Published online 25 October 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D999–D1008***
doi:10.1093/nar/gkm844

AtPID: *Arabidopsis thaliana* protein interactome database—an integrative platform for plant systems biology

Jian Cui[1,4], Peng Li[1,4], Guang Li[1,4], Feng Xu[1,4], Chen Zhao[1,4], Yuhua Li[1,4,*], Zhongnan Yang[5], Guang Wang[5], Qingbo Yu[5], Yixue Li[3,*] and Tieliu Shi[2,3,4,6,*]

[1]College of Life Sciences, the Northeast Forestry University, Harbin, Heilongjiang 150040, [2]Shanghai Information Center for Life Sciences, Chinese Academy of Sciences, Shanghai 200031, [3]Bioinformatics Center, Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, [4]Daqing Institute of Biotechnology, Northeast Forestry University, Daqing, Heilongjiang 163316, [5]College of Life and Environmental Sciences, Shanghai Normal University, Shanghai 200234 and [6]Bioinformatics Center, Shanghai University, Shanghai 200444, China

Received August 15, 2007; Revised and Accepted September 25, 2007

ABSTRACT

***Arabidopsis thaliana* Protein Interactome Database (AtPID) is an object database that integrates data from several bioinformatics prediction methods and manually collected information from the literature. It contains data relevant to protein–protein interaction, protein subcellular location, ortholog maps, domain attributes and gene regulation. The predicted protein interaction data were obtained from ortholog interactome, microarray profiles, GO annotation, and conserved domain and genome contexts. This database holds 28 062 protein–protein interaction pairs with 23 396 pairs generated from prediction methods. Among the rest 4666 pairs, 3866 pairs of them involving 1875 proteins were manually curated from the literature and 800 pairs were from enzyme complexes in KEGG. In addition, subcellular location information of 5562 proteins is available. AtPID was built via an intuitive query interface that provides easy access to the important features of proteins. Through the incorporation of both experimental and computational methods, AtPID is a rich source of information for system-level understanding of gene function and biological processes in *A. thaliana*. Public access to the AtPID database is available at http://atpid.biosino.org/.**

INTRODUCTION

At the cellular level, a network of molecular interactions is representative of life. Cellular transport such as the movement of molecules and macromolecules from one location to another within cells and the formation of complex molecular structures make the properties of the network more intricate. However, all of this apparent complexity can be systematically illustrated as a simple interaction network, particularly through an understanding of protein–protein interaction (PPI) networks.

The collection of all protein interactions in an organism is typically referred to as an interactome (1). PPIs are fundamental to virtually every aspect of cellular function (2). PPI provides useful information of functional linkage between interacting partnerships within cells (3). Therefore, PPI can help to reveal signal transductions (4,5), post-translational modifications and developmental processes (6). In addition, it can serve to aid in the identification of novel regulatory components and pathways, and provide a valuable approach to understand functional specificities at the molecular level.

Sequence-based annotation efforts have led to the identification of a number of cellular components, which can be regarded as a one-dimensional annotation. Accumulated information regarding interactions, and advancement of various high-throughput technologies make it possible to generate systematical, or two-dimensional annotations (7), such as interaction maps.

*To whom correspondence should be addressed. Tel: +86 21 54922980; Fax: +86 21 5492 0143; Email: tlshi@sibs.ac.cn
Correspondence may also be addressed to Yixue Li. Tel: +86 21 54920089; Fax: +86 21 5492 0143; E-mail: yxli@sibs.ac.cn
Correspondence may also be addressed to Yuhua Li. Tel: +86 451 8219 1737; Fax: +86 451 8219 1733; E-mail: lyhshen@mail.hl.com
The authors wish it to be known that, in their opinion, the first five authors should be regarded as joint First Authors

The past several years have witnessed an exponential increase in the amount of biological data, mainly due to the development and application of high-throughput technologies, including gene expression microarrays and mass spectrometry to characterize DNA, RNA and proteins. Currently, interactomes have been created for several model organisms, such as *Saccharomyces cerevisiae* (8,9), *Drosophila melanogaster* (10), *Caenorhabditis elegans* (11) and *Homo sapiens*, among others (12). In the plant kingdom, *Arabidopsis thaliana* has been widely employed as a model organism to elucidate important biological principles. In fact, several years ago the entire sequence of the *Arabidopsis* genome was reported, and most of its genes annotated (13). However, there are still 30% of these gene products yet to be characterized because sequence homology was not effective at assigning gene function. Only a few interactions of specific protein families in *A. thaliana* have been reported (14,15), however, an enhanced understanding of PPIs can suggest important future directions for researchers to study gene/protein relationships and functions.

Meanwhile, many experimental procedures have been developed to analyze PPIs, including biochemical methods [e.g. protein affinity chromatography (16–18), affinity blotting, coimmunoprecipitation and cross-linking (19–22)]; molecular biological methods [e.g. protein probing, the two-hybrid system (23–25) and phage display (26)] and genetic methods [e.g. the isolation of extragenic suppressors and synthetic mutants (27)]. High-throughput experimental techniques have enabled us to study PPIs at the proteome scale. This is achieved via systematic identification of physical interactions among all proteins in an organism. The ever-increasing volume of PPI data is becoming the foundation for new biological discoveries. However, these data are distributed in numerous sources and it has been confirmed that some data are noisy, data quality varies significantly, and data often cannot be verified against each other. Bioinformatics and computational approaches have been used to assess the reliability of high-throughput results and to gain confidence in published data (28). The methodology can also integrate raw data into useful information and provide experimentally testable hypotheses, thereby expanding our knowledge about new mechanisms in biological processes (29–32). Other computational prediction methods based on known protein structural interactions can also be useful to analyze large-scale PPI rules. This prediction methodology evaluates interaction rules among complete genomes using protein structural interactome maps (33). Consequently, numerous researches using computational methods have been carried out to investigate gene and protein functions, PPIs and gene regulation relationships(34–39). These approaches have been applied to interactomes of *H. (40), S. cerevisiae* (41), *C. elegans* (37), *Plasmodium falciparum* (42,43), among other organisms.

However, rapidly increasing amounts of biological data generated from genome-wide and proteome technologies on modern biochemistry and molecular biology need to be well stored, compararable, organized and accessible. An appropriate repository and maintenance system for these data can facilitate future data mining and functional investigations.

AtPID was developed using *A. thaliana* as the model system for a comprehensive resource of PPIs. All data in AtPID were deposited from either manual text mining or bioinformatics predictions. This database contains 28 062 interaction pairs, of which 3866 involve 1875 proteins obtained from the literature and 800 pairs were from enzyme complexes in KEGG. In addition, bioinformatics predictions or literature surveys provided 5562 proteins with subcellular location information. Intuitive and user-friendly query interfaces have made all the features of AtPID easily accessible. This database provides invaluable resources for researchers to study PPIs and protein functions in *Arabidopsis*, data can also be used to address questions regarding gene functions and biological processes in other taxa. AtPID is a non-commercial public access database (http://atpid.biosino.org/) that provides data download services for standalone analyses or data mining, including protein interaction properties and other areas of interest in plant biology.

DATA SOURCES AND PROCESSING

Data resources for reconstruction of interaction network

The power and expressivity of any network lies largely in the data model used to represent molecular interactions. From a computational perspective, we applied uniform systematic benchmarks and statistical approaches to specifically train our PPI network for *Arabidopsis*. In addition, to assure data quality, we treated each resource separately as weighted features and reconstructed the PPI network through the proper integration of various protein interaction datasets according to the Naïve Bayesian Network theory. In this way, meaningful biological data is made available through AtPID. Here protein interaction data are generated in the following ways: experimental results are obtained from related papers in PubMed and other available databases; and data are made accessible from bioinformatics predictions. The details of interaction data generation are described below.

Manually collected protein interactions were extracted from not only thousands of published articles, but also IntAct (44), BIND (45) and TAIR databases (13). We deposited protein interaction data possessing physical evidence or experimental references related to the association between two proteins into AtPID. To ensure the reliability of these data, we also conducted a validation process. First, we mapped PPI collected from the literature lacking AGI locus identifiers to IPI (46) and removed symbols without a match. We applied uniform AGI symbols to proteins in AtPID for further analysis. We found 3866 PPI pairs involving 1875 proteins using this filtration process. Additionally, protein pairs in enzyme complexes were also inferred as a part of GSP based on the assumption that subunits in an enzyme complex have high functional association and potential physical interactions. Enzyme complexes from KEGG (47) were obtained to extract the intersection of

Table 1. Overview of GSP resources

	PPI Resources	Number of.PPI pairs	Number of proteins in PPI pairs
GSP PPI	[1] Literatures from PubMed	1259	740
	[2] InAct	1528	677
	[3] BIND	1475	538
	[4] TAIR	1073	698
	[1]~[4]	3866	1875
Protein complexes	[5] KEGG (enzyme complex)	1700	856
Total	[1]~[5]	4666	2285

[1] Manually collected protein interactions are extracted directly from thousands of published articles in PubMed. [2] InAct provides a freely available, open source database system for protein interaction data in EMBL-EBI. All interactions are derived from literature curation or direct user submissions. [3] BIND is a new resource to perform cross-database searches of available sequence, interaction, complex and pathway information. It integrates a range of component databases including Genbank and BIND, the Biomolecular Interaction Network Database. [4] TAIR provides 'Tair Protein Interaction' file by Matt Geisler at its FTP (ftp://ftp.arabidopsis.org/home/tair/Proteins/). [5] KEGG, a reference knowledge base linking genomes to biological systems and environments, provides resourceful enzyme complex information. [1]~[4] After mapping various symbols to AGI, we found 3866 PPI pairs involving 1875 proteins with literature supports. [1]~[5] combined with enzyme complexes from KEGG, the total number of GSP is up to 4666 involving with 2285 proteins.

interactions from text mining and complexes of enzymes directly garnered from the KEGG database. We subsequently used the decision tree to determine how many proteins belonging to an enzyme complex resulted in a less false positive and higher accuracy. Because many subunits or components of an enzyme complex are mapped from sequence similarity with other species or orthologs, we compared true protein interaction data to reduce noise and redundant information. Eventually, 800 unique pairs were obtained through enzyme complex after excluding the redundancies from the 3866 pairs via text mining. Consequently, a total of 4666 interaction pairs involving 2285 proteins were generated (Table 1). Such protein interaction resources, called GSP (Golden Standard Positive) are stored in AtPID and used to score the interaction network that assigns each predictive interaction pair with quantized measures.

For predicting PPIs in AtPID, we applied computational approaches, including conserved protein interactions (i.e. interologs) (48), gene expression data (49,50), genomic context (i.e. gene neighbor algorithm) (51,52), gene fusion (Rosetta Stone method) (53,54), phylogenetic profiles (55,56) and GO annotation. The optimized phylogenetic profiles were constructed and assessed using the method of Sun *et al*. (57). Orthologous PPIs in *A. thaliana* were obtained according to ortholog function transfer. Ortholog map files in Inparanoid (58) and DIP interaction data for other organisms were also collected (59). To infer *Arabidopsis* protein interactions, we mapped several model organismal (e.g. *S. cerevisiae*, *D. melanogaster*, *C. elegans* and *H. sapiens*) protein interaction data and orthologs to *Arabidopsis*. In addition, we used the atlas of *Arabidopsis* development microarray data (Acc.no: ME00319) from the TAIR database (13) to identify co-expressed genes.

Non-redundant proteins with GO annotation from the Gene Ontology Consortium were identified. These data were used to calculate the Shared Smallest Biological Processes (SSBP) value of each pair for all proteins employing GO annotation methods (40). Interacting proteins often function in the same biological process. Therefore, proteins involved in the same process are more likely to interact than proteins in distinct processes. Furthermore, proteins exhibiting high functional specificity are more likely to interact than proteins functioning in more comprehensive processes. Based on this assumption, we first identified all biological process terms shared by two proteins. Subsequently, we counted how many other proteins are assigned to each of the common terms and produced the shared biological process term with the smallest count (SSBP). In general, the smaller the SSBP count, the more specific the biological process term, and the greater the functional similarity between two proteins. In this way, we used SSBP to predict PPIs.

We also investigated the assumption that some of the operons contained within a particular organism may be conserved across other organisms based on the Gene Neighbors method. The conservation of an operon's structure provides additional evidence that genes within an operon are functionally coupled and are perhaps components of a protein complex or pathway.

Finally, we adopted the gene fusion method. The underlying assumption of the method is that if a composite protein is uniquely similar to two component proteins in another species, the component proteins are most likely to interact (53). Gene-fusion events were identified in complete genomes, based solely on sequence comparisons. These data enable the inference of functional associations among proteins.

The Bayesian Networks approach

The predictive datasets from such individual methods were integrated employing Naïve Bayesian Networks (40). The Bayesian Networks approach was used to integrate more than seven predictive data sources and to subsequently build a model to infer novel PPIs for *Arabidopsis*. The essence of the approach is to provide a mathematical rule, given some predictive evidence, to explain how to adjust the odds that a pair of proteins interacts, either in a true interaction instance (GSP) or correspondingly, in negative protein interactions, known as GSN (Golden Standard Negative). No direct information regarding the absence of specific protein interactions is available. However, protein localization data provides indirect evidence, given we assume that proteins in different cellular compartments do not interact (60). Hence, GSN values were constructed based on this assumption using subcellular localization data from the SUBA database (61). Individual likelihood ratios were easily calculated by counting the number of protein pairs with values that overlap with the GSP and GSN sets in the predictive dataset.

The confidence scores (LR) for each inferred PPI pair were the sum of the logarithmic form of all seven individual likelihood ratios from corresponding methods. The AtPID querying results page depicts the LR score from each method with open, partially or completely filled circles that indicate positive correlations with the confidence level of the interaction relationship. The detailed number of each predictive dataset is shown in Table 2 and all predictive datasets can be downloaded from the AtPID website.

DATABASE CONTENTS AND USAGE

All of the information in AtPID is derived from expert curation and deliberate computation. The process of creating a release AtPID database begins with extracting the published and other relevant information from various databases (Figure 1). Automated and manual quality assurance procedures are administered to verify data completeness and consistency. If necessary, material in the development database is revised and a new database version is generated.

Ortholog maps, domain attributes and network displays are developed with crosslinks to other relevant external resources (62). Following the final testing of data and the web server, the new database was made available via the public website. The latest release (14 July 2007) contains 28 062 PPI pairs involving 12 506 proteins. Of the PPI pairs, 23 396 pairs were inferred by the integration of several methods, while the other 3866 pairs involving 1875 proteins were manually curated from the literature and other 800 pairs were determined from enzyme complexes from KEGG. In addition to protein interaction data, we added subcellular location annotations to nearly 5000 proteins from SwissProt and SUBA databases (61,63) and popular prediction tools, including TargetP (64), Predotar (65) and MitoProtII (66), which can promote subproteome and protein function research.

Table 2. Overview of the number of individual predictive dataset

	Number of predictive PPI pairs	Number of proteins in the PPI pairs
O: Ortholog interaction datasets	3045	1359
G: Shared biological function:GO Ontology	553	523
E: Co-expression	14 837	8024
F: Gene fusion method	6570	5671
N: Gene neighbors method	2008	1637
P: Phylogenetic profile method	15 723	8751
D: Enriched domain pair	2182	1288
AtPID[a]	28 062 (putative PPI with GSP)	12 506
	23 396 (putative PPI without GSP)	11 706

[a]Through integration by Naïve Bays Network, AtPID achieved 28 062 PPI pairs with 23 396 pairs from prediction methods. There are seven individual datasets from various approaches, identified by O, G, E, F, N, P and D. The details of each method can be browsed on AtPID FAQ.

AtPID spans roughly 41% of the estimated 30 480 peptides with interaction annotations in the *Arabidopsis* genome and reflects the labor-intensive nature of manual curation. Our future plans are to manually mine thousands of protein interactions to acquire information through bulk importation of data from other sources or experimental results. Thus increasing PPI information and power as a resource. PPI will also provide enhanced resource training data for reconstructing interaction networks with higher accuracy and larger coverage of the *Arabidopsis* genome. In turn, the database can aid users in querying more detailed information about interaction pathways or maps comprising of interesting protein attributes.

Practical applications of AtPID: querying interactions

The AtPID website can be browsed similar to an online library. The website's home page, depicted in Figure 1, features the 'QUICK START' main interaction querying box with links to each of the seven method theories used in AtPID, the AtPID database statistics, and announcements regarding the function of the website.

PPI query is the main function of AtPID, which makes available manually collected PPI data and predicted PPI through integrated data resources. Query flow is illustrated in Figure 2 and demonstrates how querying a protein name or protein pair on the query page accesses PPI information (http://atpid.biosino.org/query.php). AtPID allows several types of query keywords used by other databases, including UniProtKB/Swiss-Prot ID, TAIR AGI, Entrez Gene name, REFSEQ PROVISIONAL ID (NCBI) or International Protein Index (IPI) symbols. We defined three types of submissions. (i) 'Simple search' allows the user to submit a single protein. This search is appropriate when the user would like to know which other protein(s) have the highest probability of interacting with the protein of interest. The search results include the GSP and PPI predictions. (ii) 'Pair search' allows the user to submit a protein pair to ascertain if an interaction between two proteins has been documented. (iii) 'Multiple search' allows a user to query more than two proteins. A comma separate format is required to access an interaction network among multiple proteins. All returned pages inform the user of related protein annotations by text and graphs. For example, the user is interested in the HAP3A protein, which encodes a subunit of the CCAAT-binding complex and binds to the CCAAT box motif present in some plant promoter sequences. The 'Search Results' show a summary of the protein attributes in the first table, including the 'Locus', gene/protein symbol, the number of interactions (six from GSP and seven by inference), function description and database cross-references to Entrez, TAIR, IPI, UniProtKB/TrEMBL, UniParc and KEGG (Figure 3).

The second table of the 'Search Results' presents inferred PPI pairs belonging to GSP listed with supporting evidence, including literature references from PubMed and experimental detection methods from text mining. Each interactant can be linked to a new 'Pair Search

Figure 1. The home page of AtPID.

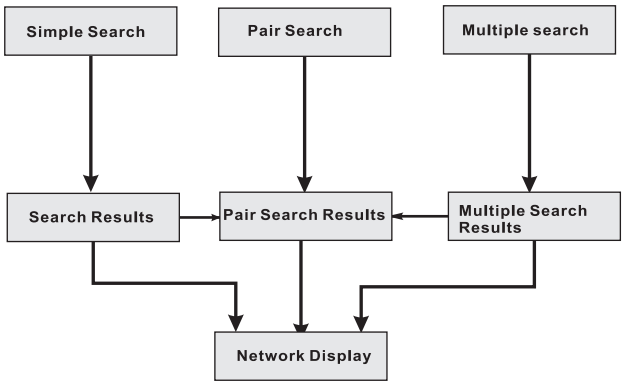

Figure 2. An overview of querying flow.

Results' window. The third table displays any potential interactant of HAP3A (Figure 3). Each of the seven prediction approaches is depicted by a letter acronym within a circle. When the user places the cursor over each circle, it displays the full name of the method. The circle under each method indicates the confidence strength for the predicted method and the related protein. The more the circle is filled, the more likely the pair of proteins is to interact. The corresponding score for the specific method is displayed when the cursor is held over the circle.

'Network Display' above the information table provides the link to a new window that displays the interaction network about HAP3A (Figure 4). In the 'Network Display' page, the query protein is represented as a triangle;

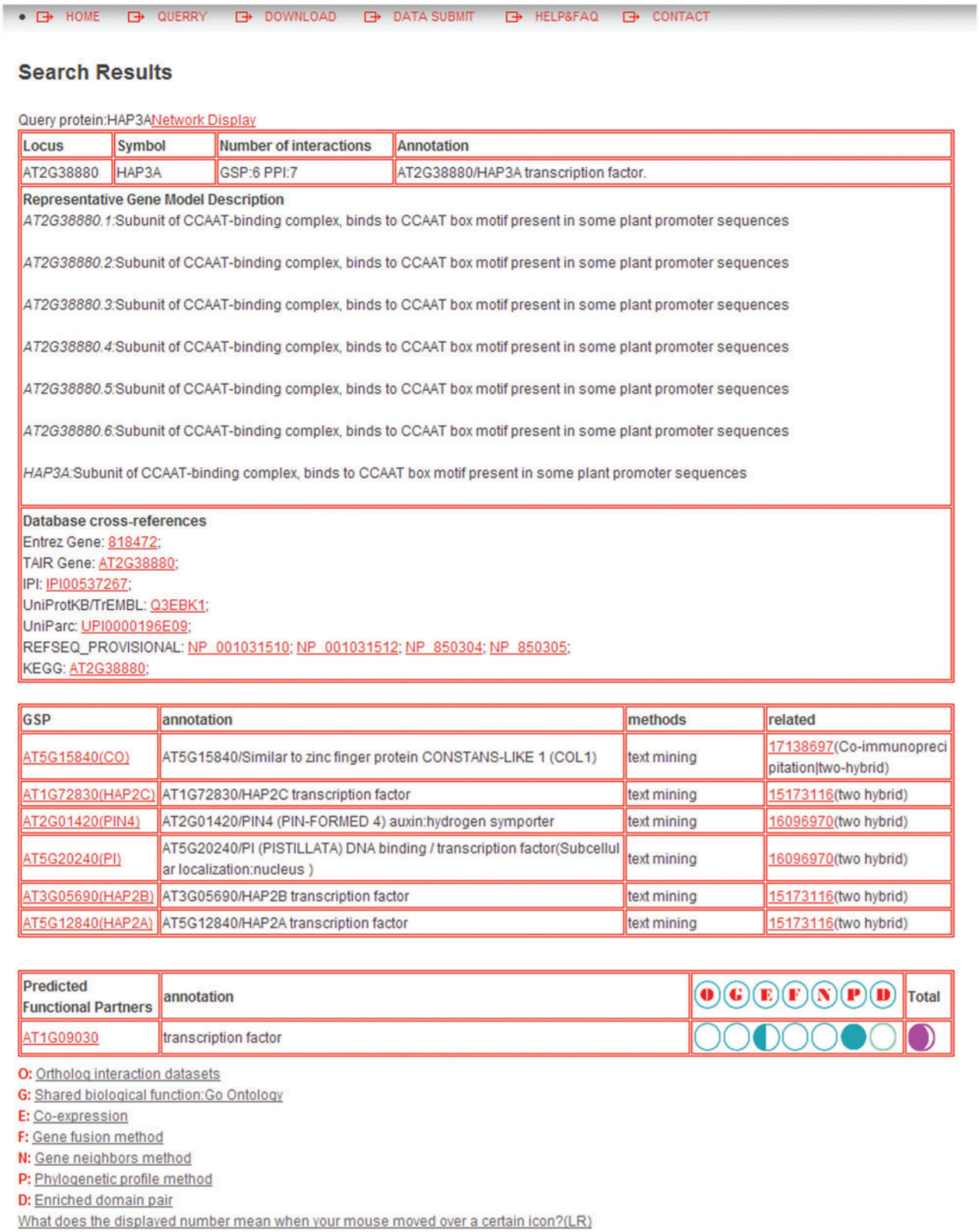

Figure 3. The AtPID interface of PPI Simple Search Results for the queried protein, HAP3A. There are generally three tables: (**1**) protein attributes with gene model descriptions and PPI summary, showing how many protein pairs can be predicted and the inferred interactions overlapped with GSP. (**2**) The second table represents interactants of the queried protein, HAP3A. Users can view what the experimental evidence is to support this consequence and link to corresponding literature(s). (**3**) The third table represents potential interactants of the queried protein, HAP3A, by Naïve Bays integration across the seven approaches. Each approach is graphically represented by circles. When the cursor is over a circle, it will display the corresponding score (LR) for the method.

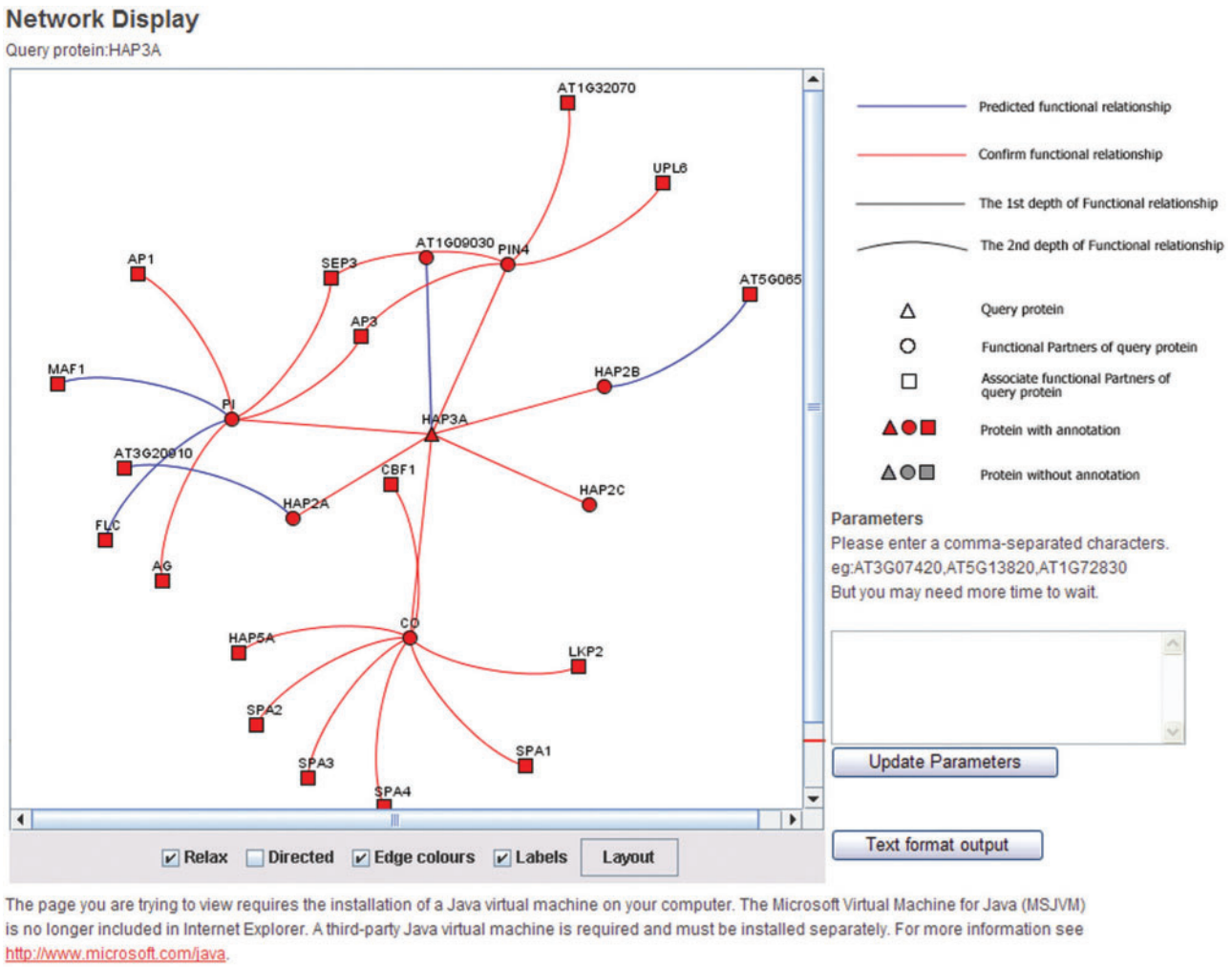

Figure 4. AtPID interface of the (Network Display) window for the queried protein, HAP3. In the (Network Display) window, we can view the interaction network involved with queried protein(s) and the neighbor component(s) intuitively. (Text Format Output) can export the interaction pair information for further analysis.

the functional Partners of the queried protein are represented by a circle, derived from the first level displayed in the PPI network that directly linked to the query protein. The associated functional partners of the queried protein are shown as squares, derived from the second level in the network. A red node represents a protein with known function (i.e. annotated), whereas a gray node represents an unknown functional protein (i.e. without annotation). The line between each protein indicates the functional relationship; a red line infers the interaction from text mining, and a blue line indicates the predictive function relationship. By holding the cursor over each protein, the related annotation for the protein is displayed and allows the user to navigate the network and easily check a proteins' relationship. 'Text Format Output' will export the interaction pair information in text format.

In the 'Pair Search Results' window when users submit potential interaction pairs, the domain attribute of one partner protein (e.g. AT1G09030) can be viewed graphically (Figure 5). Each module is linked to Pfam and when the user places the cursor over the module, details of the domain will also be displayed. Thus, AtPID provides comprehensive knowledge through a friendly and convenient interface that should be easy to use by biologists.

Software development

The database server is located at SIBS (Shanghai Institute of Biological Science) data service platform. Therefore, clients around the world can readily access the AtPID database. The AtPID development environment is apache + php + mysql, which allows for more efficient calculation rate performances and augmentation of the program.

CONCLUSIONS

AtPID is an online repository of *A. thaliana* protein interactions. AtPID serves as a major reference site for PPIs using *Arabidopsis* as a model plant system. The database collection will regularly integrate new accessions

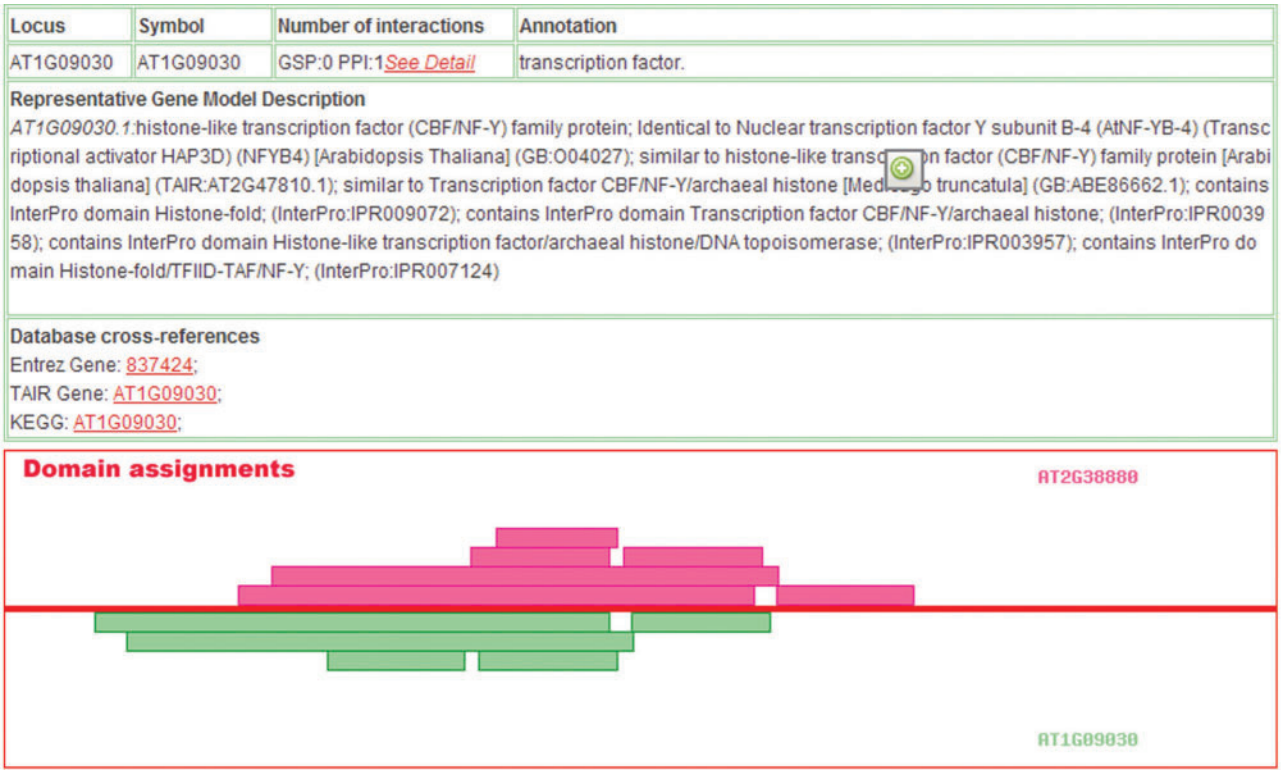

Figure 5. Domain attributes of one interactant of the queried protein, HAP3A.

as they become available. A number of new features and applications are currently under construction, such as a gene regulation and dynamic PPI network that function under different conditions with increased gene expression and proteomics data.

Currently, the subcellular localization predictions of *A. thaliana* are available for both the chloroplast and mitochondrion and the predictive organellar proteins have been added into AtPID. In addition, we plan to conduct further assessments of proteins to other cellular and/or subcellular locations, including nuclear, cytoplasmic and extracellular proteins.

Another important field of research is to elucidate the relationships between phenotype and genotype. For example, we plan to collect data relevant to mutants and their respective phenotypes. These highly varied types of data will be available through AtPID in the near future.

ACKNOWLEDGEMENTS

This work was supported by the State Key Program of Basic Research of China grants (2002CB713807, 2007CB108800), the National High Technology Research and Development Program of China (863 project) (Grant No. 2006AA02Z313, 2006AA10Z129) and National Natural Science Foundation of China grants (90408010 and 30571510). Funding to pay the Open Access publication charges for this article was provided by National Natural Science Foundation of China and the State Key Program of Basic Research of China.

Conflict of interest statement. None declared.

REFERENCES

1. Magdalena,S. (2005) Network biology: a protein network of one's own proteins. *Nat. Rev. Genets.*, **6**, 800.
2. Uhrig,J.F. (2006) *Planta*, pp. 1–11.
3. Schwikowski,B., Uetz,P. and Fields,S. (2000) A network of protein-protein interactions in yeast. *Nat. Biotechnol.*, **18**, 1257–1261.
4. Liu,Y. and Zhao,H. (2004) A computational approach for ordering signal transduction pathway components from genomics and proteomics Data. *BMC Bioinformatics*, **5**, 158–163.
5. Schulze,W.X., Deng,L. and Mann,M. (2005) Phosphotyrosine interactome of the ErbB-receptor kinase family. *Mol. Syst. Biol.*, **0008**, msb4100012:E4100011–E410001.
6. Sakakibara,H., Takei,K. and Hirose,N. (2006) Interactions between nitrogen and cytokinin in the regulation of metabolism and development. *Trends Plant Sci.*, **11**, 440–448.
7. Reed,J.L., Famili,I., Thiele,I. and Palsson,B.O. (2006) Towards multidimensional genome annotation. *Nat. Rev. Genet.*, **7**, 130–141.
8. 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.
9. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M. *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.
10. Giot,L., Bader,J., Brouwer,C., Chaudhuri,A., Kuang,B., Li,Y., Hao,Y.L., Ooi,C.E., Godwin,B. *et al.* (2003) A protein interaction map of Drosophila melanogaster. *Science*, **302**, 1727–1736.
11. Li,S., Armstrong,C., Bertin,N., Ge,H., Milstein,S., Boxem,M., Vidalain,P.O., Han,J.D., Chesneau,A. *et al.* (2004) A map of the interactome network of the metazoan C. elegans. *Science*, **303**, 540–543.
12. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A. *et al.* (2005)

A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
13. Huala,E., Dickerman A.W., Garcia-Hernandez,M., Weems,D., Reiser,L., LaFond,F., Hanley,D., Kiphart,D., Zhuang,M., Huang,W. *et al.* (2001) The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.*, **29**, 102–105.
14. de Folter,S., Immink,R., Kieffer,M., Parenicová,L., Henz,S.R., Weigel,D., Busscher,M., Kooiker,M., Colombo,L. *et al.* (2005) Comprehensive interaction map of the Arabidopsis MADS Box transcription factors. *Plant Cell*, **17**, 1424–1433.
15. Hackbusch,J., Richter,K., Müller,J., Salamini,F. and Uhrig,J.F. (2005) A central role of Arabidopsis thaliana ovate family proteins in networking and subcellular localization of 3-aa loop extension homeodomain proteins. *Proc. Natl Acad. Sci. USA*, **102**, 4908–4912.
16. Formosa,T., Barry,J., Alberts,B.M. and Greenblatt,J. (1991) Using protein affinity chromatography to probe structure of protein machines. *Meth. Enzymol.*, **208**, 24–45.
17. Miller,K.A. and Alberts,B.M. (1989) F-actin affinity chromatography: technique for isolating previously unidentified actin-binding proteins. *Proc. Natl Acad. Sci. USA*, **86**, 4808–4812.
18. Miller,K.G., Field,C.M. and Alberts,B.M. (1991) Use of actin filament and microtubule affinity chromatography to identify proteins that bind to the cytoskeleton. *Meth. Enzymol.*, **196**, 303–319.
19. Baird,B.A. and Hammes,G.G. (1976) Chemical cross-linking studies of chloroplast coupling factor 1. *J. Biol. Chem.*, **251**, 6953–6962.
20. Bragg,P. and Hou,C. (1980) A cross-linking study of the Ca2+, Mg2+-activated adenosine triphosphatase of Escherichia coli. *Eur. J. Biochem.*, **106**, 495–503.
21. Cover,J., Lambert,J.M. and Norman,C.M. (1981) Identification of proteins at the subunit interface of the Escherichia coli ribosome by cross-linking with dimethyl 3,3'-dithiobis(propionimidate). *Biochemistry*, **12**, 2843–2852.
22. Krieg,U., Johnson,A.E. and Walter,P. (1989) Protein translocation across the endoplasmic reticulum membrane: identification by photocross-linking of a 39-kD integral membrane glycoprotein as part of a putative translocation tunnel. *J. Cell Biol.*, **109**, 2033–2043.
23. Fields,S. and Sternglanz,R. (1994) The two-hybrid system: an assay for protein-protein interactions. *Trends Genet.*, **10**, 286–292.
24. Parrish,J.R., Gulyas,K. and Finley,R.L. (2006) Yeast two-hybrid contributions to interactome mapping. *Curr. Opin. Biotechnol.*, **17**, 387–393.
25. Chien,C., Bartel,P.L. and Sternglanz,R. (1991) The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc. Natl Acad. Sci. USA*, **88**, 9578–9582.
26. Clackson,T., Hoogenboom,H.R. and Griffiths,A.D. (1991) Making antibody fragments using phage display libraries. *Nature*, **352**, 624–628.
27. Phizicky,E.M. and Fields,S. (1995) Protein-protein interactions: methods for detection and analysis. *Microbiol. Rev.*, **59**, 94–123.
28. Deane,C.M., Salwinski,Ł., Xenarios,I. and Eisenberg,D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics*, **1**, 349–356.
29. Fu,L.D. and Tsamardinos,I. (2005) A comparison of Bayesian network learning algorithms from continuous data. *AMIA Annu. Symp. Proc.*, pp. 960.
30. Li,J., Li,X., Su,H., Chen,H. and Galbraith,D.W. (2006) A framework of integrating gene relations from heterogeneous data sources: an experiment on Arabidopsis thaliana. *Bioinformatics*, **22**, 2037–2043.
31. Needham,C.J., Bradford,J.R., Bulpitt,A.J. and Westhead,D.R. (2006) Inference in Bayesian networks. *Nat. Biotechnol.*, **24**, 51–54.
32. Kato,T., Tsuda,K. and Asai,K. (2005) Selective integration of multiple biological data for supervised network inference. *Bioinformatics*, **21**, 2488–2495.
33. Daeui,P., Semin,L., Dan,B., Michael,S., Michael,L., Donghoon,O. and Jong,B. (2005) Comparative interactomics analysis of protein family interaction networks using PSIMAP (protein structural interactome map). *Bioinformatics*, **21**, 3234–3240.
34. Walhout,A.J. and Vidal,M. (2001) Protein interaction maps for model organisms. *Nat. Rev. Mol. Cell Biol.*, **2**, 55–62.
35. Ben-Hur,A. and Noble,W.S. (2005) Kernel methods for predicting protein-protein interactions. *Bioinformatics*, **21**(Suppl. 1), i38–i46.
36. Buckingham,S.D. (2005) Data mining for protein-protein interactions in invertebrate model organisms. *Invert. Neurosci.*, **5**, 183–187.
37. Zhong,W. and Sternberg,P.W. (2006) Genome-wide prediction of C. elegans genetic interactions. *Science*, **311**, 1381–1382.
38. Moon,H.S., Bhak,J., Lee,K.H. and Lee,D. (2005) Architecture of basic building blocks in protein and domain structural interaction networks. *Bioinformatics*, **21**, 1479–1486.
39. Patil,A. and Nakamura,H. (2005) Filtering high-throughput protein-protein interaction data using a combination of genomic features. *BMC Bioinformatics*, **18**, 100.
40. Rhodes,D.R., Tomlins,S.A., Varambally,S., Mahavisno,V., Barrette,T., Kalyana-Sundaram,S., Ghosh,D., Pandey,A. and Chinnaiyan,A.M. (2005) Probabilistic model of the human protein-protein interaction network. *Nat. Biotechnol.*, **23**, 951–959.
41. Valente,A.X. and Cusick,M.E. (2006) Yeast protein interactome topology provides framework for coordinated-functionality. *Nucleic Acids Res.*, **34**, 2812–2819.
42. Date,S.V. and Stoeckert,C.J. (2006) Computational modeling of the Plasmodium falciparum interactome reveals protein function on a genome-wide scale. *Genome Res.*, **16**, 542–549.
43. Suthram,S., Sittler,T. and Ideker,T. (2005) The Plasmodium protein network diverges from those of other eukaryotes. *Nature*, **438**, 108–112.
44. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A. *et al.* (2007) IntAct – open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**(Database issue), D561–D565.
45. Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248–250.
46. Paul,J.K., Jorge,D., Allyson,W., Youla,K., Ewan,B. and Rolf,A. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.
47. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**(Database issue), 357.
48. Matthews,L.R., Vaglio,P., Reboul,J., Ge,H., Davis,B.P., Garrels,J., Vincent,S. and Vidal,M. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res.*, **11**, 2120–2126.
49. Ge,H., Liu,Z., Church,G.M. and Vidal,M. (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat. Genet.*, **29**, 482–486.
50. 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 Res.*, **29**, 3513–3519.
51. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.
52. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
53. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
54. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.
55. Huynen,M., Snel,B., Lathe,W.III and Bork,P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, **10**, 1204–1210.
56. Pellegrini,M., Marcotte,E., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.

57. Sun,J., Xu,J., Liu,Z., Liu,Q., Zhao,A., Shi,T. and Li,Y. (2005) Refined phylogenetic profiles method for predicting protein-protein interactions. *Bioinformatics*, **21**, 3409–3415.
58. O'Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**(Database issue), D476–D480.
59. Salwinski,L., Miller,C., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**(Database issue), D449–D451.
60. Jansen,R., Yu,H., Greenbaum,D., Kluger,Y., Krogan,N.J., Chung,S., Emili,A., Snyder,M., Greenblatt,J.F. *et al.* (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.
61. Heazlewood,J.L., Verboom,R., Tonti-Filippini,J., Small,I. and Millar,A.H. (2007) SUBA: the Arabidopsis subcellular database. *Nucleic Acids Res.*, **35**, D213–D218.
62. Hooper,S.D. and Bork,P. (2005) Medusa: a simple tool for interaction graph analysis. *Bioinformatics*, **21**, 4432–4433.
63. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
64. Emanuelsson,O., Brunak,S., von Heijne,G. and Nielsen,H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.*, **2**, 953–971.
65. Small,I., Peeters,N., Legeai,F. and Lurin,C. (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**, 1581–1590.
66. Claros,M.G. and Vincens,P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.*, **241**, 779–786.