genes often interact [8]. Besides many researchers link those identification proteins each other to turn them into a connected network graph. It supports a visual tool to appear relationship of proteins [9]. Some researcher applies graph mining to find novel protein complexes [10].

4. Current Achievements of PPI Prediction

Presently plenty research teams effort in PPI prediction, but all methods can not predict it perfectly. Nevertheless, the successful percentage of prediction enhances constantly. Two research teams, Uetz and Ito, used Y2H system to make high throughput of PPI even since [2,3]. The prediction of protein-protein interaction is not so far. It is known that we can get a large network including 3186 mostly novel interactions among 1705 proteins [1]. This interaction network can be applied to find interactions linking uncharacterized gene products and human disease protein about regulatory cellular pathways. It offers us a new viewpoint of protein-protein interaction. Edward M. Marcotte et al. brings "Rosetta Stone" approach. They apply this approach to use gene sequences for inferring PPI [4]. In the past, the protein functions depend on their structures, but investigators can recognize the whole through the observation of the part by using homology. It creates a distant path. A database was built by applying homology to extending PPIs experimentally verified in module organism to the orthologous proteins in *Homo sapiens* [5]. Besides Minghua Deng et al. apply data mining, e.g. MLE method, Association Method, to predict [7]. Dong-Soo Han et al. use more accurate computational approach, it is domain used, to do PPI prediction. In their way, the sensitivity is 77% and the specificity is 95% [11,12]. Graph mining is used by Xiao-Li Li et al. to build LCMA algorithm for prediction of PPI [10]. See-Kiong Ng et al. and Minghua Deng et al. infer protein domain-domain interactions from protein-protein interactions [6,7]. Their way differs from others. See-Kiong Ng's team use homology and text mining, too. They create the InterDom database. Table 2 lists some different methods from different teams.

Table 2. Lists of Different Methods

Team	Methods	Result or Verification
Edward M. Marcotte et al.[4]	Rosetta Stone Domain fusion homology	By filtering promiscuous domains, they make true interactions by 47% over the unfiltered predictions.†
Xiao-Li Li et al.[10]	Graph Mining (LCMA algorithm)	LCMA was compared with MCODE‡. For the F-measure§, LCMA is 15.99% higher score than MCODE.
Dong-Soo Han et al.[11,12]	(domain combination based protein-protein interaction prediction method)	Sensitivity = 77% Specificity = 95%
Minghua Deng et al.[7]	Data Mining (MLE Method)	Compare with the result of association method and another database
See-Kiong Ng et al.[6]	Homology Rosetta Stone Text Mining	(Probabilistically-weighted Odd Ratios)

† This data was about *E. coli.*

‡ MCODE is a proposed created by Bader and Hogue [13].

§ They defined $Recall = |TP| / (|TP| + |FN|)$ and $Precision = |TP| / (|TP| + |FP|)$. F-measure takes into account of both precision and recall and is defined as $F\text{-}measure = 2 * Precision * Recall / (Precision + Recall)$.

5. Discussion

Attempts of uncovering protein-protein interactions is an accumulation of human wisdom. PPI prediction should be the greatest topic of this generation. Kinds of Protein complexes are numerous to lead to a huge degree of difficulty. Although now technology can not decode it, human continuous study and the high throughput of relative information would make it truth about eradicating cancer, stopping aging and curing inheritable disease some day in the future.

The current trend is to compare some relationship of interaction for prediction and it gets a considerable achievement. Even so, there are many job needed to be strived. Horng JT. et al have supported a approach about applying computational method with data mining for protein motif. Their team wants to predict protein structure [14]. Here we try a way about applying protein motif to predict PPI without prediction of structure. Investigators know there is one or more motifs in the protein. Those motifs concern with protein functions thickly. For the reason, we find respective motifs out from known interacted proteins. Next, we may be able to look for the relationship of protein-protein interaction. Finally, we could use the relationship found to do PPI prediction. Now we have collect data of DIP, BIND, PIR and NCBI and use PHP programs we made to transfer them to our form. Now DIP and the non-redundant human PPI of BIND are completed. Total is 18,617 and store them in our MySQL database. The web-based

database (Figure 7) is building. Its functions are fragmentary. InterDom is our direction. We hope that will support another useful tool when the job is finished.

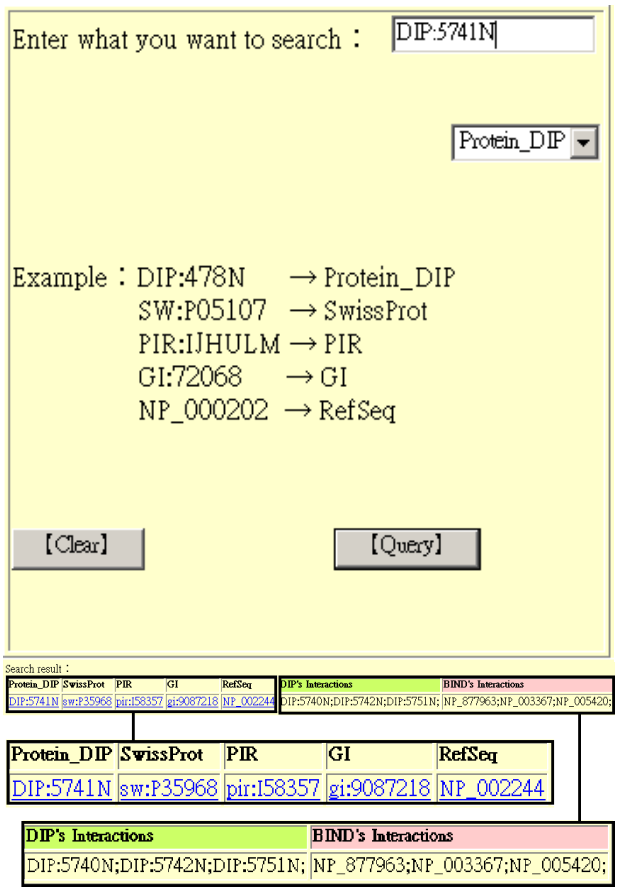

Figure 7. Our Database

6. References

[1] Stelzl U., Worm U., Lalowski M. *et al.*, "A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome", *Cell*, 2005, Vol. 122, p. 957–968.

[2] Uetz P., Giot L., Cagney G. *et al.*, "A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae", *Nature*, 2000, Vol. 403, p. 623-627.

[3] Ito T., Chiba T., Ozawa R. *et al.*, "A comprehensive two-hybrid analysis to explore the yeast protein interactome", *Proc. Natl. Acad. Sci.*, 2001, Vol. 98, p. 4569-4574.

[4] Marcotte EM., Pellegrini M., Ng HL. *et al.*, Detecting Protein Function and Protein-Protein Interactions from Genome Sequences, *SCIENCE*, 1999, Vol. 285, p. 751-753.

[5] Cesareni G., Ceol A., Gavrila C. *et al.*, "Comparative interactomics", *FEBS Letters*, 2005, Vol. 579, p. 1828–1833.

[6] Ng SK., Zhang Z., Tan SH. *et al.*," InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes", *Nucleic Acids Research*, 2003, Vol. 31(1), p. 251-254.

[7] Deng MH., Mehta S., Sun FZ. *et al.*, "Inferring Domain-Domain Interactions From Protein-Protein Interactions", *Genome Research*, 2002, Vol. 10, p. 1540-1548.

[8] Segal E., Wang H., Koller D. *et al.*, "Discovering molecular pathways from protein interaction and gene expression data", *Bioinformatics*, 2003, Vol 19, p. 264-272.

[9] Hu Z., Mellor J., Wu J. *et al.*, "VisANT: an online visualization and analysis tool for biological interaction data", *BMC Bioinformatics*, 2004, Vol. 5,p. 17-24.

[10] Li XL., Tan SH., Ng SK. *et al.*, "Interaction Graph Mining for Protein Complexes Using Local Clique Merging", *Genome Informatics*, 2005, Vol. 16(2), p. 260-269.

[11] Han DS., Kim HS., Seo JM. *et al.*, "A Domain Combination Based Probabilistic Framework for Protein-Protein Interaction Prediction", *Genome Informatics*, 2003, Vol. 14, p. 250-259.

[12] Han DS., Kim HS., Jang WH. *et al.*, "PreSPI: Design and Implementation of Protein-Protein Interaction Prediction Service System", *Genome Informatics*, 2004, Vol. 15(2), p. 171-180.

[13] Bader GD. and Hogue CWV., "An automated method for finding molecular complexes in large protein interaction networks", *BMC Bioinformatics*, 2003, Vol. 4(1):2.

[14] Horng JT., Huang HD., Wang SH. *et al.*, "Computing Motif Correlations in Proteins", *Comput Chem*, 2003, Vol. 24, p. 2032-2043.