D1130–D1133 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 29 October 2010*
doi:10.1093/nar/gkq959

AtPID: the overall hierarchical functional protein interaction network interface and analytic platform for *Arabidopsis*

Peng Li[1,2,3], Weidong Zang[2,3], Yuhua Li[2,3,*], Feng Xu[2,3], Jigang Wang[1,2] and Tieliu Shi[1,4,*]

[1]Center for Bioinformatics and Computational Biology, and The Institute of Biomedical Sciences, School of Life Science, East China Normal University, 500 Dongchuan Road, Shanghai 200241, [2]College of Life Sciences, the Northeast Forestry University, Harbin, Heilongjiang 150040, [3]Daqing Institute of Biotechnology, Northeast Forestry University, Daqing, Heilongjiang 163316 and [4]Shanghai Information Center for Life Sciences, Chinese Academy of Science, 319 Yueyang Road, Shanghai 200031, China

Received August 26, 2010; Accepted September 30, 2010

ABSTRACT

Protein interactions are involved in important cellular functions and biological processes that are the fundamentals of all life activities. With improvements in experimental techniques and progress in research, the overall protein interaction network frameworks of several model organisms have been created through data collection and integration. However, most of the networks processed only show simple relationships without boundary, weight or direction, which do not truly reflect the biological reality. *In vivo*, different types of protein interactions, such as the assembly of protein complexes or phosphorylation, often have their specific functions and qualifications. Ignorance of these features will bring much bias to the network analysis and application. Therefore, we annotate the *Arabidopsis* proteins in the AtPID database with further information (e.g. functional annotation, subcellular localization, tissue-specific expression, phosphorylation information, SNP phenotype and mutant phenotype, etc.) and interaction qualifications (e.g. transcriptional regulation, complex assembly, functional collaboration, etc.) via further literature text mining and integration of other resources. Meanwhile, the related information is vividly displayed to users through a comprehensive and newly developed display and analytical tools. The system allows the construction of tissue-specific interaction networks with display of canonical pathways. The latest updated AtPID database is available at http://www.megabionet.org/atpid/.

INTRODUCTION

With the improvement of modern biological research technology and the advances in studying the model organism—*Arabidopsis thaliana*, a large amount of data related to proteins, such as the data of proteomics, subcellular localization, (1,2) three-dimensional structures, the tissue-specific gene expressions (3,4), etc. have been published in corresponding literature. Progress in functional genomics has allowed large amounts of data about mutants to be reported. The mutant related data are partially curated and mapped to the protein-coding genes with the information of germplasm and phenotype information in TAIR (5) and other seed resource databases (6). These annotated data are valuable resources for researchers to further comprehensively understand the gene/protein functions at multi-levels.

Meanwhile, the research of protein–protein interaction (PPI) in *Arabidopsis* has achieved significant results, both experimentally and computationally (7). High-throughput data, such as protein phosphorylation (8,9) has also been reported. In addition, the accumulation of data related to signal transduction and transcriptional regulatory mechanisms (10,11) has allowed a great quantity of protein–protein interactions to be annotated in detail. The integration of various data from the AtPID database has been updated accordingly and the new properties about the network have been established, the annotation assignment of each node, along with the direction of edges and the type/style of edges all added to the previous protein–protein interaction network.

Besides the accomplished text mining and general data collection that expand the contents of AtPID database, new network display programs have also been developed to help researchers focus on the protein that they are interested in. In our latest AtPID 4.0 version, a more advanced query mode, allowing retrieval of a whole

*To whom correspondence should be addressed. Tel: +86 21 54345020; +86 21 54344922; Email: tlshi@sibs.ac.cn
Correspondence may also be addressed to Yuhua Li. Tel: +86 451 8219 1737; Fax: +86 451 8219 1733; Email: lyhshen@126.com

pathway, is implemented by certain optimized algorithms. All of the improvements and updates will accelerate researchers in exploiting information in our protein–protein interaction network in an effective and comprehensive way.

DATA SOURCES

Rather than storing data from other databases in our system, we retrieve data directly from these other data resources through a data interface; this eliminates the potential bias due to delayed updates and inconsistent data integration. In the process of literature data assembly, we consider accuracy as the first priority, not just quantity of the data. At the same time, we also standardize data mining processes, determine and record data manually for data from each literature with archiving the related literature information and descriptive sentences. Moreover, we tap the literature PubMed link for text-mining data, and provide the cross-links to other related databases. Users can view these resources from the query result pages and network display page.

For each queried results, besides displaying them in a table format layout, we also display them by integrating all related data in a network way with our newly developed powerful tool, and the intuitive visualization provides users a convenient way to navigate the data and check the relationships between proteins, the relationships include the interaction between proteins, the transcription regulation between the transcription factors and their target genes, the pathway in which the proteins are involved and the phosphorylation status of the protein, etc.

Protein annotation

Phenotype data. Mutants have been widely used in functional genomics research, considering the advantages in seed mutagenesis, genetic modification and tissue culture, plants are easier than animals for obtaining stable traits, and thus, have generated rich resources for mutants. So far, a large number of characterized stable *Arabidopsis* mutants has been reported in research literature (12), and some seed resource databases. We have integrated information and experimental results extracted from research literature, seed resource database and TAIR-released phenotype data into our AtPID database; at the same time, we also have classified those data and annotated the phenotypes for mutants based on plant ontology (13) (Table 1)

Proteomics data of 12 tissues. We have collected proteomics data from several MS experiments (14) and integrated them into AtPID database. Those MS data are generated from 12 different tissues and some of those are tissue-specific proteins that come from flower bud, flower, cotyledon, juvenile leaf, root, seed, carpel, silique, cell suspension culture, shoot, rosette and pollen. There are a total of 13 970 non-redundant proteins identified from those 12 different tissues (Table 2), which can be regarded as a set with high confidence.

Table 1. Protein annotation information

Annotation type	Data sources	The amount of data
Subcellular localization	GO, SUBA, Text-mining	10 429 proteins
Functional annotation	TAIR, GO, PO, NCBI	40 000 proteins
Mutant information	Tair, NASC, RAPID, Text-mining	5121 mutants, 3431 genes
Pathway information	KEGG	142 pathways, 5514 proteins
Phosphoprotein	PhatPhos	2514 proteins
Tissues	PRIDE*	12 tissues, 13 970 proteins

*Statistics Date: 1 September 2010.

Table 2. Proteomics data of 12 tissues

Tissue	Protein count
Carpel	3946
cell suspension culture	8698
Cotyledon	3665
Flower	5215
flower bud	5104
juvenile leaf	3892
Pollen	3511
Root	6125
Seed	3789
Silique	5779
shoot	3264
rosette	4219
total	13 970

Moreover, based on the collected proteomics data, we have integrated the identified proteins from each of 12 tissues with the protein interaction data in AtPID database; users can display the protein interaction network in a selected tissue. The network view of proteins in each tissue provides an intuitive way for users to explore the protein function, the protein interaction relationships, the function module of the protein involved and the potential regulatory relationship based on the current available data.

Interaction annotation

Transcriptional regulation is one of the research fields making dramatic progress recently with the adoption of new technologies and is an important resource for us to understand the mechanisms of various biological processes and activities. Therefore, we have also integrated such related information into our AtPID database to help the community to explore the regulatory relationships between the transcription factors and their target genes. In the AtPID database, the gene transcription regulation information has been embedded into the protein interaction network and we believe that it makes the overall network more consistent with the biological reality through the transcriptional regulation information integrated into the static protein interaction network. Moreover, we extend our PPI data through the integration and text mining, and have collected 770 new relationships

involved in 522 proteins into our AtPID database (Table 3).

DISPLAY

For the purpose of maximizing display annotation information and optimizing web data transmission, we developed new online display tools, which provide more detailed network information output and incorporate some useful online analysis tools. With the optimization of algorithms, the components of the displayed network can be expanded to 1500 nodes maximally or until up to the limit of the dataset. In this way, users can analyze and visualize the query results in a global view (Figure 1). We plan to launch the off-line version in the near future in order to provide the opportunity for the users to better display and analyze relevant network information.

Table 3. Interaction annotation information

Annotation type	Data sources	The amount of data
Protein complex	Swiss-port, KEGG	238 complex, 590 proteins
Transcriptional regulation	Text mining	8070 relations, 5211 proteins
PPI with evidence	Text mining	4642 relations, 2341 proteins

CONCLUDING REMARK

As the platform of protein–protein interaction for *Arabidopsis thaliana*, AtPID database will continue to enrich the data concerning protein–protein interactions through text mining and analysis, focus on the integration of different types of data efficiently as well as developing new display tools, which will enable researchers to analyze, utilize and visualize the data in convenient ways. As more direct evidence for protein–protein interactions in *A. thaliana* cells becomes available, it will be possible to refine the networks we have defined and make them more useful for testing hypotheses about the mechanisms of various physiological activities. Our next step is to integrate several high-throughput data types, such as gene expression data generated from deep sequence technology and proteomics profiles, and annotate the network to provide direction. We also plan to carry out quantification analysis for protein interaction network via statistical approaches; at the same time, we will dynamically display the network based on time course data on developmental or real-time responses to reflect tissue-specific gene expression.

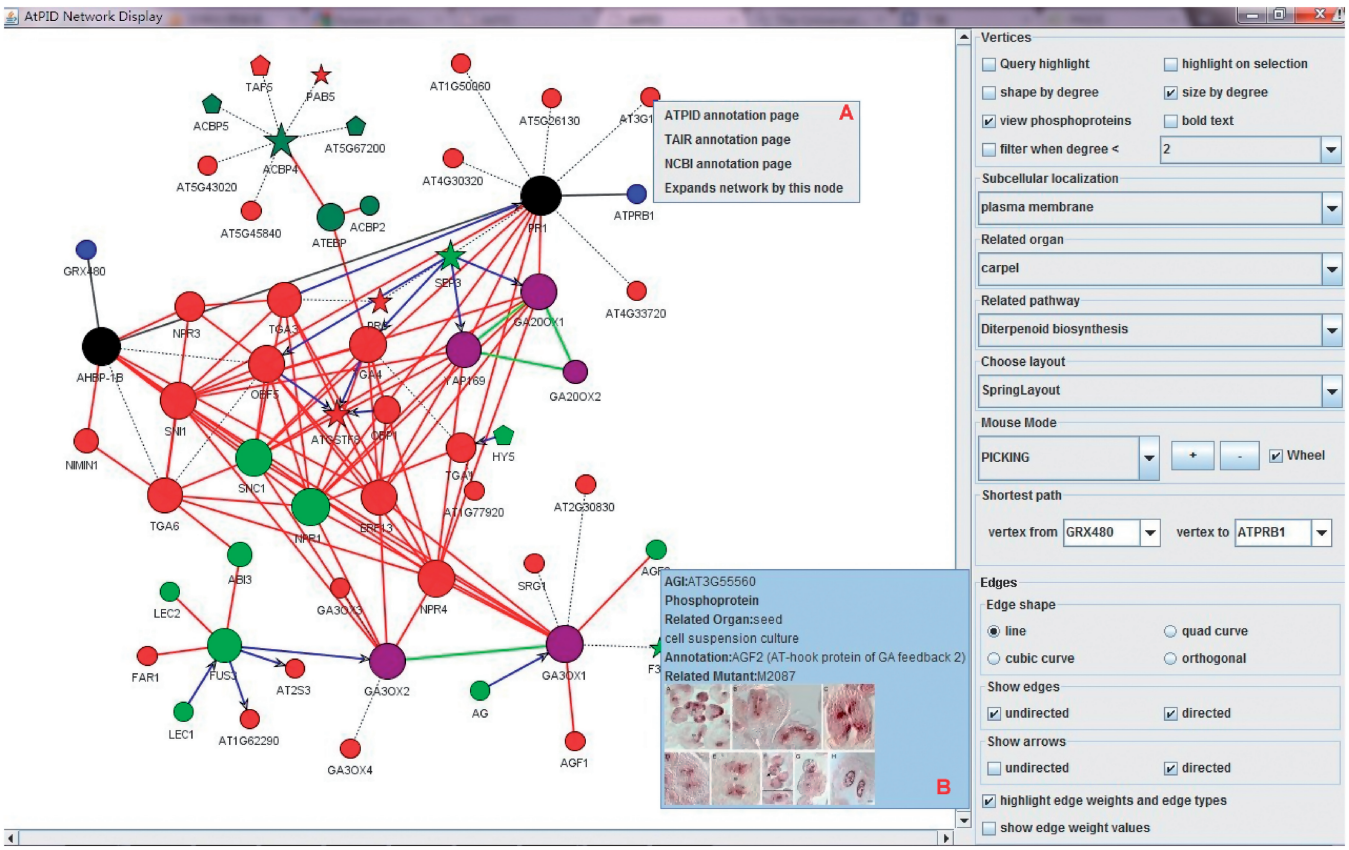

Figure 1. Protein–protein interaction network display. Edges color: red, PPI with evidence; green, proteins in small complex; blue, transcriptional regulation. Vertices color: green, gene with mutant info; purple, genes with selected subcellular localization info; Dark green, genes with selected pathway info; blue, shortest path starting and ending protein; black, shortest path via protein. Vertices shape: regular pentagon, phosphoproteins; 5-pointed star, proteins with selected tissue information. (**A**) Node right menu. (**B**) Node mouse hovering annotation.

FUNDING

Funding for open access charge: State Key Program of Basic Research of China grants (2007CB108800, 2010CB945400); National High Technology Research and Development Program of China (863 project) (Grant No. 2006AA10Z129, 2006AA02Z313); National Natural Science Foundation of China grants (30870575, 30730078); Science and Technology Commission of Shanghai Municipality (06DZ22923).

Conflict of interest statement. None declared.

REFERENCES

1. Heazlewood,J.L., Verboom,R.E., Tonti-Filippini,J., Small,I. and Millar,A.H. (2007) SUBA: the Arabidopsis Subcellular database. *Nucleic Acids Res.*, **35(Database issue)**, D213–D218.
2. Marion,J., Bach,L., Bellec,Y., Meyer,C., Gissot,L. and Faure,J.D. (2008) Systematic analysis of protein subcellular localization and interaction using high-throughput transient transformation of Arabidopsis seedlings. *Plant J.*, **56**, 169–179.
3. Tamura,T., Asakura,T., Uemura,T., Ueda,T., Terauchi,K., Misaka,T. and Abe,K. (2008) Signal peptide peptidase and its homologs in *Arabidopsis thaliana*–plant tissue-specific expression and distinct subcellular localization. *FEBS J.*, **275**, 34–43.
4. Ckurshumova,W., Koizumi,K., Chatfield,S.P., Sanchez-Buelna,S.U., Gangaeva,A.E., McKenzie,R. and Berleth,T. (2009) Tissue-specific GAL4 expression patterns as a resource enabling targeted gene expression, cell type-specific transcript profiling and gene function characterization in the Arabidopsis vascular system. *Plant Cell. Physiol.*, **50**, 141–150.
5. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al*. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36(Database issue)**, D1009–D1014.
6. Scholl,R.L., May,S.T. and Ware,D.H. (2000) Seed and molecular resources for Arabidopsis. *Plant Physiol.*, **124**, 1477–1480.
7. Kim,S.H., Kwon,S.I., Saha,D., Anyanwu,N.C. and Gassmann,W. (2009) Resistance to the Pseudomonas syringae effector HopA1 Is governed by the TIR-NBS-LRR protein RPS6 and is enhanced by mutations in SRFR1. *Plant Physiol.*, **150**, 1723–1732.
8. Heazlewood,J.L., Durek,P., Hummel,J., Selbig,J., Weckwerth,W., Walther,D. and Schulze,W.X. (2008) PhosPhAt: a database of phosphorylation sites in Arabidopsis thaliana and a plant-specific phosphorylation site predictor. *Nucleic Acids Res.*, **36(Database issue)**, D1015–D1021.
9. Li,H., Wong,W.S., Zhu,L., Guo,H.W., Ecker,J. and Li,N. (2009) Phosphoproteomic analysis of ethylene-regulated protein phosphorylation in etiolated seedlings of Arabidopsis mutant ein2 using two-dimensional separations coupled with a hybrid quadrupole time-of-flight mass spectrometer. *Proteomics*, **9**, 1646–1661.
10. Shin,R., Burch,A.Y., Huppert,K.A., Tiwari,S.B., Murphy,A.S., Guilfoyle,T.J. and Schachtman,D.P. (2007) The Arabidopsis transcription factor MYB77 modulates auxin signal transduction. *Plant Cell*, **19**, 2440–2453.
11. Steffens,N.O., Galuschka,C., Schindler,M., Bülow,L. and Hehl,R. (2005) AthaMap web tools for database-assisted identification of combinatorial cis-regulatory elements and the display of highly conserved transcription factor binding sites in *Arabidopsis thaliana*. *Nucleic Acids Res.*, **33(Web server issue)**, W397–W402.
12. Kuromori,T., Wada,T., Kamiya,A., Yuguchi,M., Yokouchi,T., Imura,Y., Takabe,H., Sakurai,T., Akiyama,K., Hirayama,T. *et al*. (2006) A trial of phenome analysis using 4000 Ds-insertional mutants in gene-coding regions of Arabidopsis. *Plant J.*, **47**, 640–651.
13. Jaiswal,P., Avraham,S., Ilic,K., Kellogg,E.A., McCouch,S., Pujar,A., Reiser,L., Rhee,S.Y., Sachs,M.M., Schaeffer,M. *et al*. (2005) Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comp. Funct. Genomics*, **6**, 388–397.
14. Jones,P., Côté,R.G., Martens,L., Quinn,A.F., Taylor,C.F., Derache,W., Hermjakob,H. and Apweiler,R. (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.*, **34(Database issue)**, D659–D663.