PROTEIN INTERACTION NETWORK – INFERENCE TOOL
CAPSTONE PROJECT REPORT

BY

DIVYA N. RAO

ADVISOR: Dr. FILIPPO MENCZER
CO-ADVISORS: Dr. ALESSANDRO FLAMMINI
Dr. HAIXU TANG

SCHOOL OF INFORMATICS
INDIANA UNIVERSITY
BLOOMINGTON, INDIANA

TABLE OF CONTENTS

INTRODUCTION

Proteins along with carbohydrates, lipids and nucleic acids form the building blocks of all organisms. Proteins are macromolecules composed of amino acids that are connected by peptide bonds. They play an integral role in the structure and function of all organisms. For example: membrane proteins are a part of the membrane of a cell or organelle whereas hormones help in the regulation of various biological processes. The amino acid components of a protein specify its structure and hence its function.

The interactions between proteins are important for many biological functions. These interactions can be of different kinds – enzymes react with substrates during a metabolic reaction, different proteins may interact in order to form a protein complex, proteins may bind to each other in order to transfer a molecule or a protein may interact with another protein in order to modify it.

Anomalies in the interactions between proteins can cause an imbalance in the proper functioning of the organism. For example: Huntington Disease is a progressive neurodegenerative disorder caused by an elongated polyglutamine tract in a large protein Huntingtin, of unknown function. In a study of the proteins involved in the disease, a network of 86 proteins involved in 188 protein-protein interactions was constructed. Studying this network it was found that alterations in the process of Huntingtin interacting with other proteins involved in vesicle trafficking, cytoskeletal organization and transcriptional regulation are important for disease development. The protein-protein interaction network of Huntington Disease provides a valuable basis for the identification of new drug targets for therapeutic intervention (Goehler *et al*., 2004). The study of protein-protein interactions is an important aspect in understanding the complex cellular processes of an organism and in identifying targets for drug development.

The physical structure of a protein is specified by its amino-acid composition. The DNA or protein sequences can be compared to see if two proteins homologous based on their sequence similarity. Sequence homology can indicate shared ancestor, common function, or simply random chance. Homologous sequences are said to be Orthologous if they were separated by a speciation event, if a gene exists in a species, and that species diverges into two species, then the

copies of this gene in the resulting species are orthologous. In the fig. 1 below if species 1 has protein domains a & b and species B has orthologous domains A & B, then proteins A & B most probably interact to generate the same function as species 2. This method has been used to predict thousands of interactions in various organisms, as their genomes are sequenced (Marcotte, E. M. *et al* 1999).

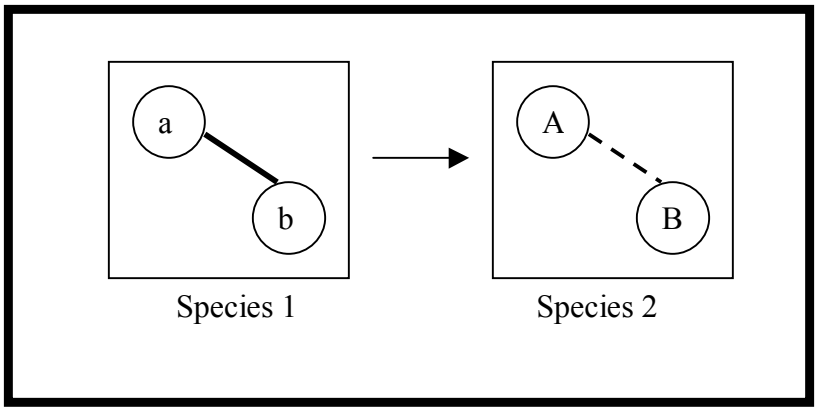


Figure 1. Orthologous proteins

Protein-protein interactions help in understanding the biochemical processes underlying cellular function and lifecycle. Protein-protein interaction maps provide a valuable framework for a better understanding of the functional organization of the proteome. In addition, protein sequences help in identifying homologs that can be used to validate functional relationships between proteins. In this paper we will discuss a tool that was built to infer protein-protein interactions based on known protein interactions and the homology between proteins. The tool would serve as a starting point in identifying protein-protein interactions.

BACKGROUND

Initially protein interactions were identified using the top-down, hypothesis driven approaches of genetics, biochemistry and biophysics. Though experimental methods such as Immunoprecipitation produced high quality results, they were excessively time consuming. In the past two years, with the advances in proteomics technology new bottom-up interaction detection approaches such as the yeast two-hybrid system and mass spectrometry have been developed to detect large scale protein interactions. Experimentally verified interactions have been compiled in various large scale protein–protein interaction datasets (Gavin *et al*., 2002; Ito *et al*., 2001; Xenarios *et al*., 2002; Bader *et al*., 2003).

The Yeast Two-Hybrid (Y2H) system utilizes genetically engineered strains of yeast (*Saccharomyces cerevisiae*) to identify protein-protein binding (fig. 2). The mutant strains of yeast are incapable of synthesizing certain nutrients (amino acids such as Tryptophan or nucleic acids) and cannot survive when grown on media lacking these nutrients. New DNA can be incorporated into these organisms with the use of plasmids. Two kinds of plasmids are introduced into the mutant yeast strains – one encoding the protein that will fuse to the DNA-binding domain and the other encoding the protein that fuses with the transcription activating domain. Any interaction between the proteins will lead to transcription of the proteins and the mutant strain will be able to grow on the selective media. The Y2H is powerful tool that can be applied in a high-throughput manner to detect interactions across the entire proteome of an organism. It has been used to detect proteome wide interactions in model organisms such as *H. pylori* (Rain *et al*., 2001), *S. cerevisiae* (Ito *et al*., 2001; Uetz *et al*., 2000), *C. elegans* (Li *et al*., 2004) and *D. melanogaster* (Giot *et al*., 2003). As it is an *in vivo* technique, it allows for the detection of unstable and transient interactions. However, the disadvantages of this technique are that it allows for testing only two proteins at a time and as it takes place in the nucleus many proteins are not in their native state and the interactions do not take into account the physiological setting. It has been found that for filtered yeast two-hybrid dataset the fraction of false positives is predicted to be of the order of 50% (von Mering *et al.*, 2002).

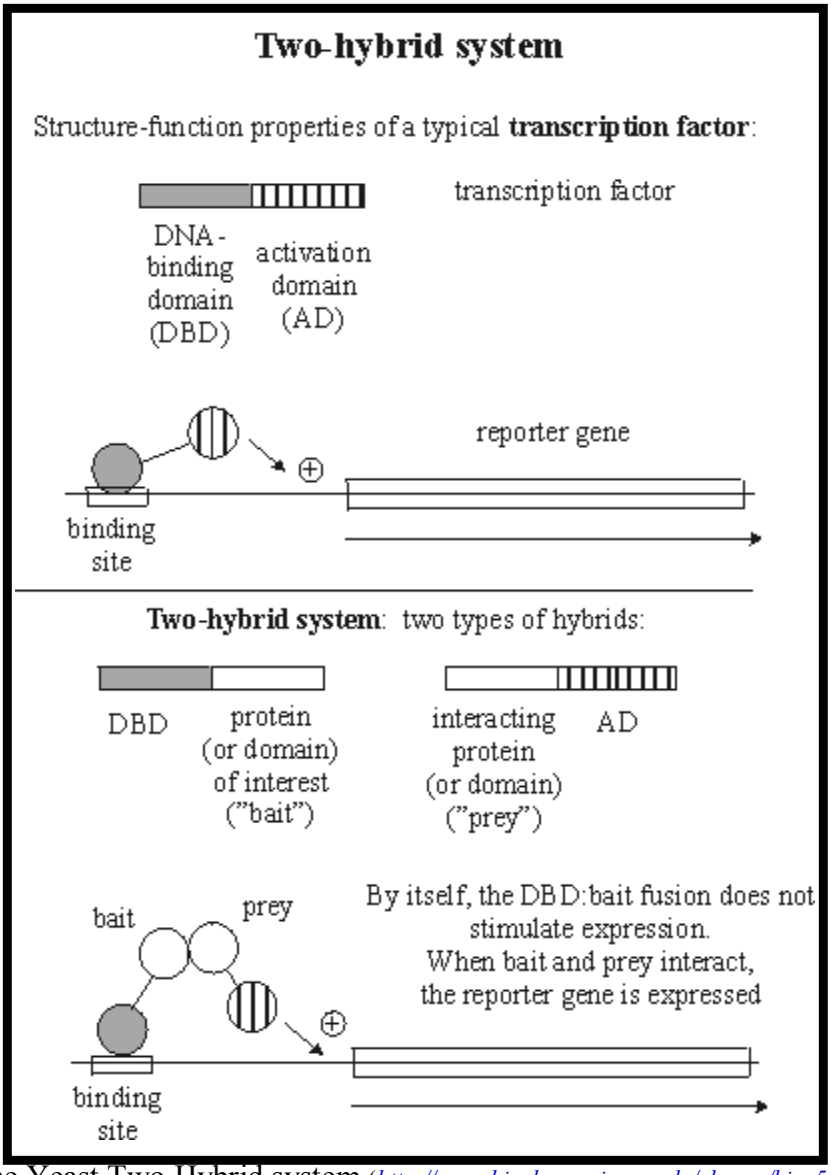


Figure 2. The Yeast Two Hybrid system (http://www.biochem.arizona.edu/classes/bioc568/bioc568.htm)

Mass spectrometry uses individual proteins that are tagged as "hooks" to biochemically purify whole protein complexes. These are then separated and their components identified by mass spectrometry. Tandem Affinity Purification (TAP) and High-throughput Mass Spectrometric Protein Complex Identification (HMS-PCI) are two of the protocols used. The advantages of using this system of detection are that several members of a complex can be tagged providing an internal check for consistency and it detects complexes in their physiological settings. However, the disadvantages of this approach are that some of the proteins may not be present in the given conditions and may be overlooked and the tagging may disrupt the formation of complexes (von Mering *et al.*, 2002).

Though the high-throughput methods have generated large amounts of interaction data, their results include a large number of false-positive and false-negative associations. However, they are still extremely invaluable to interpret protein–protein interactions and construct protein–protein networks (Salwinski and Eisenberg, 2003; Lu *et al*., 2002). Computational methods can address protein–protein interactions at different levels. They may focus on in-depth analysis or carry out a broad scale analysis across large datasets. Through genomic and protein sequence analysis, they may infer whether proteins do interact (Salwinski and Eisenberg, 2003; Lu *et al*., 2002). Methods using genomic and protein sequence data include analysis of presence or absence of genes in related species, conservation of gene neighborhood, gene fusion events, similarity of phylogenetic trees, correlated mutations on protein surfaces and co-occurrence of sequence domains (Salwinski and Eisenberg, 2003).

Thus, we find that several databases exist such as DIP (Salwinski and Eisenberg, 2003) and BIND (Bader *et al*., 2003) whose main purpose is to collect and curate direct experimental evidence about protein–protein interactions. Other databases such as KEGG (Kanehisa, M., *et al.,* 2004) and MetaCyc (Krieger, C.J., *et al.,* 2004) take a more generalized perspective on proteins and their associations, by functionally grouping proteins into metabolic, signaling or transcriptional pathways. Finally, a third class of resources attempts to fill gaps in both datasets, by predicting protein–protein associations *de novo*, using a variety of computational techniques. InterWeaver, Protein Interactions by Structural Matching (PRISM), Automated Detection and Validation of Interaction by Co-Evolution (ADVICE) and Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) are some of the currently available tools to predict protein-protein interactions

InterWeaver is a web server aimed at predicting potential protein interaction partners based on various online resources and prediction methods (Zhang and Ng, 2004). Currently, they use two approaches – homology based and domain based approach to identify potential interaction partners for the user's uncharacterized source protein. In the homology based approach, they use data mined from various online protein databases and biomedical literature to find homologs in different species using BLAST for the source protein. Then, they mine the online protein interaction databases (DIP, BIND) and protein complex database (PDB) to find

experimentally verified protein interactions and complexes for the source protein. They also use text mining techniques to extract interaction information from abstracts of biomedical literature in the PubMed database. In the domain based approach they use computationally derived domain fusion events and domain-domain interactions to infer protein that putatively interact with the source protein.

ADVICE is a web server providing Automated Detection and Validation of Interaction based on the Co-Evolutions between interacting proteins (Tan, S.H., *et al.* 2004). They use sequence analysis to determine proteins' evolutionary histories in order to detect co-evolved interacting proteins (fig. 3). The first step involves the search for homologs. The pair of sequences submitted by the user is used to search sequence databases for orthologous sequences based on sequence similarities. Identified orthologous sequences will be used to compute each input protein's evolutionary history. ADVICE allows users the option to search for orthologous sequences either from one of the four kingdoms of life (Eukaryota, Prokaryota, Archaebacteria and Viridae) or from the Swiss-Prot and/or TrEMBL databases. BLAST is used to search these databases and the user can control the sensitivity of the search by setting an *E*-value threshold for the BLAST hits. In the second step, ADVICE will then construct the distance matrix for both othologous group of sequences. The distance matrices are derived from multiple sequences alignments using ClustalW. In the final step, they use the Pearson's Correlation Coefficient formula to calculate the similarities between the two distance matrices. The result *r* will fall into -1 to 1. Previous studies have indicated that interacting proteins share similarity in their evolutionary histories and have high r-value (>=0.8)

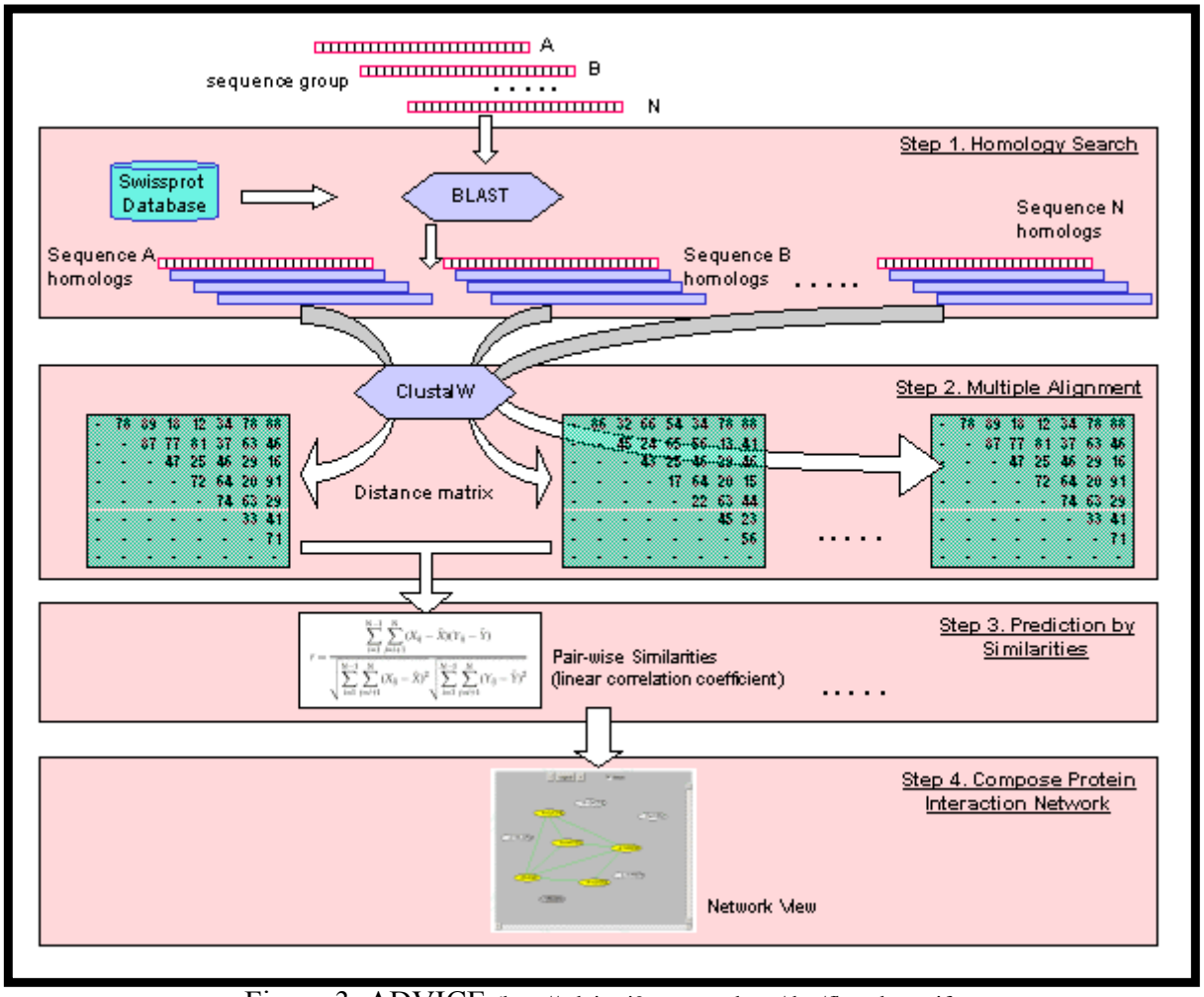


Figure 3. ADVICE (http://advice.i2r.a-star.edu.sg/doc/flowchart.gif)

Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) predicts protein associations based on the genomic context (von Mering 2005). They search completely sequenced genomes for conserved genomic neighborhoods, gene fusion events and co-occurrence of genes across genomes in order to identify pairs of genes which appear to be under common selective pressures during evolution and which are therefore thought to be functionally associated. These associations are assigned a confidence score by comparing with the KEGG reference set. They also use literature mining and functional genomics data to derive the protein-protein associations.

Protein Interactions by Structural Matching (PRISM) predicts pairs of polypeptide chains that may potentially interact in a target dataset of protein structures by comparing them with a

template dataset of protein interfaces which is a structural and evolutionary representative of all biological and crystal interactions present in the PDB (Aytuna 2005). PRISM is developed and maintained by the Center for Computational biology at Koc University in England. As the name suggests, they use protein structure information to predict interactions. In this tool, they find every possible binary interaction between pairs of structures in the target dataset. To do this, they extract surfaces of target proteins and perform successive structural alignments between these surfaces and the partner chains of interfaces in template interface dataset, in an all-against-all manner. This allows them to measure the "structural similarity" of a target structure to a template binding site. If surfaces of two target proteins (A and B) contain regions "similar" to complementary partner chains of a template interface, they say A and B may interact through these "similar" regions. Further, they check for the presence of hot spots on the target structure. Hot spot match ratio is used for the calculation of an "evolutionary similarity score" whereas structural match ratio is used for a "structural similarity score". Combination of these scores contributes to the overall prediction score (fig. 4).

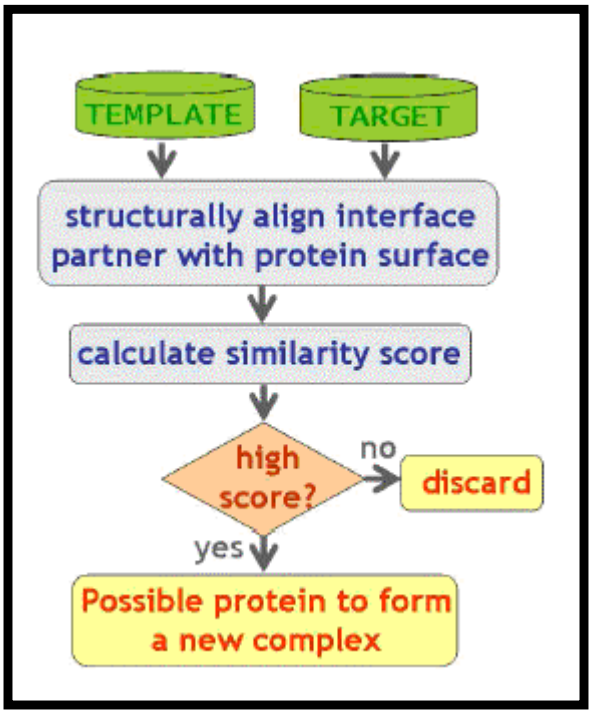


Figure 4. PRISM (http://gordion.hpc.eng.ku.edu.tr/prism/tutorial.php#predictions.php)

It can be seen that currently available protein-protein interaction prediction tools such as InterWeaver, ADVICE and STRING use protein homology and others such as PRISM use

protein structure information to predict protein – protein interactions. Thus, there is a need for a tool that can use both the protein sequences and known experimental evidence to predict protein-protein interactions. The Protein Interaction Network – Inference Tool represents an effort to predict protein-protein interactions based on experimental evidence and protein homology.

METHODS

Data Sources

With the advances in molecular biology and genetic techniques biologists are able to obtain large amounts of information about proteins, their structure, function and their interacting partners. All of this information is now available through public databases on the World Wide Web. The National Center for Biotechnology Information (NCBI) is a large data warehouse containing information not only about proteins but also DNA and RNA. There are protein specific databases such as the Protein Information Resource (PIR) that specialize in proteins – their families, domains, structure, prost-translational modifications, UniProt is an extensive curated database of protein function, classification and cross-reference. There are also databases that specialize in protein interaction information. The Database of Interacting Proteins (DIP), Biomolecular Interaction Network Database (BIND) and MIPS Mammalian Protein-Protein Interaction Database (MPPI) are databases that focus on protein-protein interaction information. The data for this tool was obtained from both DIP and BIND.

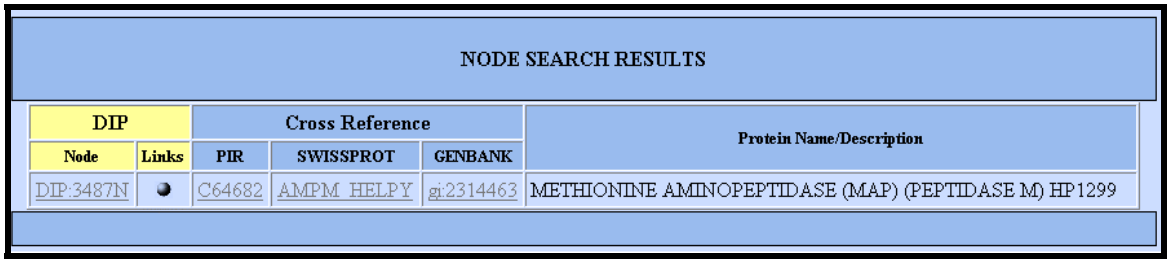
Database of Interacting Proteins (DIP)

The Database of Interacting Proteins (DIP) is maintained by the Molecular Biology Institute at UCLA. It contains experimentally verified protein-protein interaction data. The information is curated from scientific literature and archives, both manually by experts as well as through computationally automated procedures. For each protein involved in an interaction, DIP provides the gene name, description, enzyme code, cellular localization as well as cross references to other protein sequence databases such as Swiss-Prot, PIR and GenBank. The interaction information includes the ranges of amino acids, protein domains of the interacting proteins and the experiments used to detect the interaction. In addition, DIP also lists the literature citations associated with the identification of the interaction (Xenarios *et al*., 2000). Table 1 lists the DIP database statistics as of March 2006.

Table 1: DIP database statistics obtained from the DIP website

Number of proteins	19051
Number of organisms	110
Number of interactions	55732
Number of data sources (including articles)	3077

We ran a sample query for the protein Methionine Aminopeptidase present in *Helicobacter pylori*. The GenBank ID for this protein is: 2314463. As searching DIP with the GI number did not produce any results, we had to determine the protein's accession number and SWISS-PROT ID to use as a query for DIP. Fig. 5 shows the results obtained from DIP for this query. They provide cross-references to PIR, SWISS-PROT and GenBank and a short description and name of the protein.



NODE SEARCH RESULTS

DIP		Cross Reference			Protein Name/Description
Node	Links	PIR	SWISSPROT	GENBANK	
DIP:3487N		C64682	AMPM_HELPY	gi:2314463	METHIONINE AMINOPEPTIDASE (MAP) (PEPTIDASE M) HP1299

Figure 5. Query results from DIP

Clicking on the Node link provides more detailed cross-reference information including links to protein domain database (PRINTS), protein families (Pfam) database among others. They also provide a link to graphically visualize the protein's interaction network as shown in Fig. 6.

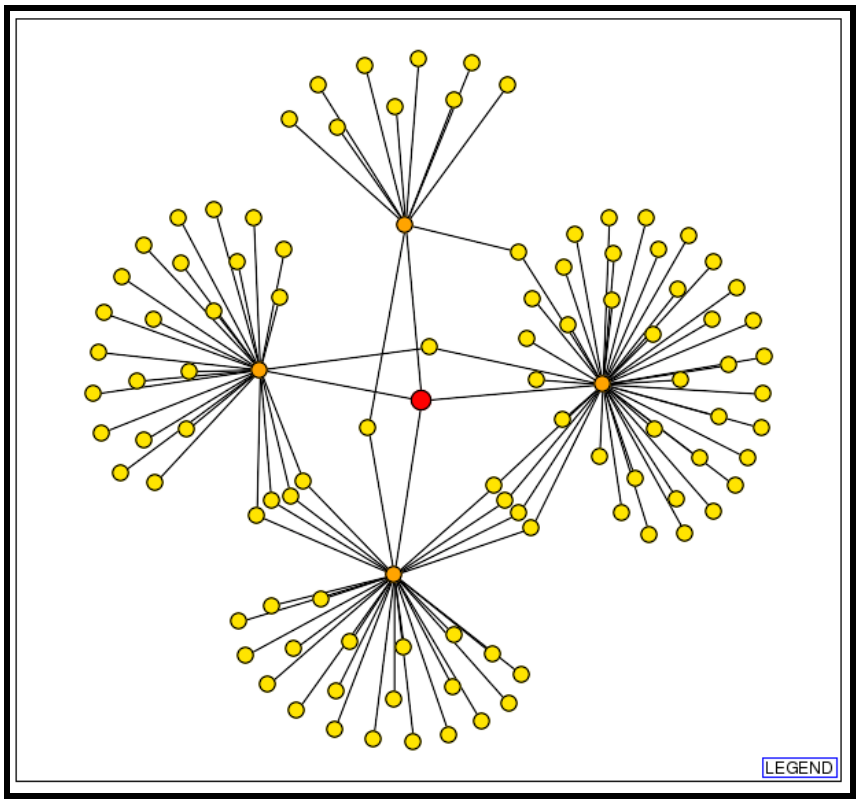
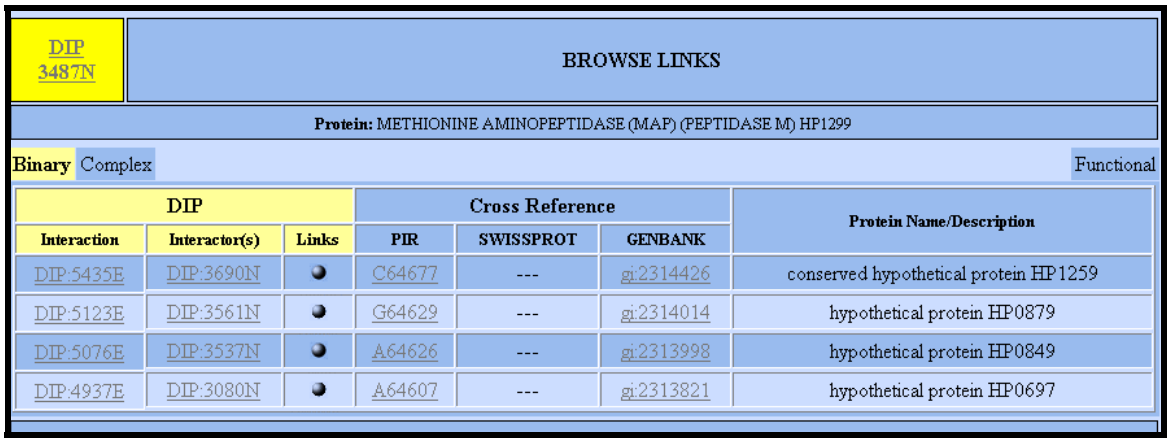


Figure 6. DIP interaction network for protein: 2314463

In the network image, the starting node or the query protein is represented as a red circle. All of the first-shell edges and neighbors of the query protein are drawn. The first shell nodes are represented as orange circles. The second-shell neighbors of the query protein are represented as yellow circles. Only the edges linking the first shell nodes to the second shell nodes are drawn. Edges between the second shell nodes are omitted for clarity. The thickness of the edges is used to represent the number of independent experiments identifying the interaction. The edges are also color coded to represent the reliability of the interaction evidence such that green is used to represent core interactions whereas red is used to represent the unverified results of high-throughput screens. The nodes trace back to the DIP node information page for that node (Xenarios *et al*., 2002). Information about the immediate neighbors of the query protein is also available in a tabular format as shown in fig. 7. By clicking the "Links" section of the DIP node information page, one can obtain a list of the first-shell interacting neighbors of the query protein. Each interaction is given a DIP edge identification number.



DIP 3487N

BROWSE LINKS

Protein: METHIONINE AMINOPEPTIDASE (MAP) (PEPTIDASE M) HP1299

Binary Complex Functional

DIP			Cross Reference			Protein Name/Description
Interaction	Interactor(s)	Links	PIR	SWISSPROT	GENBANK	
DIP:5435E	DIP:3690N		C64677	---	gi:2314426	conserved hypothetical protein HP1259
DIP:5123E	DIP:3561N		G64629	---	gi:2314014	hypothetical protein HP0879
DIP:5076E	DIP:3537N		A64626	---	gi:2313998	hypothetical protein HP0849
DIP:4937E	DIP:3080N		A64607	---	gi:2313821	hypothetical protein HP0697

Figure 7. DIP interaction information in tabular format

We downloaded interaction information from DIP in the XIN format. XIN is based on the XML format and can be used to describe an arbitrary annotated graph (Xenarios *et al., 2002)*. Fig. 8 below illustrates the node and edge information stored in XIN format. Each protein node has a unique DIP ID of the form DIP: #N and a node id of the form G: #, where # is a number. Similarly, each edge or interaction has a unique identifier of the form DIP: #E. The edge is defined in term of the node id as being from G:x to G:y, where x and y are numbers. Each node provides cross-reference information, organism source, a taxon ID and a short description of the protein. Each edge contains links to the experimental evidence.

```
<node id="G:1" uid="DIP:232N" name="BAXA_HUMAN"
class="protein">
  <xref db="DIP"  id="232N" type="src"/>
  <feature name="swp_ref" class="cref">
   <src>SwissProt</src>
   <val>SWP:Q07812</val>
   <xref db="SWP"  id="Q07812" type="src"/>
  </feature>
  <feature name="pir_ref" class="cref">
   <src>PIR</src>
   <val>PIR:A47538</val>
   <xref db="PIR"  id="A47538" type="src"/>
  </feature>
  <feature name="gi_ref" class="cref">
   <src>NCBI</src>
   <val>GI:539664</val>
   <xref db="gi"  id="539664" type="src"/>
  </feature>
  <feature name="refseq_ref" class="cref">
   <src>RefSeq</src>
   <val>RefSeq:NP_620116</val>
   <xref db="RefSeq"  id="NP_620116" type="src"/>
  </feature>
  <att name="descr">
   <val>bcl-2-associated protein x, alpha splice form</val>
  </att>
  <att name="organism">
   <val>Homo sapiens</val>
   <xref db="TXID"  id="9606" type="ont"/>
  </att>
 </node>

<edge uid="DIP:1E" id="G:3" from="G:1" to="G:2" class="link">
  <xref db="DIP"  id="1E" type="src"/>
  <feature uid="DIP:1X"  name="evidence" class="exp:s">
   <src>PMID:9194558</src>
   <val>Experimental</val>
   <xref db="DIP"  id="1X" type="src"/>
   <xref db="DO"  id="DO:00045" type="ont"/>
   <xref db="PSI"  id="MI:0045" type="ont"/>
  </feature>
 </edge>
```

Figure 8. DIP data format for a node and an edge

Sequence information for all the proteins contained in DIP was downloaded in the FASTA format.

Biomolecular Interaction Database (BIND)

BIND is hosted at the Mount Sinai Hospital in Toronto. The information is obtained from high-throughput data submissions and manually curated information obtained from scientific literature. It contains records of molecular associations – interactions between molecules, molecular complexes formed from one or more interactions and pathways defined by a sequence of two or more interactions (Bader *et al*., 2001). BIND includes interaction information for various biological molecules such as proteins, DNA, RNA, ligand, molecular complex, gene and photon. Table 2 shows the amount of protein-protein interaction data available in BIND as of March 2006.

Table 2: BIND database statistics obtained from the BIND website

Number of proteins	52467
Number of organisms	1566
Number of interactions	83517
Number of data sources (including articles)	23376

We used the same protein - Methionine Aminopeptidase (GI: 2314463) as a query for BIND. As they provide different options such as: BIND ID, PubMed ID, ID's of various databases it was easy to search for the protein as compared with DIP. Fig. 9 shows the interaction results obtained. BIND uses OntoGlyphs to represent annotation information. OntoGlyphs are pictorial representation of gene ontology information. As seen in the box within fig. 9, OntoGlyphs provide a "visual approach for automated retrieval and representation of gene product annotation from various datasets" (Alfarano *et al*., 2005). In addition to the interaction information, they provide links to the experimental evidence, the binding sites on the proteins and the authors of the published work.

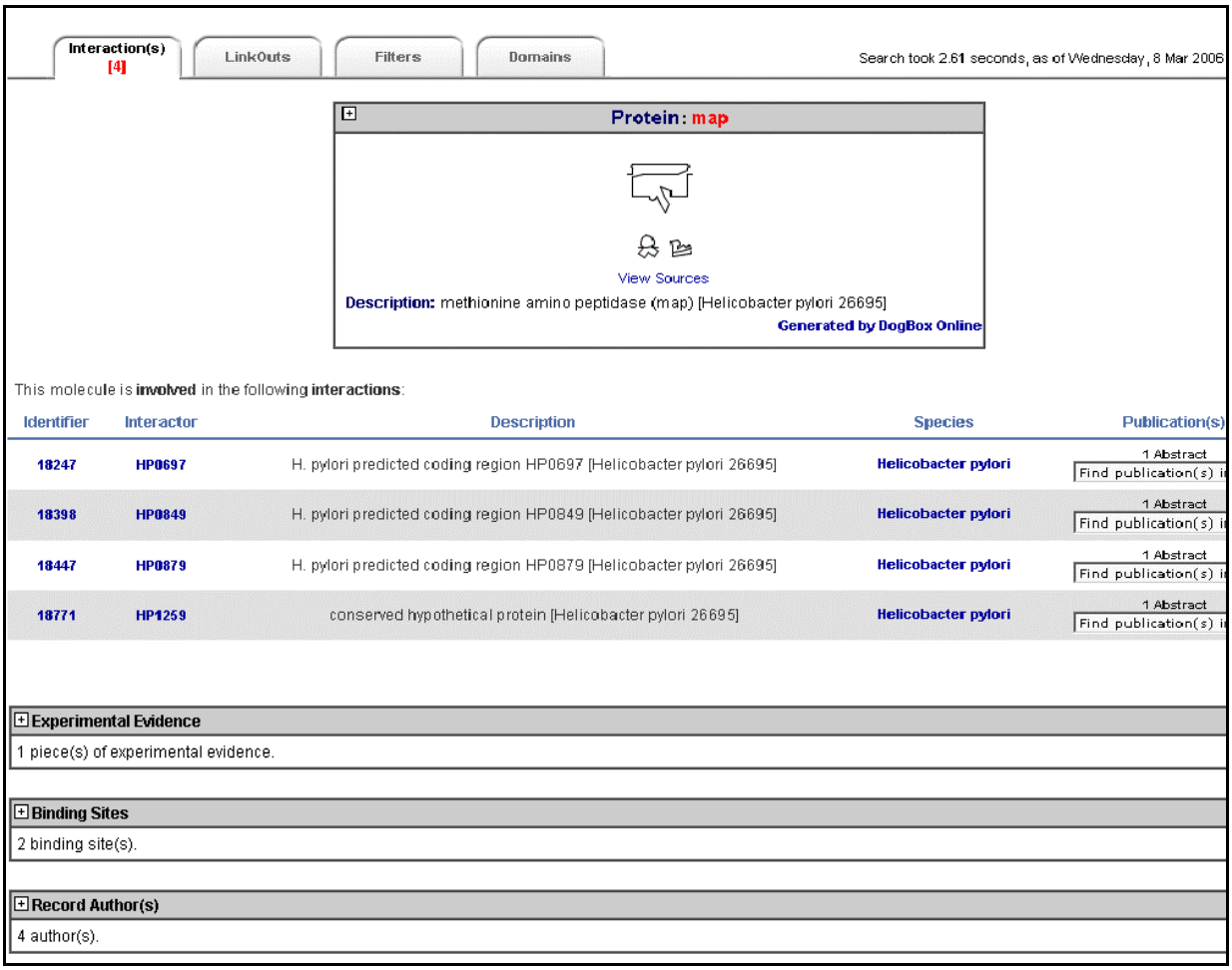


Figure 9. BIND results for protein query

Each interaction is identified by a BIND identifier. Clicking on the identifier links provides additional information about the interaction (Fig. 10). A description of the interaction together with links to the publications from which this interaction was identified is provided. For each of the proteins in the interaction they provide a short description, source organism, links to different databases such as NCBI, SwissProt, UniProt and SeqHound (BIND's sequence database). They also include automatically annotated information about cross references to GenBank, RefSeq, Swiss-Prot, and tigr, GO terms and protein Domains.

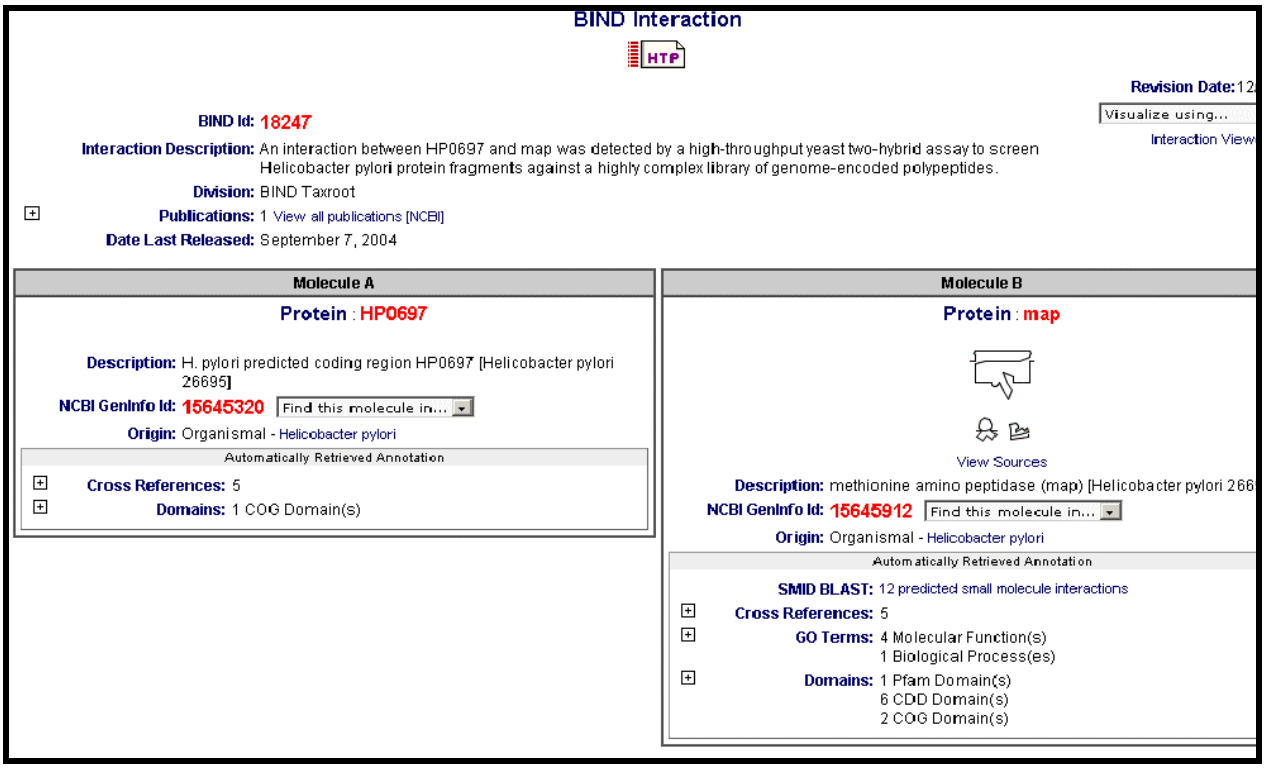


Figure 10. BIND interaction information

Each interaction can also be visualized by downloading and installing the BIND Interaction Viewer. This is a Java based application. Fig. 11 shows the visualization for the above interaction using BIND's Interaction Viewer. Each protein is represented as a rectangular node. The node contains the OntoGlyphs defining the protein's function – this feature can be turned on or of. The viewer provides a legend for the various OntoGlyphs used to represent a molecule's function, localization and binding. Clicking on a node provides information about that node – the name of the protein, the number of interactions it is involved in, synonyms and a short description. For a given node one can also display all of its interactions where available. Fig. 12 shows all the interactions available for the query protein – Methionine aminopeptidase. The query protein is highlighted in red.

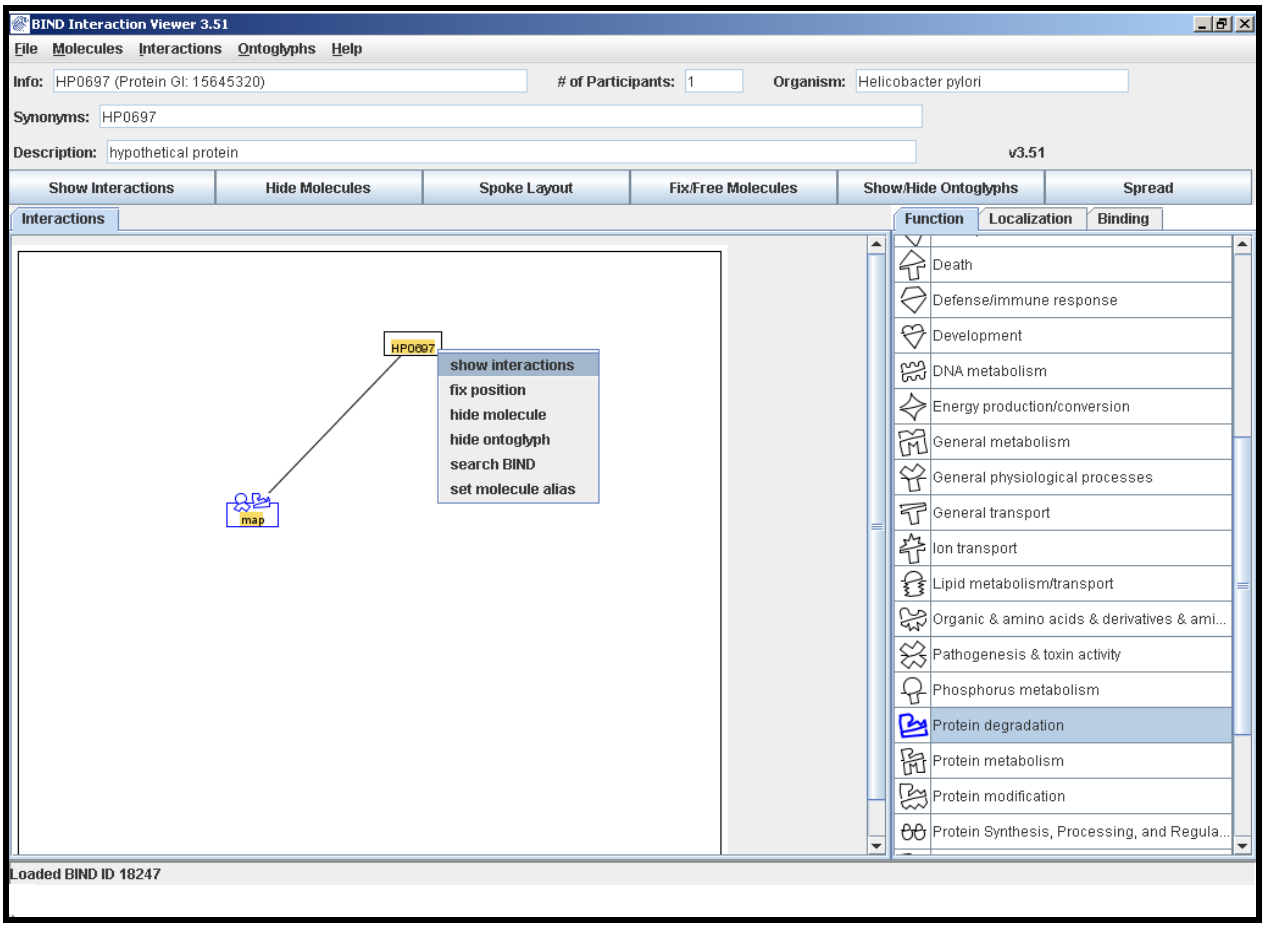


Figure 11. Bind Interaction Viewer

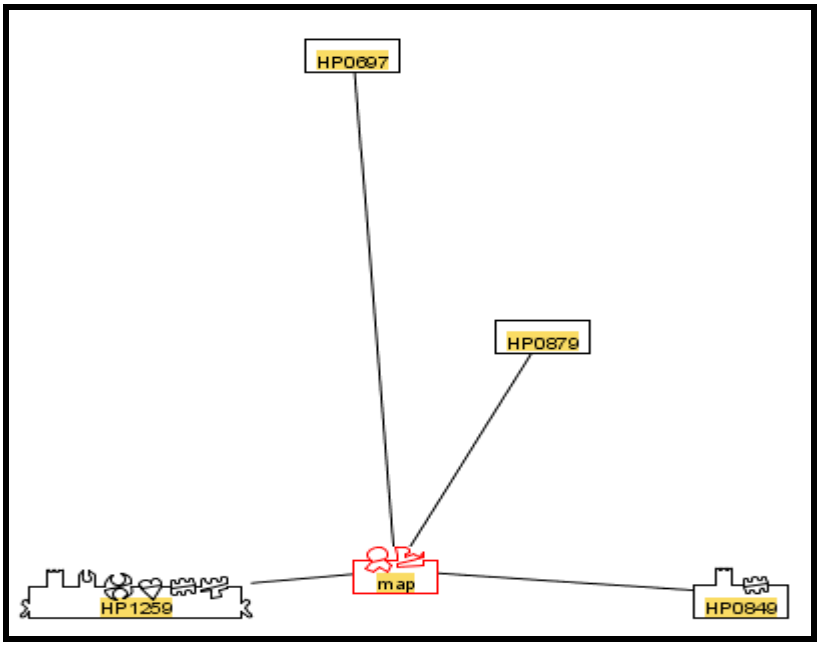


Figure 12. BIND Interaction Viewer showing all the interactions for GI 2314463

The non-redundant list of interaction information from BIND was downloaded in the available tab-delimited format (Fig. 13). This dataset contains information about all the different kinds of interaction information present in BIND. Hence, it includes interactions between proteins and other molecules such as DNA, RNA as well as interactions between non-proteins. Each line represents the interaction between two molecules which we will call A and B for the purpose of this discussion. Each interaction is identified by a unique number which represents the BIND ID. Columns 2 and 7 specify the type of molecule (protein, DNA, RNA, gene, small molecule) interactors A and B are. Columns 3 and 8 denote the database source, columns 4 and 9 the accession ID for the molecules. Columns 5 and 10 contain the GI number for the molecule and columns 6 and 11 contain the taxon ID for the organism source [5]. We also download the list of taxon ID's and the associated organism's name from BIND.

1	protein	GenBank	NP_208350	2314742	85962	protein	GenBank	NP_207285	2314275	85962
2	protein	GenBank	NP_000448	31077205	9606	DNA	GenBank	NA	14249383	9606
3	protein	GenBank	Q9JKN6	33516943	10090	RNA	GenBank	NM_025683	13385145	10090
4	protein	GenBank	NA	0	9606	protein	GenBank	NP_006779	15341887	9606
5	DNA	GenBank	1NK8|C	47168401	0	protein	GenBank	P52026	3041672	1422

Figure 13. BIND interaction data format

As the interaction data does not contain experimental evidence information, we download the entire BIND experimental system dataset in the FASTA format. This dataset consists of FASTA formatted files for each experiment such as Two-hybrid test, Immunoblotting, etc. The files contain the sequences of all the molecules whose interaction was discovered using that experimental system. We also download a FASTA file containing the non-redundant list of sequences of the proteins in BIND.

The datasets from BIND are downloaded monthly through an automated process. We only download files that have been updated since the previous download. Currently, we do not have an automated download process set up for the DIP files as DIP does not provide remote FTP access. File access in DIP is provided through the web-interface where an extremely short-life (10-20 seconds) FTP password is randomly generated for each file download. Hence, the DIP data files are manually downloaded every month.

Data Parsing

From the DIP XIN file we extract the interaction information and map each DIP node ID to its GI number where available. We also obtain the names of the experiments used to identify the interaction and the taxon ID's of the two proteins. Similarly, from the BIND interaction data file we extract all the protein-protein interactions, the GI numbers for each protein and their taxon ID's. We then parse the BIND experimental system dataset and map each interaction to its identifying experiment. We check to ensure that both proteins in an interaction are specified in the same experiment file. This data is stored in a tab-delimited text file (Fig. 14). Each line corresponds to an interaction between proteins. Columns 1 and 2 contain the GI numbers of the proteins, columns 3 and 4 their taxon ID's, column 5 contains the experiment names separated by "*" and column 6 specifies the source of the interaction – D for interactions from DIP, B for BIND and D*B for those present in both. Table 3 shows the statistics for the PIN-IT data file as of March 2006.

101452	723899	4932	4932	in vitro binding*tandem affinity purification*	D
24644797	23397623	7227	7227	two hybrid test*	D
2314567	2313643	85962	85962	two hybrid test*	D*B
47168915	47168915	1426	1426	not specified*	B

Figure 14. PIN-IT interaction data format

Table 3. PIN-IT data statistics

Number of proteins	43753
Number of organisms	1033
Number of interactions	94659

BLAST

A significant measure of PIN-IT is the expansion of known protein interaction networks through the addition of similar proteins. Sequence alignments provide a powerful way to compare novel sequences with previously characterized genes. Both functional and evolutionary information can be inferred from well designed queries and alignments. We carry out sequence alignments with the use of the Basic Local Alignment Search Tool.

BLAST is used to compare sequences – amino acid sequences of proteins or nucleotide sequences of DNA and determine regions of similarity. It can be used to either compare two sequences or to compare a query sequence against a database or library of sequences. BLAST emphasizes regions of local alignment to detect relationships among sequences which share only isolated regions of similarity (Altschul et al., 1990). Sequence alignments are determined using a heuristic approach to the Smith-Waterman algorithm and the statistically significant alignments are then displayed to the user.

In the first stage, BLAST searches for exact matches of a small fixed length W between the query and the sequences in the database. For example, given the sequences in figure 14 and a word length W=3, BLAST would identify the substring TTA that is common to both the sequences. In the second stage, BLAST tries to extend the match in both directions starting at the seed. The un-gapped alignment process extends the initial seed match length of W in each direction in an attempt to boost the alignment score. Insertions and deletions are not considered during this stage. For our example, the un-gapped alignment between the sequences AGTTAC and ACTTAG centered on the common word TTA would now include A. If a high scoring un-gapped alignment is found, the database sequence is passed on to the third stage. In the third stage, BLAST performs a gapped alignment between the query sequence and the database sequence using a variation of the Smith-Waterman algorithm. Statistically significant results are then displayed to the user. One of the statistical score used is the E-value or the expectation value. It represents the number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. Thus, the lower the E-value, the more significant the score. We use the E-value to identify similar proteins.

Figure 15. BLAST sequence alignment

We obtained the stand-alone version of BLAST from the NCBI website. We use the BLASTP program to search our protein sequence files for similar proteins.

Algorithm to Identify the Inference Interaction

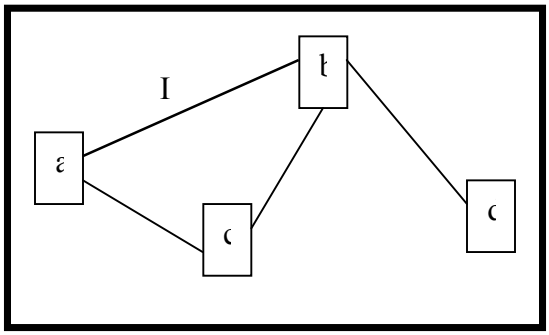


Figure 16. Interaction edge between nodes

Using the interaction data for a given query protein, we construct a network where each protein is represented as a node and the edges between the nodes are based on the experiments identifying the interaction between them (Fig. 16). Each experiment is associated with a false-positive rate. This is a guesstimated value ranging from 0.0 to 1.0. The values were determined by Dr. Haixu Tang based on his knowledge of the accuracy of these experiments. Table 4 lists the different experiments and their associated false-positive rates.

Table 4. Experiments and their associated false-positive rates

EXPERIMENT NAME	FALSE POSITIVE VALUE	EXPERIMENT NAME	FALSE POSITIVE VALUE
Adhesion Assay	0.35	Hybridization	0.6
Affinity Chromatography	0.35	Immunoblotting	0.1
Alanine Scanning	0.35	Immunofluorescence	0.1
Atomic Force Microscopy	0.35	Immunolocalization	0.1
Autoradiography	0.35	Immunoprecipitation	0.1
Biochemical	0.9	Immunostaining	0.1
Biophysical	0.9	In Vitro Binding	0.6
Calcium Mobilization Assay	0.35	In Vivo Kinase Activity	0.35
Chemotaxis	0.35	Ion-Exchange Chromatography	035
Co Immunoprecipitation	0.1	Lambda Fusion	0.35
Co Purification	0.1	Light Scattering	0.35
Co-localization	0.35	Mass Spectrometry	0.6
Competition Binding	0.1	Membrane Filtration	0.6
Cosedimentation	0.35	Microarray	0.6
Cross Linking	0.1	Microtiter Plate Binding Assay	0.35

Denaturing Gel Electrophoresis	0.6	Monoclonal Antibody Blockade	0.35
Density Gradient Sedimentation	0.35	Native Gel Electrophoresis	0.35
Deuterium Hydrogen Exchange	0.35	NMR	0.1
Electron Microscopy	0.1	Not Specified	0.9
Electron Spin Resonance	0.1	Other	0.9
ELISA	0.35	Phage Display	0.6
Enhancement Test	0.35	Resonance Energy Transfer	0.35
Equilibrium Dialysis	0.35	Split-Ubiquitin system	0.35
Experimental	0.9	Surface Plasmon Resonance	0.35
Far Western	0.35	Synthetic Lethal/Sick Test	0.35
Filter Overlay Assay	0.35	Tandem Affinity Purification	0.35
Fluorescence Anisotropy	0.35	Three Dimensional Structure	0.1
Footprinting	0.35	Transcription Assay	0.35
FRET Analysis	0.35	Transient Coexpression	0.35
Gel Filtration Chromatography	0.35	Two Hybrid Test	0.6
Gel Retardation Assay	0.35	X-Ray Crystallography	0.1
Genetic	0.9	X-Ray Diffraction	0.35
Gradient Sedimentaion	0.35	X-Ray Scattering	0.35

The edge weight between two interacting proteins a and b is calculated as:

$$I_{(a,b)} = 1 - \prod r_{\exp(a,b)}$$

Where r is the false-positive rate for the experiment's identifying the interaction.

The interaction network is now extended through BLASTP to include proteins similar to each of the protein in the interaction network. The addition of proteins is limited by the e-value chosen by the user as well as their desire to include proteins from different organisms. The newly added similar proteins are connected to their "parent" protein through edges based on the e-value (Fig. 17). The edge weight between a protein "a" and protein "a1" which is similar to it is computed as:

$$S_{(a1,a)} = E_{(a1,a)} / n$$

where E is the e-value obtained from BLAST and n is the size of the database used for BLAST.

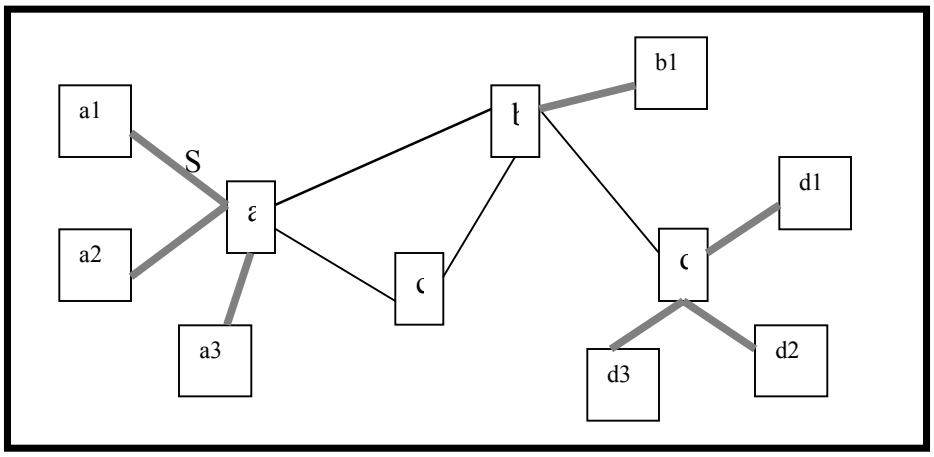


Figure 17. Similarity edges added through BLASTP

Using the interaction edge weight and the similarity edge weight, we determine the inference interaction (Fig. 18a). Given two interacting proteins a and b, such that protein a1 is similar to protein a, and protein b1 is similar to protein b, the interaction distance between a1 and b1 is computed as follows:

$$D_{(a1,b1)} = S_{(a1,a)} \times I_{(a,b)} \times S_{(b1,b)}$$

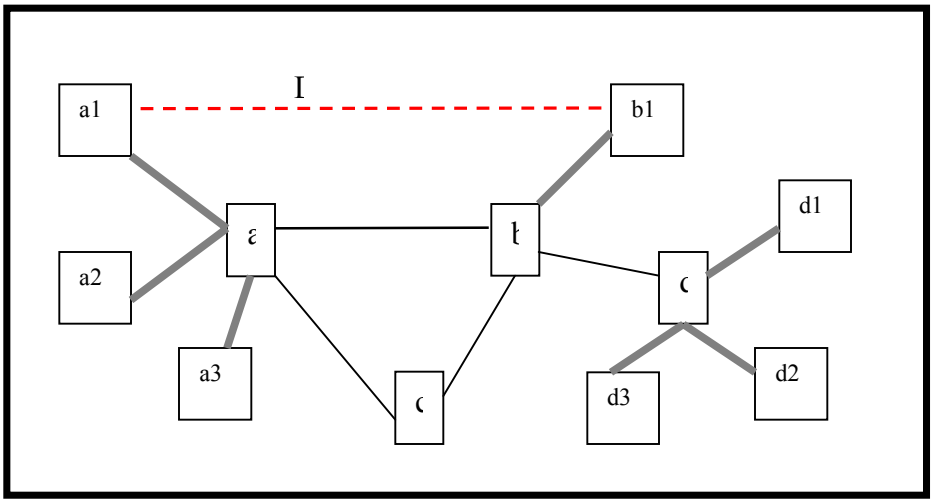


Figure 18a. Inference edge inferred between proteins a1 and b1

In addition to looking at proteins that have similar proteins (proteins a and b in figure 18a), we also look at cases where one of the interacting proteins does not have proteins similar to it as in protein a in figure 18b. We use the edge weight between proteins a, b and b1 (figure 18b) to compute the interaction distance.

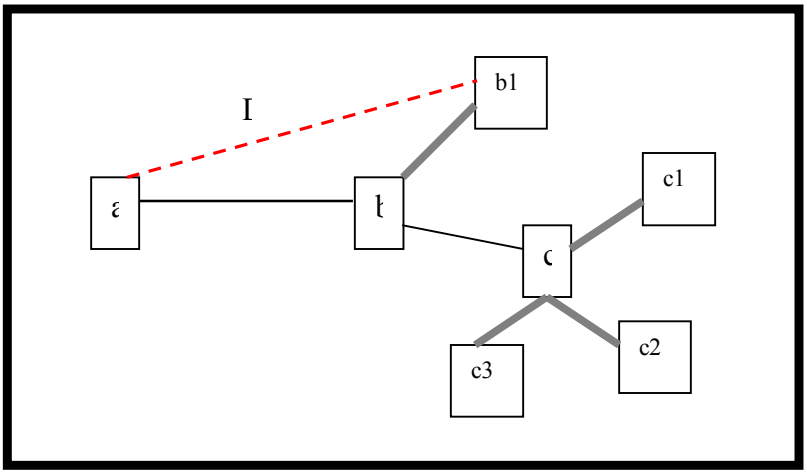


Figure 18b. Inference edge inferred between proteins a and b1

We compare all the similar proteins between each pair of interacting partners to identify the two proteins with the maximum interaction distance.

$$\text{Suggest } (a^*, b^*) = \arg\max_{\{a,b\}} D_{(a,b)}$$

We hypothesize that the two proteins with the maximum interaction distance are most likely to interact.

Extending the Network

The generated protein interaction network can easily be extended to include interactions from a different protein in the network. When the user clicks on the "extension" portion of a node, we re-generate the entire protein network starting from that node. The extension portion is the "e" section of each node in the protein interaction network (Figure 19). The same BLAST e-value threshold and the choice of including proteins from different organisms as in the original query are used. We retain the original interactions and similar proteins and generate a new network image combining the two. The inference-distance is re-computed for this new extended network.

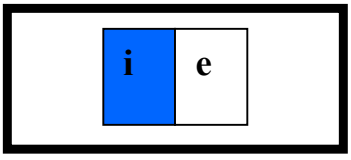


Figure 19. A query node representation showing the extension portion

Online Implementation

The Protein Interaction Network – Inference Tool (PIN-IT) is available at *http://homer.informatics.indiana.edu/pinit/*.

The tool is implemented in Perl & CGI. JavaScript was used to create the pop-up information for the nodes and edges. The data is automatically downloaded on the 15th of every month. We check the source websites – DIP & BIND to see if they have updated their data files since the previous download. Files are downloaded only if newer ones are available at the source. The downloaded files are saved as text files. The interaction information is parsed from DIP & BIND and saved to a text file – data.txt (Fig.14). This file contains a non-redundant list of interactions, taxon ID's and experimental evidence. The FASTA format sequence files are combined to form sequence.fsa. This file is formatted for BLAST by running the formatdb command. The tool contains a page – Datasets which is automatically updated each time the data files are downloaded (Fig.20). The page contains statistic information for each of the downloaded datasets as well as the combined dataset used by PIN-IT.

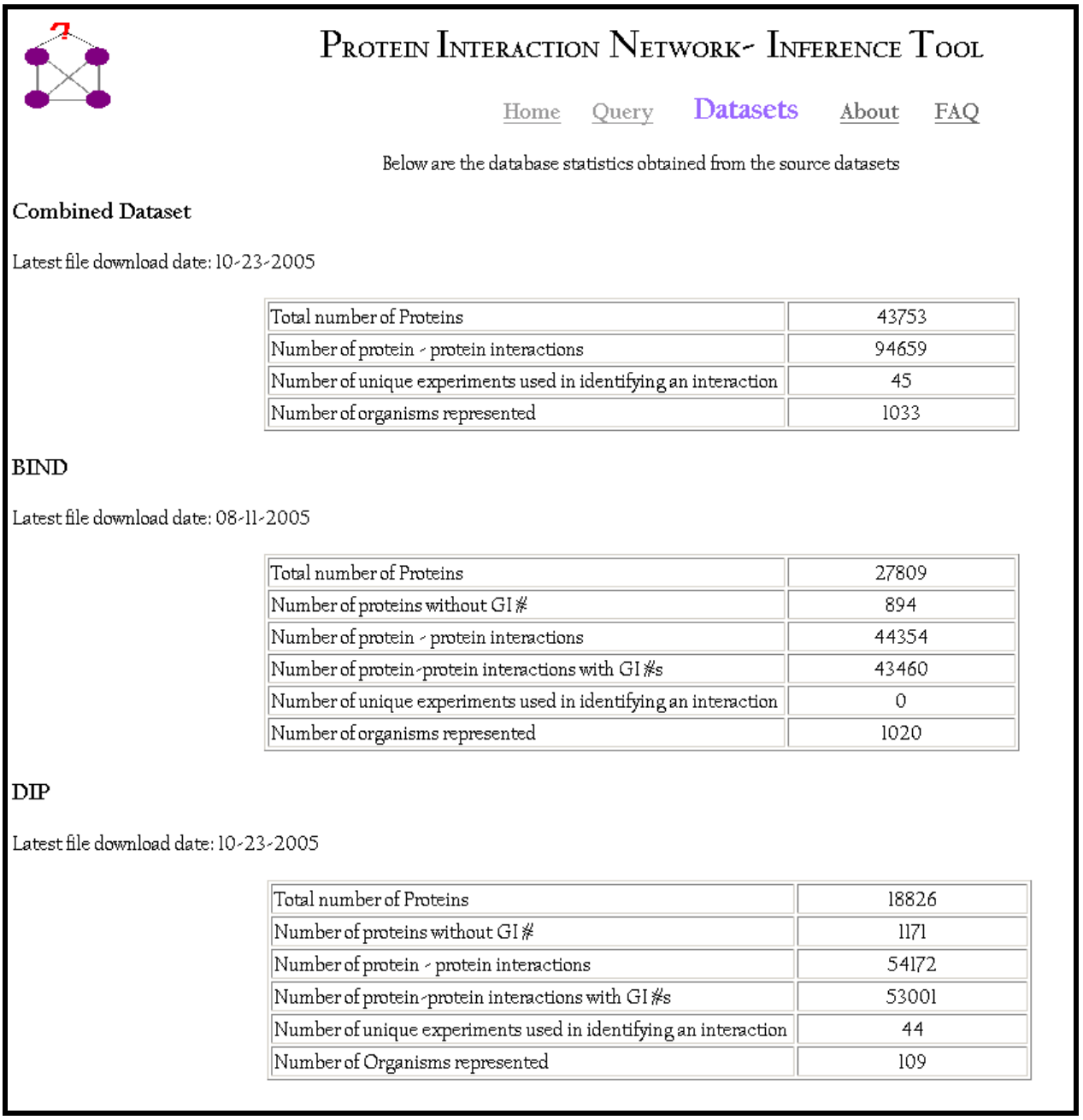


Figure 20. PIN-IT Datasets page

Using the combined interaction data we create two tables – one that maps protein ID's to the taxon ID of their source organism and the other that maps the GenBank or reference ID to the NCBI ID. These tables are stored as BerkeleyDB files in order to have efficient access. We also use a BerkeleyDB file to store the results each time a user searches for a protein. This greatly

minimizes the time required to generate the results when the same protein is queried in the future.

PIN-IT has a very simple interface to search for protein interactions. The user is required to search using the protein's GenBank ID. This identifier was used as it is one of the most universally used identifier and also because it is unlikely to be replaced by something else. We queried PIN-IT with the same query protein – Methionine aminopeptidase (GI: 2314463) and obtained the results as shown in fig.21. The results obtained do not include proteins from other organisms and an e-value of 1e-2 is used.

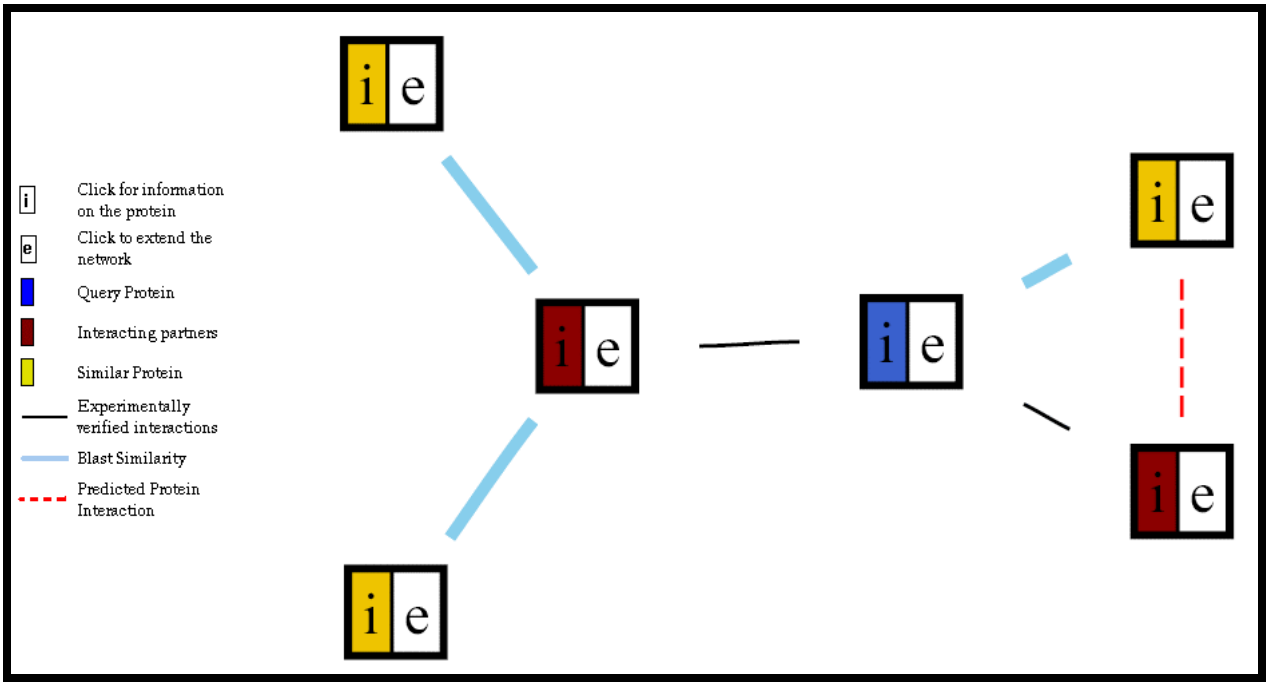


Figure 21. PIN-IT query results

All the proteins are represented as rectangular nodes consisting of two parts – the Information part - "i" and the Extension part – "e". The color of the information portion of each node is used to represent the proteins – query, interactors and similar proteins. The query node is represented in blue, its interactors in red and all the similar proteins in yellow. Clicking on the "i" portion of any node opens up a new window linking to the NCBI information for that protein (Fig. 22).

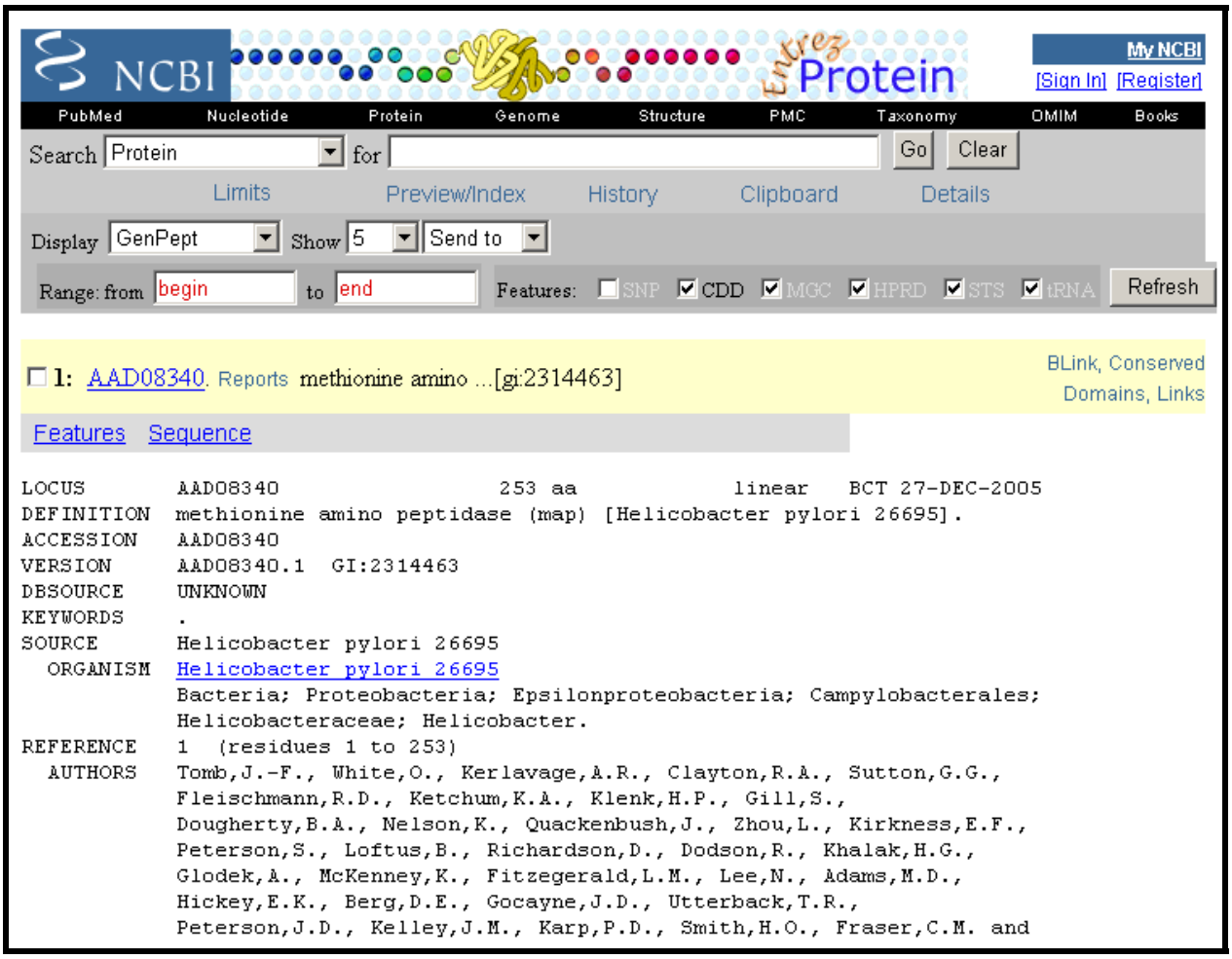


Figure 22. PIN-IT node information from NCBI

The interacting proteins are connected with a black edge between them. A protein and its similar protein are connected with a thick light blue edge. A darker blue color is used to connect proteins whose similarity is higher. The inference edge is represented with a dashed red line. Clicking on each edge provides information about that edge. Clicking on an interaction edge between two proteins provides information about the experimental evidence used to identify the interaction (Fig. 23). Clicking on a similarity edge provides information about the e-value obtained from BLAST for the two proteins connected by that edge (Fig. 24). The inference edge shows how the inference distance was calculated for that pair of proteins (Fig. 25). The inference edge is drawn between the pair of proteins whose inference distance was the maximum.

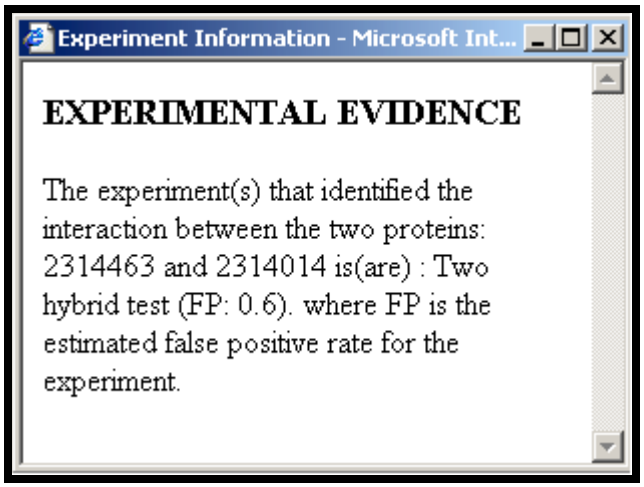


Figure 22. PIN-IT Interaction Edge information

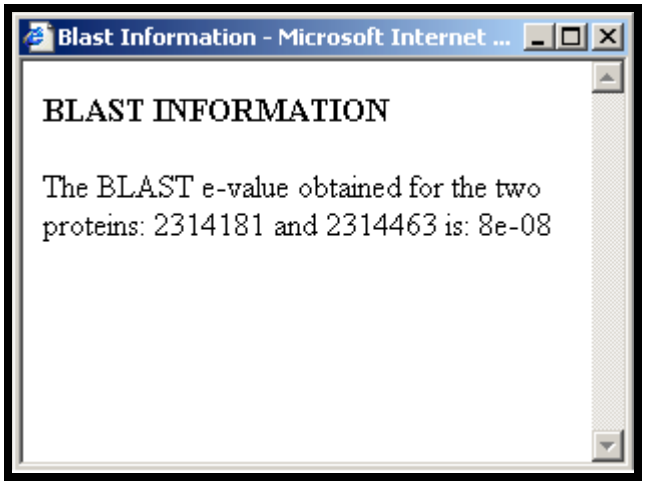


Figure 23. PIN-IT Similarity Edge information

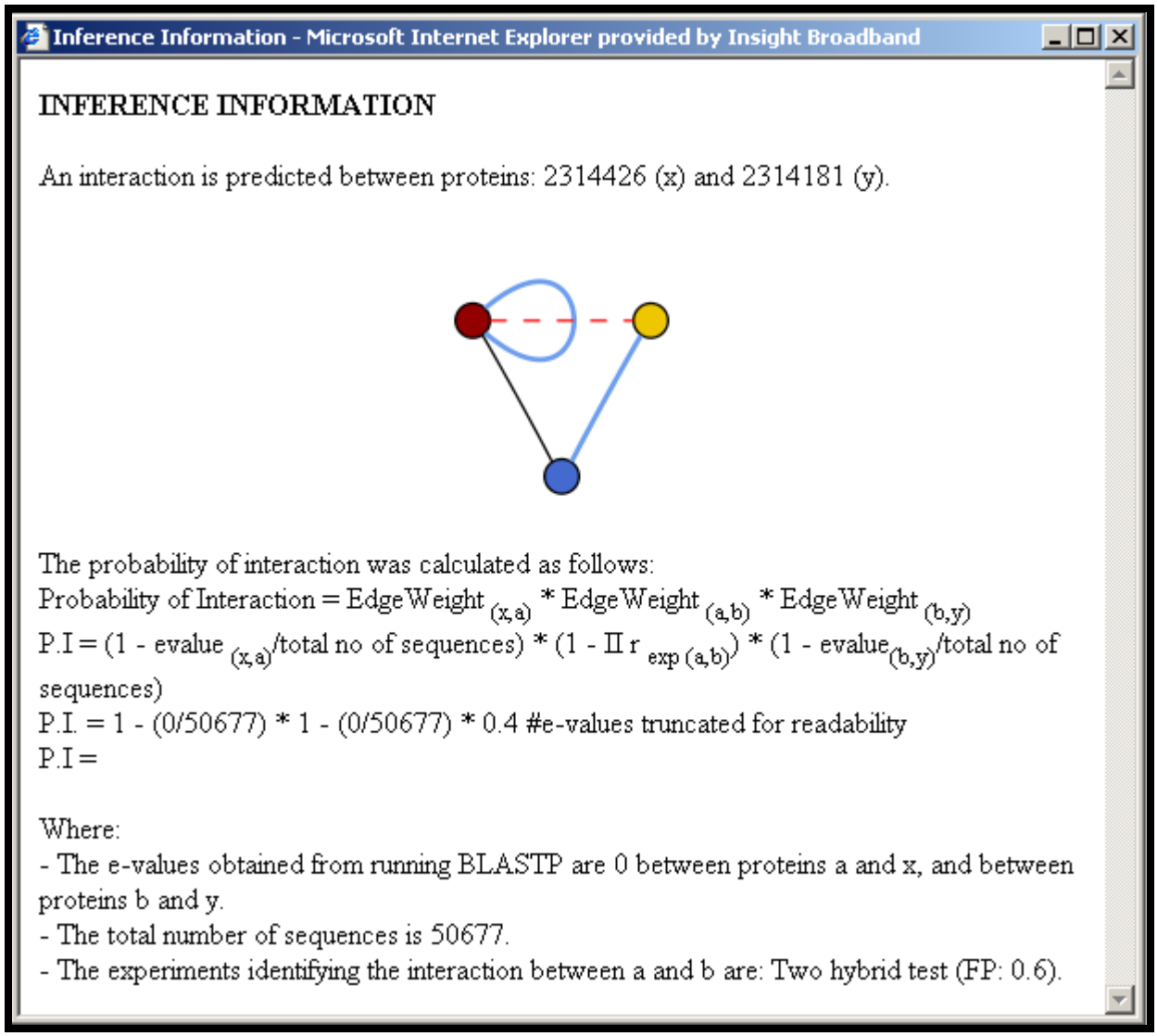


Figure 24. PIN-IT Inference edge information

The interaction network was extended by clicking on the protein similar to the query protein (GI: 2314181). A new interaction network was obtained (Fig. 25). This network did not have an inference protein due to the lack of similar proteins between interacting pairs.

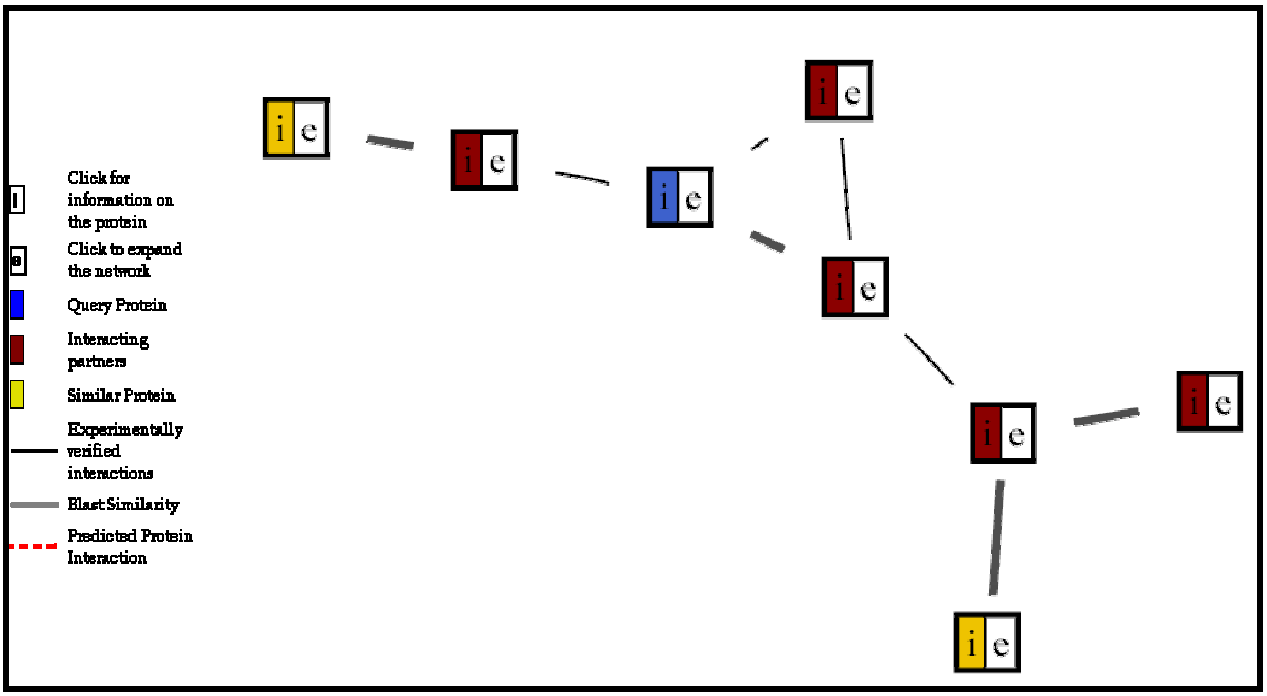


Figure 25. PIN-IT Extended interaction network

The network images are rendered in SVG format. The SVG format was chosen as it is compact, allows for high quality graphics and allows for interactivity which is essential when interaction images such as the one in fig. 26 are obtained. Fig. 26 shows the interaction network obtained for the Tat Human Immunodeficiency virus 1 (GI: 9629358). The image was rendered small in order to show all the interactions. SVG allows the user to zoom in or zoom out, in order to obtain a better view of the interactions. Some of the other interactive features available in SVG images are shown in the table 5 below.

Table 5. SVG image manipulation features

CURSOR	FUNCTION
	Click nodes to learn more
	Resize window to scale map
	CTRL-click to zoom in

	SHIFT-CTRL-click to zoom out
	ALT-click to scroll
	RIGHT-click for a menu

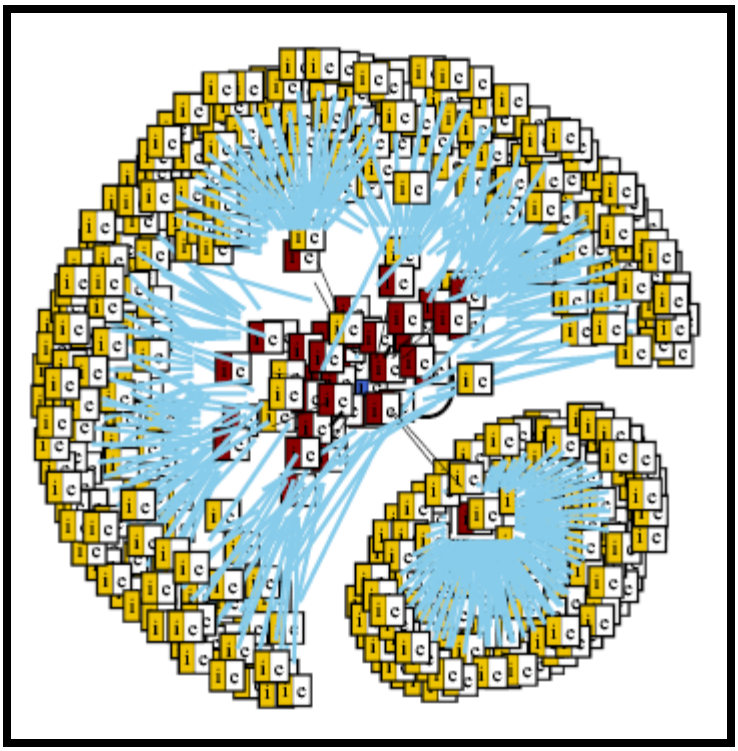


Figure 26. PIN-IT interaction network for protein GI: 9629358 with an E-value limit of 1e-5

RESULTS

Protein Interaction Network – Inference Tool predicts uncharacterized protein-protein interactions based on known protein interactions and the sequence similarity between proteins. The difficulty faced with this tool along with the other protein interaction prediction tools such as InterWeaver and ADVICE is a method to validate the predictions. The best way to validate the predictions would be to carry out biological experiments to verify the predictions. This requires expertise in molecular biology and time. However, in the meantime we can validate the predictions made by analyzing the proteins that have been predicted to interact.

We tested PIN-IT using as query the putative integral membrane E3 ubiquitin ligase (GI: 6324320) as query. From the interaction network that was obtained, an inference interaction was drawn between proteins GI: 6321260 and GI: 6323766. Looking up these proteins in the NCBI database, we found that protein 6321260 is a protein that specifically binds to mRNAs encoding chromatin modifiers and spindle pole body components, and has roles in longevity, in maintenance of cell wall integrity, and in sensitivity to and recovery from pheromone arrest. Protein 6323766 is a putative integral membrane E3 ubiquitin ligase whose genetic interactions suggest a role in negative regulation of amino acid uptake. Thus it can be seen that both the proteins 6321260 and 6323766 are highly likely to interact based on their functions and locations in the cell.

We carried out another sample query with protein GI: 6323144 which is a subunit of the mRNA cleavage and polyadenylation factor (CPF). This protein is required for pre-mRNA cleavage, polyadenylation and poly(A) site recognition. From the interaction network, an inference interaction was found between the query protein (GI: 6323144) and the protein GI: 6323307. This protein 6323307 is a putative endonuclease subunit of the mRNA cleavage and polyadenylation specificity complex and is required for 3' processing of mRNAs. The prediction indicates that these two proteins which are involved in the processing of mRNA's interact – which is highly likely.

DISCUSSION

Various online databases such as InterWeaver (Zhang 2004), ADVICE (Tan 2004), PRISM (Aytuna 2005) and STRING (von Mering 2005) predict putative protein-protein interactions per-computed using various computational methods. PIN-IT differs from these approaches as it utilizes sequence homology in addition to the experimental and literature data to predict protein-protein interactions. The use of false-positive rates and sequence similarity values provide a higher level of confidence in predicting putative interactions.

As this is only the first iteration in the development of PIN-IT, there are certain features that will be improved in the future. One of the problems with regards to the usability of PIN-IT is

the time taken to generate results. Whenever a new query is issued the program can take a long time to present the user with the interaction network. This is dependent on the number of interactions the protein is involved in and the number of similar proteins available for each of the proteins in the interaction network. In order to minimize this processing time, we set up the BerkeleyDB files for looking up taxonomic information, and mapping NCBI ID's. However, the most time intensive step in the entire process is the generation of the Perl Graph containing all the interactions and parsing that graph to find the first-degree neighbors and the connections between them.

Other features that can be added to PIN-IT include:

a. Providing different search options such as the PIR ID, Swiss-Prot ID in addition to the NCBI GI number.
b. Allowing for the user to upload a protein sequence and include it in constructing an interaction network to determine whether it is similar to one of the proteins in a known interaction.
c. Allowing for the user to upload new interactions and run queries.

ACKNOWLEDGMENTS

We thank Dr. Sun Kim for pointing us in the direction of protein interaction networks. We thank Dr. Alessandro Flammini and Dr. Haixu Tang for providing helpful feedback – especially with the biologically relevant aspects of the project. We are grateful to the people who created and maintain DIP & BIND – the backbone of this tool.

REFERENCES

- Alfarano C. *et al.* (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucl. Acids Res.* **33**(Database Issue):D418-D424
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**(3):403-10, 1990
- Aytuna,A.S, Gursoy,A. Keskin,O., (2005). Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics* **21**(12):2850-2855.

- Bader G.D., Donaldson I., Wolting C., Ouellette B.F., Pawson T., Hogue C.W. (2001). BIND--The Biomolecular Interaction Network Database. *Nucleic Acids Res* 29(1):242-5
- Bader, G.D., et al. (2003) BIND: the biomolecular interaction network database. *Nucleic Acids Res.*, **31**, 248–250
- Gavin, A.C., et al. (2002) Functional organization of the yeast genome by systematic analysis of protein complexes. *Nature*, **415**, 141–147
- Giot, L., Bades, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., *et al*. (2003). A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727-1736.
- Goehler, H., Lalowski, M., Stelzl, U., Waelter, S., Stroedicke, M., Worm, U., Droege, A., Linderberg, K.S., Knoblich, M., Haenug, C., et al (2004). A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol. Cell* **15**, 853-865.
- Ito T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
- Kanehisa M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
- Krieger C.J., Zhang,P., Mueller,L.A., Wang,A., Paley,S., Arnaud,M., Pick,J., Rhee,S.Y. and Karp,P.D. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **32**, D438–D442.
- Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., *et al*. (2004). A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540-543.
- Lu, L., et al. (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein–protein interactions by multimeric threading. *Proteins*, **49**, 350–364
- Marcotte, E. M., Matteo, P., *et al.* (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*. **285**: 751-753
- Rain, J.C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Peterl, F., Wojick, J., Schachter, V. *et al*. (2001). The protein-protein interaction map of *Helicobacter pylori*. *Nature* **409**, 211-215.

- Salwinski, L. and Eisenberg, D. (2003) Computational methods of analysis of protein–protein interactions. *Curr. Opin. Struct. Biol.*, **13**, 377–382
- Tan, Soon-Heng., Zhang, Zhuo., Ng, See-Kiong. (2004). ADVICE: Automated Detection and Validation of Interaction by Co-Evolution. *Nucl. Acids Res*. **32**: W69-72
- Uetz P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
- von Mering C, Jensen, L.J., Snel, B., Hoper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., Bork, P. (2005). STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucl. Acids Res.* **33** (Database Issue):D433-437.
- von Mering C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.
- Xenarios I., Rice D.W., Salwinski L., Baron M.K., Marcotte E.M., Eisenberg D. (2000) DIP: The Database of Interacting Proteins. *Nucl. Acids Res.* **28**:289-91
- Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M., Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucl. Acids Res.* **30**: 303-305
- Zhang, Zhuo, Ng, See-Kiong. (2004). InterWeaver: interaction reports for discovering potential protein interaction partners with online evidence. *Nucl. Acids Res*. 2004 **32**: W73-75

WEBSITES

- Protein Interaction Network – Inference Tool (PIN-IT): *http://homer.informatics.indiana.edu/pinit*
- InterWeaver: *http://interweaver.i2r.a-star.edu.sg/*
- Protein Interactions by Structural Matching (PRISM): *http://gordion.hpc.eng.ku.edu.tr/prism/index.php*
- Automated Detection and Validation of Interaction by Co-Evolution (ADVICE): *http://advice.i2r.a-star.edu.sg/*

- Search Tool for the Retrieval of Interacting Genes/Proteins (STRING): *http://string.embl.de/*
- Biomolecular Interaction Network Database (BIND): *http://www.bind.ca/*
- Database of Interacting Proteins (DIP): *http://dip.doe-mbi.ucla.ed*
- GenBank: *http://www.ncbi.nih.gov/Genbank/*
- Munich Information Center for Protein Sequences (MIPS): *http://mips.gsf.de/*
- National Center for Biotechnology Information (NCBI): *http://www.ncbi.nlm.nih.gov/*
- Protein Information Resource (PIR): *http://pir.georgetown.edu/*
- Swiss-Prot: *http://www.expasy.ch/sprot/sprot-top.html*
- Universal Protein Resource (UniProt): *http://www.pir.uniprot.org/*
- PRINTS: *http://umber.sbs.man.ac.uk/dbbrowser/PRINTS/*