PAIR: the predicted *Arabidopsis* interactome resource

Mingzhi Lin[1], Xueling Shen[2] and Xin Chen[1,2,*]

[1]Department of Bioinformatics and [2]Institute of Biochemistry, Zhejiang University, Hangzhou, PR China, 310058

ABSTRACT

The predicted *Arabidopsis* interactome resource (PAIR, http://www.cls.zju.edu.cn/pair/), comprised of 5990 experimentally reported molecular interactions in *Arabidopsis thaliana* together with 145 494 predicted interactions, is currently the most comprehensive data set of the *Arabidopsis* interactome with high reliability. PAIR predicts interactions by a fine-tuned support vector machine model that integrates indirect evidences for interaction, such as gene co-expressions, domain interactions, shared GO annotations, co-localizations, phylogenetic profile similarities and homologous interactions in other organisms (interologs). These predictions were expected to cover 24% of the entire *Arabidopsis* interactome, and their reliability was estimated to be 44%. Two independent example data sets were used to rigorously validate the prediction accuracy. PAIR features a user-friendly query interface, providing rich annotation on the relationships between two proteins. A graphical interaction network browser has also been integrated into the PAIR web interface to facilitate mining of specific pathways.

INTRODUCTION

Protein–protein interactions (PPIs) are the major components of many fundamental cellular processes. Identification of the interactions involving a protein is often a key step toward understanding its functions in the cellular context. Many efforts have been made to chart PPI maps in several model organisms. In *Saccharomyces cerevisiae* (1,2), *Drosophila melanogaster* (3), *Caenorhabditis elegans* (4) and *Homo sapiens* (5,6), genome-wide yeast two-hybrid screens and large-scale affinity purification/mass spectrometry studies have been conducted to map their interactomes. Meanwhile, a number of databases, such as IntAct (7), BioGRID (8), BIND (9) have been established as repositories for inter-

action data. The STRING database (10) collected a large set of known and predicted protein interactions. The quality of experimentally reported interactions has been rigorously assessed (11). However, to date, no large-scale experiment aiming to map a plant interactome has been reported (12). The number of plant PPIs in these databases remains very limited.

In the plant kingdom, *Arabidopsis thaliana* is arguably the most important model organism. Even for this best-studied model, less than 6000 experimentally reported interactions can be found in major data repositories. Therefore, the need for predicted interactions has been recognized by several groups and led to a series of efforts to predict the *Arabidopsis* interactome. Geisler-Lee *et al.* (13) predicted approximately 20 000 *Arabidopsis* interactions from homologous interactions in other species (the 'interolog' approach). A recent study went one step further to use functional association data to improve prediction reliability, which resulted in approximately 18 000 filtered predictions (14). Though useful, these interolog-based approaches are limited to detecting evolutionarily conserved protein interactions, whereas a significant number of *A. thaliana* proteins do not have orthologs in other model organisms with rich interactome information. Another work, the AtPID database (15), predicted approximately 23 000 interactions from multiple indirect evidences using a Naïve Bayesian approach. This work represented a conceptual advance in interaction prediction, yet the number of predicted interactions seemed to be too small to represent a comprehensive interactome. In addition, the AtPIN database (16) integrated experimentally reported interactions from the major data repositories and the predicted interactions from the Geisler-Lee data set and the AtPID database.

However, in all the prediction efforts, accuracies were not rigorously assessed with external benchmark data sets, nor could they give a reasonable estimation of the *Arabidopsis* interactome size. It has been suggested that the yeast interactome includes approximately 18 000 PPIs involving approximately 6000 genes (17). Assuming the same rate of interaction between genes, approximately 200 000 *Arabidopsis* PPIs would be expected between the

---

*To whom correspondence should be addressed. Tel/Fax: (+86) 571 88208595; Email: xinchen@zju.edu.cn

approximately 20 000 *Arabidopsis* genes. Therefore, the number of experimentally reported interactions and the sizes of available predicted interactomes are approximately an order of magnitude less than the expected size of *Arabidopsis* interactome.

Here, we present the predicted *Arabidopsis* interactome resource (PAIR), which contains the most comprehensive data set of the *Arabidopsis* interactome to date. These interactions are expected to cover 24% of the entire *Arabidopsis* interactome with a reasonably high reliability (confidence of each predicted interaction) of 44%. PAIR features an information-rich and user-friendly interface and an integrated, graphical interaction network browser to facilitate mining of specific pathways.

THE DATABASE

Overview

The PAIR project started as a simple effort to infer *Arabidopsis* interactions by homology mapping. In 3 years, it gradually evolved into a dedicated effort aiming to provide the most accurate interactome predicted by the state-of-the-art machine learning approach. The current version of PAIR (V3.3) contains 5990 experimentally reported molecular interactions together with 145 494 predicted interactions in *A. thaliana*.

The 5990 experimentally reported interactions were collected from three interaction repositories, i.e. IntAct (7), BioGrid (8) and BIND (9), as of 23 July 2010, and the interaction data set compiled by TAIR curators (18). All these major interaction data repositories manually curate data from literature. However, the *Arabidopsis* interactions collected by these repositories showed a small overlap (19). In our compilation, only 131 interactions were shared by all repositories. Therefore, integration of their data is necessary and helpful. In this effort, PPIs in these repositories with experimental evidence were extracted. Protein identifiers in respective repositories were mapped to the *Arabidopsis* gene loci according to the conversion tables provided by TAIR (Version 9) (18). Interactions involving proteins that could not be unambiguously mapped were discarded. This resulted in 5990 experimentally reported PPIs involving 2824 proteins.

The 145 494 interaction predictions were made at an earlier time, 1 February 2010, as part of the PAIR V3.0 major release. Due to the heavy computational requirements, interaction predictions were only updated with each major data release. These interactions were predicted by a support vector machine (SVM) model that integrates several indirect evidences for interaction, such as gene co-expressions, domain interactions, shared GO annotations, co-localizations, phylogenetic profile similarities and interologs (20). The SVM model was trained using a set of example interactions known as the Gold Standard Positives (GSPs), which is a collection of interactions in the major repositories as of 15 June 2009. The prediction accuracy was validated using two external benchmark data sets, containing interaction examples that were not available at the time when we trained our prediction

model (detailed below). The algorithmic details on how the indirect evidences were computed, how the prediction models were trained, and how the prediction accuracies were evaluated can be found in the Help/FAQ page of the PAIR website. Altogether, 145494 interactions involving 9480 proteins were predicted by the PAIR V3 prediction model. These predicted interactions were expected to cover 24% of the entire *Arabidopsis* interactome, and the reliability of each predicted interaction was estimated to be 44%. These predicted interactions had 1584 (26.44%) overlap with the 5990 experimentally reported interactions mentioned above.

Altogether, the PAIR 3.3 release contains 149 900 interactions involving 10 380 proteins. They can be queried through a user-friendly web interface, downloaded in a number of widely-used data formats or mined with a graphical interaction network browser integrated within the PAIR website.

The quality of predicted interactions

Two external benchmark data sets were used to verify the accuracy of predicted interactions. Before showing the assessment results, it needs to be clarified that the interaction data set searchable at the PAIR website (PAIR V3.3) is essentially a compilation including the PAIR V3 predictions and the experimentally reported interactions deposited in the major interaction databases before 23 July 2010 (update V3.3). In the accuracy assessments below, only the PAIR V3 predictions without the additional experimentally-reported interactions were evaluated.

The first benchmark data set contained newly reported interactions that were not included in the major interaction databases at the time (15 June 2009) our positive interaction examples were assembled. This independent evaluation set was retrieved from an update of the BioGRID database (8) (as of 27 December 2009), which included 448 new interactions that had been double-checked to avoid any overlap with our GSPs used in model training. As shown in Table 1, 115 (26%) of these new interactions were successfully recognized by our prediction model. This sensitivity (26%) was comparable to the expected sensitivity (24%). In contrast, these new interactions showed much less overlap with other predicted interactomes. Only 50 interactions could be found in the Geisler-Lee data set (the Interologs data set) (13), 20 in the De Bodt data set (the Filtered Interologs data set) (14), 16 in the AtPID database and 57 in the predicted interactions in the AtPIN database (16). Detailed results are provided in Supplementary Table S1.

The other benchmark data set was reported by a recent (April 2010) article in Plant Cell (21), published 2 months after PAIR V3 was released. In this report, 917 protein pairs involving 58 core cell-cycle proteins were tested by two complementary interaction assays, bimolecular fluorescence complementation and high-confidence yeast-two-hybrid, resulting in 357 interactions, of which 293 had not been reported before (21). Among the 357 reported interactions, PAIR predicted 170 (48%) of them. Among the 293 newly reported interactions, 140 (48%) were correctly predicted. Again, in this data set, PAIR predicted many

Table 1. Assessments of prediction accuracy using independent benchmark data sets

Independent example data set	PAIR[a]	AtPID[b]	Interologs[c]	Filtered Interologs[d]	AtPIN[e]
New interactions in BioGrid (448)[f]	115	16	50	20	57
All cell-cycle protein interactions (357)[g]	170	22	27	11	37
New cell-cycle protein interactions (293)[h]	140	16	16	2	26

[a]The PAIR V3 predicted interactions.
[b]The predicted interactions in the AtPID database (15).
[c]The interactions predicted by Geisler-Lee *et al.* (19) based on homology to known interactions in other organisms (interologs).
[d]The filtered interolog-based predicted interactions by De Bodt *et al.* (14).
[e]The predicted interactions in the AtPIN database (essentially the assembly of the Interologs data set and the AtPID database) (16).
[f]The newly curated interactions in the BioGRID database (8).
[g]The 375 interactions between 58 cell-cycle proteins reported in a recent Plant Cell article (21).
[h]The 293 interactions reported in (21) that had not been reported before.

more interactions than other predicted interactomes. As shown in Table 1, the sensitivity of PAIR V3 was over four times higher than others. Detailed results are provided in Supplementary Table S2. On the other hand, PAIR predicted 338 interactions from the 917 experimentally tested protein pairs, of which 170 were confirmed. Therefore, the reliability of PAIR predictions reached 50%. In other words, over half of the predicted interactions were real in this test. However, it should be noted that the coverage and reliability observed with this data set (48/50%) were higher than our estimations (24/44%). This might have happened because the proteins tested in this experiment were all well-studied core cell-cycle proteins. Well-studied proteins tend to have more comprehensive and accurate supporting data to compute the indirect evidences based on which our predictions are made. Therefore, the high coverage and reliability observed with this data set may not apply to the entire predicted interactome. Even so, these results showed that with the ever-growing volume of protein characteristics data, PAIR has the potential to predict the *Arabidopsis* interactome at higher levels of coverage and reliability.

It is also worth noting that the size, estimated coverage and reliability of the predicted *Arabidopsis* interactome implied an estimation of the *Arabidopsis* interactome size, $\sim 2.58 \times 10^5$. In other words, 1 out of 893 random protein pairs was expected to interact. This ratio is similar to the experimentally observed ratio in yeast (1/775) (17). Considering a smaller fraction of the genome is usually expressed at the same time in higher organisms as compared to unicellular species, this estimated ratio of interacting protein pairs seemed to make sense. Details of the above results are provided at the PAIR website.

In addition, a recent Plant Cell article by another group (19) showed that the previous version of PAIR (PAIR V2)

already had the highest coverage among all available predicted interactomes. Using the newly curated interactions in the IntAct database (7) as a benchmark, this article reported that the coverage of PAIR V2 predictions was more than double the coverage of the second-best predicted interactome (19 versus 9%).

The high coverage of our predicted interactions is supported by multiple assessments. However, the high reliability of these predicted interactions is not validated with external experimental data. This is because that most negative results, pairs of proteins that do not interact, are never reported in literature. Consequently there is no reliable data source of non-interactions that is large enough to support an accurate estimation of our prediction reliability. But given that the estimated coverage and reliability led to a reasonable estimation of the interactome size, the estimated reliability (44%) should be roughly accurate.
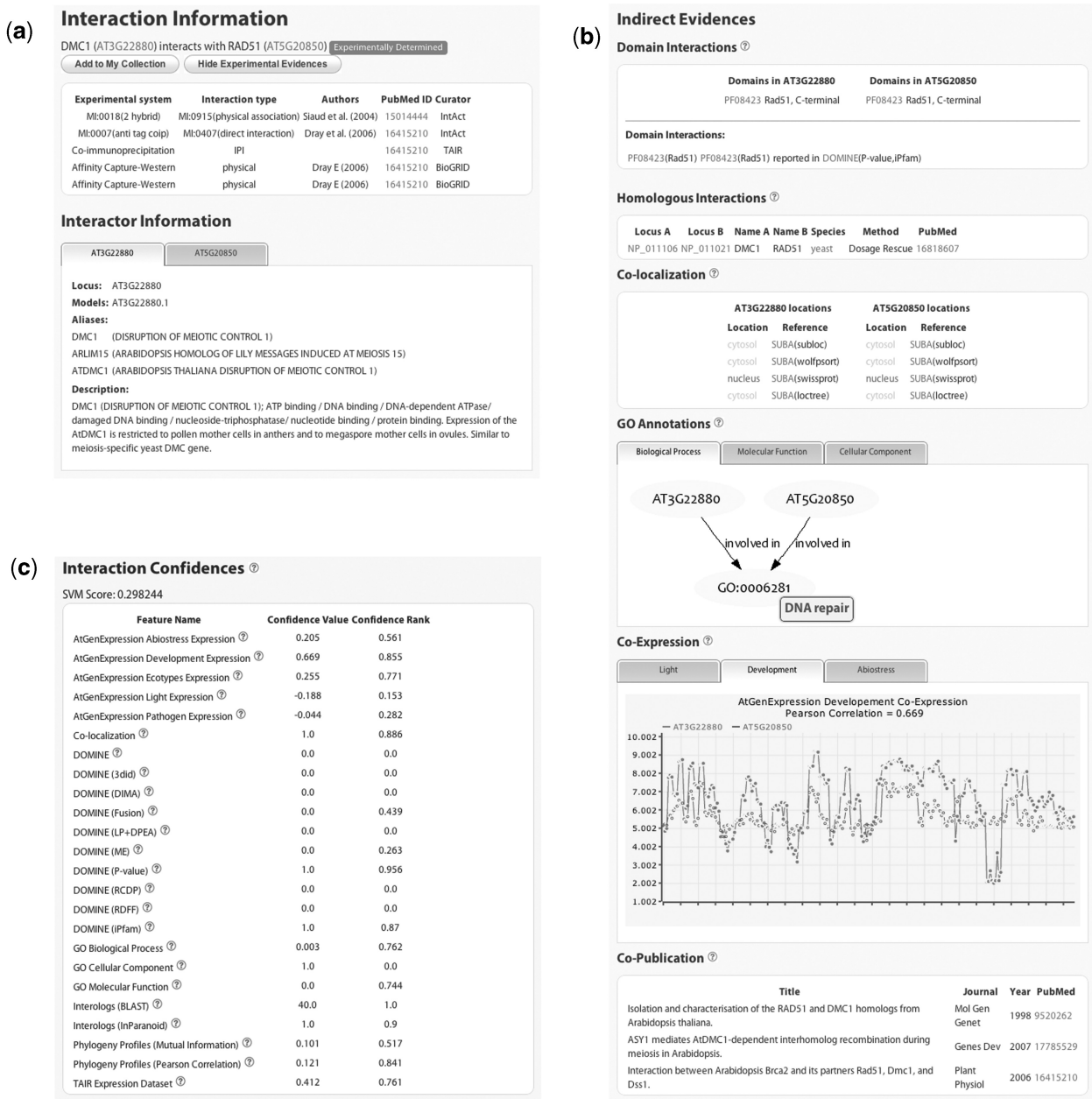
THE WEB INTERFACE

Search methods

In PAIR, interactions can be searched by specifying one or both of their component proteins or by specifying a homologous interaction (with both component proteins) in one of four other model organisms: *H. sapiens*, *S. cerevisiae*, *C. elegans* and *D. melanogaster*. Proteins may be specified by their identifiers, such as AGI codes (gene loci), UniProt accessions and RefSeq identifiers, or by keywords in their annotation texts. Alternatively, users can perform a BLAST sequence search to retrieve interactions involving a particular gene family or protein domain. In addition, PAIR supports gene set search, which allows a user to enter a number of AGI codes (gene loci). According to user option, PAIR can return interactions between the specified proteins or all interactions involving the specified proteins. This function is most useful to extract interaction sub-networks related to a specific cellular process.

Interaction information

For every PPI, PAIR provides rich annotation on the relationships between the two proteins involved. Taking the interaction between DMC1 (At3g22880) and RAD51 (At5g20850) as an example, we show in Figure 1a typical Interaction Information page. This page contains three sections. The first section shows a summary of the interaction and its component proteins (Figure 1a). If the interaction has been experimentally reported, the related experimental evidences will be shown. The second section shows the indirect evidences supporting this interaction, including gene co-expressions, domain interactions, shared GO annotations, co-localizations, phylogenetic profile similarities and homologous interactions in other organisms (interologs) (Figure 1b). For domain interactions, domains in both proteins are retrieved from Pfam (22). Known interactions between the domains were collected from the DOMINE database (23), which contains multiple domain interaction data sets reported by different approaches. The homologous interactions in four

(a)

Interaction Information

DMC1 (AT3G22880) interacts with RAD51 (AT5G20850) [Experimentally Determined]

[Add to My Collection] [Hide Experimental Evidences]

Experimental system	Interaction type	Authors	PubMed ID	Curator
MI:0018(2 hybrid)	MI:0915(physical association)	Siaud et al. (2004)	15014444	IntAct
MI:0007(anti tag coip)	MI:0407(direct interaction)	Dray et al. (2006)	16415210	IntAct
Co-immunoprecipitation	IPI		16415210	TAIR
Affinity Capture-Western	physical	Dray E (2006)	16415210	BioGRID
Affinity Capture-Western	physical	Dray E (2006)	16415210	BioGRID

Interactor Information

[AT3G22880] [AT5G20850]

Locus: AT3G22880
Models: AT3G22880.1
Aliases:
DMC1   (DISRUPTION OF MEIOTIC CONTROL 1)
ARLIM15 (ARABIDOPSIS HOMOLOG OF LILY MESSAGES INDUCED AT MEIOSIS 15)
ATDMC1 (ARABIDOPSIS THALIANA DISRUPTION OF MEIOTIC CONTROL 1)

Description:
DMC1 (DISRUPTION OF MEIOTIC CONTROL 1); ATP binding / DNA binding / DNA-dependent ATPase/ damaged DNA binding / nucleoside-triphosphatase/ nucleotide binding / protein binding. Expression of the AtDMC1 is restricted to pollen mother cells in anthers and to megaspore mother cells in ovules. Similar to meiosis-specific yeast DMC gene.

(c)

Interaction Confidences ⑦

SVM Score: 0.298244

Feature Name	Confidence Value	Confidence Rank
AtGenExpression Abiostress Expression ⑦	0.205	0.561
AtGenExpression Development Expression ⑦	0.669	0.855
AtGenExpression Ecotypes Expression ⑦	0.255	0.771
AtGenExpression Light Expression ⑦	-0.188	0.153
AtGenExpression Pathogen Expression ⑦	-0.044	0.282
Co-localization ⑦	1.0	0.886
DOMINE ⑦	0.0	0.0
DOMINE (3did) ⑦	0.0	0.0
DOMINE (DIMA) ⑦	0.0	0.0
DOMINE (Fusion) ⑦	0.0	0.439
DOMINE (LP+DPEA) ⑦	0.0	0.0
DOMINE (ME) ⑦	0.0	0.263
DOMINE (P-value) ⑦	1.0	0.956
DOMINE (RCDP) ⑦	0.0	0.0
DOMINE (RDFF) ⑦	0.0	0.0
DOMINE (iPfam) ⑦	1.0	0.87
GO Biological Process ⑦	0.003	0.762
GO Cellular Component ⑦	1.0	0.0
GO Molecular Function ⑦	0.0	0.744
Interologs (BLAST) ⑦	40.0	1.0
Interologs (InParanoid) ⑦	1.0	0.9
Phylogeny Profiles (Mutual Information) ⑦	0.101	0.517
Phylogeny Profiles (Pearson Correlation) ⑦	0.121	0.841
TAIR Expression Dataset ⑦	0.412	0.761

(b)

Indirect Evidences

Domain Interactions ⑦

	Domains in AT3G22880	Domains in AT5G20850
	PF08423 Rad51, C-terminal	PF08423 Rad51, C-terminal

Domain Interactions:
PF08423(Rad51) PF08423(Rad51) reported in DOMINE(P-value,iPfam)

Homologous Interactions ⑦

Locus A	Locus B	Name A	Name B	Species	Method	PubMed
NP_011106	NP_011021	DMC1	RAD51	yeast	Dosage Rescue	16818607

Co-localization ⑦

AT3G22880 locations		AT5G20850 locations	
Location	Reference	Location	Reference
cytosol	SUBA(subloc)	cytosol	SUBA(subloc)
cytosol	SUBA(wolfpsort)	cytosol	SUBA(wolfpsort)
nucleus	SUBA(swissprot)	nucleus	SUBA(swissprot)
cytosol	SUBA(loctree)	cytosol	SUBA(loctree)

GO Annotations ⑦

[Biological Process] [Molecular Function] [Cellular Component]

AT3G22880   AT5G20850

involved in   involved in

GO:0006281
[DNA repair]

Co-Expression ⑦

[Light] [Development] [Abiostress]

AtGenExpression Developement Co-Expression
Pearson Correlation = 0.669
— AT3G22880 — AT5G20850

Co-Publication ⑦

Title	Journal	Year	PubMed
Isolation and characterisation of the RAD51 and DMC1 homologs from Arabidopsis thaliana.	Mol Gen Genet	1998	9520262
ASY1 mediates AtDMC1-dependent interhomolog recombination during meiosis in Arabidopsis.	Genes Dev	2007	17785529
Interaction between Arabidopsis Brca2 and its partners Rad51, Dmc1, and Dss1.	Plant Physiol	2006	16415210

Figure 1. The Interaction Information page. (**a**) Summary of the interaction and its component proteins. (**b**) Indirect evidences supporting this interaction, including gene co-expressions, domain interactions, shared GO annotations, co-localizations, phylogenetic profile similarities and homologous interactions in other organisms (interologs). Co-publication (two interacting proteins appearing in the same articles) information was not used for interaction prediction, but is also presented here to facilitate literature mining. (**c**) The confidence scores for this interaction. The SVM score is the overall confidence score.

well-studied organisms (human, fruit fly, yeast and worm) are listed in the Homologous Interactions table. The sub-cellular localizations of *Arabidopsis* proteins were collected from the SUBA database (24), which also contains multiple sub-cellular localization data sets reported by different approaches. The co-localization table summarizes the shared localizations in each data set, which are distinctly colored for clarity. In the GO annotations table, the relationship between the annotation terms of two proteins that share a significant parent term is present in a graphical manner. The expression profiles for two

proteins are illustrated by charts in the co-expression table, so users may develop an intuitive idea on how they correlate under different perturbations (e.g. light, development or abiostress). Co-publication (two interacting proteins appearing in the same articles) information was not used for interaction prediction, but is also presented in this page, as it serves as a useful tool to help users look up publications discussing both proteins. The last section of this page lists the confidence scores for this interaction (Figure 1c). The SVM score is the overall confidence score for a predicted interaction. It is usually >0, indicating a

positive prediction. However, for experimentally reported interactions, this score may be negative as our prediction model may not correctly predict all real interactions. The confidence score derived from each indirect evidence is also presented. A detailed account of these scores is provided in the Help/FAQ page.

Graphical interaction network browser and 'My collection'

Each PAIR web page showing interactions includes an integrated interaction network browser, developed from the Adobe Flash^TM-based program Cytoscape Web

(http://cytoscapeweb.cytoscape.org). Figure 2 shows an example network graph obtained by searching PAIR with a set of genes in the sulfur metabolism pathway. The query proteins are displayed as triangle nodes. Other proteins are shown as circle nodes. All nodes are colored by their molecular function annotations. The color scheme is provided in the Help/FAQ page. Interactions are presented as edges between nodes. Experimentally reported interactions are in red. Predicted interactions that are homologous to known interactions in other organisms are in blue. Other predicted interactions are colored grey. The layout of the
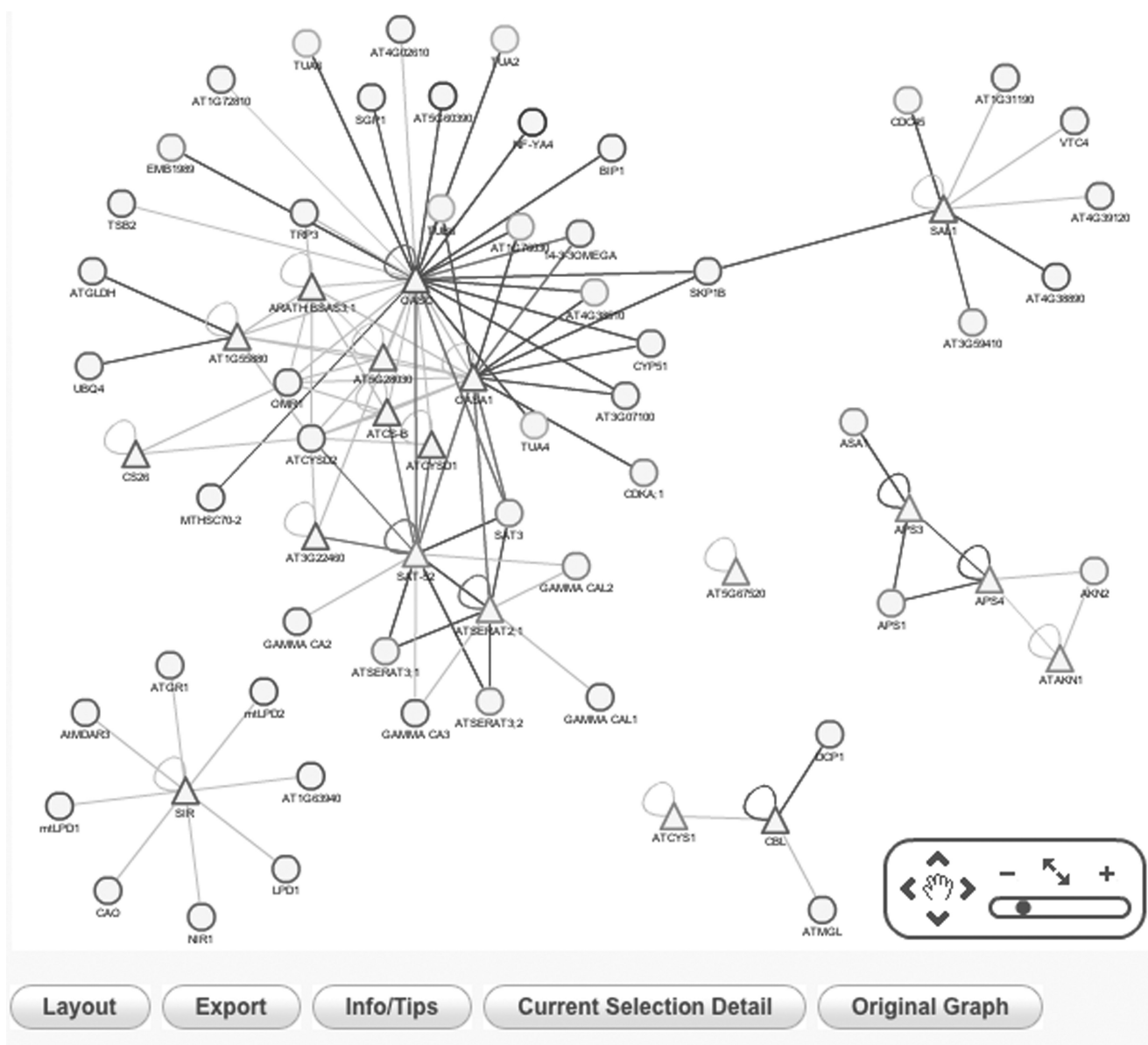
Figure 2. The graphical interaction network browser showing a network of interactions between proteins in the sulfur metabolism pathway. Several core sulfur metabolism proteins (triangle nodes) are used to search PAIR for interactions between them. This network graph is then expanded to include additional interactions. Double click on a node will bring up a window showing all interactions involving this protein and allow selection of some interactions to be added into this graph. Newly added proteins are shown as circle nodes. Nodes are colored according to their protein functions, as detailed in the Help/FAQ page. Experimentally reported interactions (edges) are in red. Predicted interactions that are homologous to known interactions in other organisms are in blue. Other predicted interactions are in grey.

network graph can be changed by applying a number of layout algorithms or by simply dragging the nodes. Double clicking on a node will bring up a window showing all interactions involving this protein. Users can select some of these interactions to be added into the network graph. Once a desired network has been created, it is possible to save it as an image or export it in various data formats, including the Microsoft Excel, Cytoscape SIF, GraphML, PSI 2.5 XML and PAIR XML formats.

We also support the use of 'My collection' where interactions of interest may be stored as browser cookies and later retrieved in the 'My collection' page. This feature will facilitate mining of specific biological pathways. In many PAIR web pages, there is a button to add selected interactions to 'My collection'. Right-clicking on an edge in the interaction network browser will also bring up an option to add this interaction to 'My collection'.

SUMMARY

The PAIR database contains experimentally reported interactions integrated from major interaction repositories and the most comprehensive prediction of the *Arabidopsis* interactome with a high reliability. These predictions were expected to cover 24% of the entire *Arabidopsis* interactome, and their reliability was estimated to be 44%. PAIR features a user-friendly query interface, providing rich annotation on the relationships between two proteins. A graphical interaction network browser has also been integrated into the web interface to facilitate mining of specific pathways. PAIR is a resource not only for large-scale mining of *Arabidopsis* interaction networks but is also an exploratory tool for cell/molecular biologists to understand more about the relationships between the proteins in specific cellular processes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

REFERENCES

1. Krogan,N.J., Cagney,G., Yu,H., Zhong,G., Guo,X., Ignatchenko,A., Li,J., Pu,S., Datta,N., Tikuisis,A.P. *et al.* (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440**, 637–643.
2. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al.* (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.
3. Formstecher,E., Aresta,S., Collura,V., Hamburger,A., Meil,A., Trehin,A., Reverdy,C., Betin,V., Maire,S., Brun,C. *et al.* (2005) Protein interaction mapping: a Drosophila case study. *Genome Res.*, **15**, 376–384.
4. 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.
5. Rual,J.F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M., Ayivi-Guedehoussou,N. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.
6. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A., Koeppen,S. *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
7. Aranda,B., Achuthan,P., Alam-Faruque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.
8. Stark,C., Breitkreutz,B.J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
9. Alfarano,C., Andrade,C.E., Anthony,K., Bahroos,N., Bajec,M., Bantoft,K., Betel,D., Bobechko,B., Boutilier,K., Burgess,E. *et al.* (2005) The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Res.*, **33**, D418–D424.
10. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.
11. Cusick,M.E., Yu,H., Smolyar,A., Venkatesan,K., Carvunis,A.R., Simonis,N., Rual,J.F., Borick,H., Braun,P., Dreze,M. *et al.* (2009) Literature-curated protein interaction datasets. *Nat. Methods*, **6**, 39–46.
12. Morsy,M., Gouthu,S., Orchard,S., Thorneycroft,D., Harper,J.F., Mittler,R. and Cushman,J.C. (2008) Charting plant interactomes: possibilities and challenges. *Trends Plant Sci.*, **13**, 183–191.
13. Geisler-Lee,J., O'Toole,N., Ammar,R., Provart,N.J., Millar,A.H. and Geisler,M. (2007) A predicted interactome for Arabidopsis. *Plant Physiol.*, **145**, 317–329.
14. De Bodt,S., Proost,S., Vandepoele,K., Rouze,P. and Van de Peer,Y. (2009) Predicting protein-protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-expression. *BMC Genomics*, **10**, 288.
15. Cui,J., Li,P., Li,G., Xu,F., Zhao,C., Li,Y., Yang,Z., Wang,G., Yu,Q. and Shi,T. (2008) AtPID: Arabidopsis thaliana protein interactome database–an integrative platform for plant systems biology. *Nucleic Acids Res.*, **36**, D999–D1008.
16. Brandao,M.M., Dantas,L.L. and Silva-Filho,M.C. (2009) AtPIN: Arabidopsis thaliana protein interaction network. *BMC Bioinformatics*, **10**, 454.
17. Yu,H., Braun,P., Yildirim,M.A., Lemmens,I., Venkatesan,K., Sahalie,J., Hirozane-Kishikawa,T., Gebreab,F., Li,N., Simonis,N. *et al.* (2008) High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**, 104–110.
18. 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**, D1009–D1014.
19. Lee,K., Thorneycroft,D., Achuthan,P., Hermjakob,H. and Ideker,T. (2010) Mapping plant interactomes using literature curated and predicted protein-protein interaction data sets. *Plant Cell*, **22**, 997–1005.
20. Lin,M., Hu,B., Chen,L., Sun,P., Fan,Y., Wu,P. and Chen,X. (2009) Computational identification of potential molecular interactions in Arabidopsis. *Plant Physiol.*, **151**, 34–46.

21. Boruc,J., Van den Daele,H., Hollunder,J., Rombauts,S., Mylle,E., Hilson,P., Inze,D., De Veylder,L. and Russinova,E. (2010) Functional modules in the Arabidopsis core cell cycle binary protein-protein interaction network. *Plant Cell*, **22**, 1264–1280.

22. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.

23. Raghavachari,B., Tasneem,A., Przytycka,T.M. and Jothi,R. (2008) DOMINE: a database of protein domain interactions. *Nucleic Acids Res.*, **36**, D656–D661.

24. 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**, D213–D218.