ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 153

International Conference on Applied Mathematics, Modeling and Simulation (AMMS 2017)

Predicted Arabidopsis Interactome Resource-A Network Modeling Method of Integration and Analysis for the Omics Data

Heng Yao and Xin Chen* and Xiaoxuan Wang

Institute of Pharmaceutical Biotechnology, Zhejiang University, 866 Yuhangtang Road, Hangzhou, P.R.China

*Corresponding author

***Abstract*—The Predicted Arabidopsis Interactome Resource (PAIR) is an online database of the functional interactions between Arabidopsis genes. PAIR is inferred by integrating six types of evidence each of which suggests a different aspect of functional associations between Arabidopsis genes and therefore enables extended analysis on the potential functional impacts of the observed omics changes at the physiological level.**

Keywords-omics annotation; arabidopsis; functional interaction network; analysis method

I. INTRODUCTION

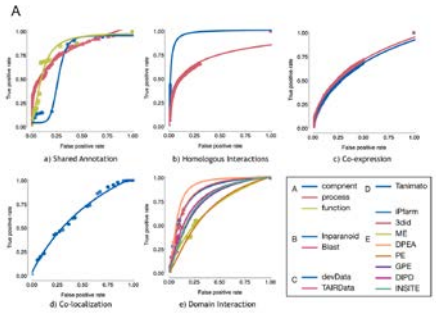
We present the Predicted Arabidopsis Interactome Resource version 5.0 (PAIR v5.0) for analyzing omics data changes. Omics experiments are more and more commonly explored in contemporary plant research [1–3]. Taking RNA-seq as example, transcriptomics results can provide a comprehensive overview on the differences between sample groups in different conditions or at different times. One aim of omics data analysis is to elucidate the biological pathways or functions that are altered to produce the molecular phenotype, so as to understand the regulatory logic behind how and why plants react to specific stimuli or interventions. To understand pathway-level regulations, a reference network that connects functionally linked genes is desired.

PAIR v5.0 is designed to meet this need by inferring a functional interaction network of Arabidopsis genes. This network is built by integrating different evidences of functional associations in multiple forms and finally infers 335301 putative functional interactions, which are expected to cover ~26% of all true protein-protein interactions with ~38% reliability. Below we describe the preparation and evaluation of PAIR v5.0.

II. RESULT

Support vector machine (SVM) [4,5] was used for the inference of significant functional associations from six types of evidence including 17907 homologous interactions in other species (interologs), 517560 gene co-localization, 16233 phylogenetic profile, 173378 shared functional annotations, 1791283 domain interactions, 22240 gene co-expression profiles. All of the data were retrieved before 12/22/2014.

31 features were extracted from these six types of evidences and their discriminatory powers to discriminate the known physically protein-protein interactions from all random protein-protein pairs were estimated by the the Area Under Curve (AUC) in a Receiver Operator Characteristics (ROC) test. Finally, 16 features whose AUC of ROC is over 0.6 were selected for the subsequent inference of functional interactions by SVM (Figure I).



B

Evidence type	Feature	AUC of ROC
Shared Annotation	Component	0.707
	Process	0.809
	Function	0.816
Homologous Interactions	Inparanoid	0.711
	Blast	0.736
Domain InteractionR	iPfam	0.724
	3did	0.768
	ME	0.645
	DPEA	0.634
	PE	0.614
	GPE	0.659
	DIPD	0.641
	INSITE	0.659
Co-localization	Tanimoto	0.649
Co-expression	Development	0.644
	TAIRDataSet	0.669

FIGURE I. RECEIVE OPERATING CHARACTERISTICS CURVES AND AUC VALUES OF THE FEATURES

In the SVM modeling process, we choose physical protein-protein interactions as positive samples. We collected 260013 experimentally reported physical protein interactions from Intact [6], BioGRID [7], BIND [8] and TAIR [9]. If the protein-protein interaction was reported by at least two researches or reported by any low-throughput experiment, this interaction was selected as high confidence interaction for use as positive examples. The 260013 protein-protein interactions were filtered and 6257 high confidence interactions were found. Using a ratio of 1:100, we use randomly generated protein-protein pairs as negative samples.

After parameter optimization using 5-fold cross-validation and grid search, we identified a set of optimal parameter $C = 10^{0.3}$, $\gamma = 0$ for SVM training. The resulting model has 25.79% ± 2.180% sensitivity, and 99.95% ± 0.0099% specificity, which were calculated as below.

$$sensitivity = \frac{TP}{FP + FN} \quad \text{precision} = \frac{TP}{TP + FP}$$

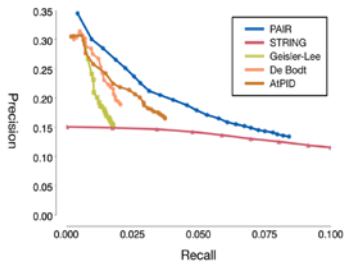
A total of 329,044 interactions were predicted with the model trained with optimal parameters. With this 329044 predicted interactions, we can estimate the size of Arabidopsis interactome by solving the equation 'True positives + False positives = All predicted interactions'. This equation is equivalent to,

$$I \times Sen + (N - I) \times (1 - Spe) = P$$

Where I is the Arabidopsis interactome size; N is the number of all possible protein pairs ($3.758 = 10^8$ pairs between 27416 *Arabidopsis* protein coding genes); P is the number of predicted interactions (329,044); and Sen and Spe are the sensitivity and specificity, respectively, of the prediction model.

Using this method, we estimate that the size of Arabidopsis interactome is $5.088 = 10^5$. Compared to the number of all the possible protein pairs, 1/739 of the Arabidopsis protein pairs were predicted to interact according to this model. This probability is very similar to the observed interaction rate in Yeast (1/775) [10]. Consequently, the reliability of these interactions being true physical protein interactions is 38%. According to this estimated Arabidopsis interactome size, we also estimated the coverage and reliability of other available Arabidopsis interactomes (Figure II-B).

A



B

Data set	Coverage	Reliability
Geisler-Lee	6.864%	16.88%
De Bodt	2.677%	26.25%
STRING	98.98%	2.992%
AtPID	1.510%	6.480%
PAIR v5.0	27.01%	37.94%

FIGURE II. EVALUATION OF NETWORKS

The quality of the inferred interaction network was evaluated, i.e., the fidelity to connect functionally related genes together was evaluated by predicting a gene's function (gene ontology term) using a term enrichment tool (PANTHER [11]) from the terms that have been annotated to this gene's first-degree network neighbors [12]. Results of this evaluation were shown as a precision-recall curve (Figure II-A). Among five interactomes including Geisler-Lee [13], De Bodt [14], STRING [15], and AtPID [16], PAIR v5.0 performance the best, which was reflected by the highest AUC in the precision-recall curves. STRING is the only interactome reached the same high recall region, however, the high recall of STRING was at the expense of very low precision in all regions.

III. METHOD

A. Data Collection and Integration

Six types of evidence were chosen to be used in the prediction. Each of them suggests a certain aspect of functional association [17,18]. Thirty-one features were computed to represent these six types of evidence by different mathematical characterizations.

Gene co-expression: Interacting proteins are often co-expressed. We computed co-expression features from six microarray expression AtGenExpress data sets: AtGenExpress light, AtGenExpress pathogen, AtGenExpress development, AtGenExpress abiotic stress, ecotypes [19–21]), and TAIR (ftp://ftp.arabidopsis.org//Microarrays/analyzed_data/affy_data_1436_10132005.zip). These data sets were all pre-normalized by a robust multi-array average method [22]. Using these data sets, we calculated the Pearson's correlation coefficients for each pair of proteins, which produced six co-expression features.

Shared ontology annotation: Functional related protein coding genes were expected to have similar descriptions in nature language in the biology knowledge-base. Considering using geneontology database as our knowledge-base, a protein-coding gene A is represented by all the geneontology terms annotated by A, define this term list as term(A). Then the functional linkage between a protein/gene pair can be described as a function F(term(A), term(B)). In our model, F is the minimum parent geneontology term shared by A and B. The number of the genes annotated by this parent term and all its children terms has been counted as the Minimum Parent Term Size (MPTS). We calculated the fraction of the MPTS and the size of the whole Arabidopsis genome as the feature score.

Domain interaction: Protein interactions involve physical interactions between their domains. It has been proposed that novel protein interactions can be inferred by known domain interactions. In our prediction method, the domain composition of each protein was annotated according to the Pfam database [23], which assigned one or more of the 2,854 distinctive domains to one or more of the 20,183 Arabidopsis proteins. Known domain-domain interactions were retrieved from the DOMINE database [24]. DOMINE contains two domain interaction datasets inferred from PDB entries (i.e. iPfam and 3did), and 13 datasets predicted by different computational approaches (i.e. ME, RCDP, Pvalue, Fusion, DPEA, PE, GPE, DIPD, RDFF, KGIDDI, INSITE, DomainGA and PP). According to each dataset, we counted the number of interacting domains in a pair of proteins as its feature value. This resulted in 15 domain interaction features.

Co-localization: Considering the physical interactions between proteins, one of the prerequisites of physical interactions is that the proteins must be in the same location in a cell. Thus, if a pair of proteins are detected in the same sub-cell location, they are more likely to interact with each other than other proteins which are not. We use the sub-location database SUBA3 [25,26] as the data source of the co-localization proteins of Arabidopsis. For every protein, we construct a PLV (protein location vector) of n×0 or 1 in order to represent the sub-location conditions. If there is a record showing the protein is in this sub-cell location, we mark it as 1, otherwise, 0. Then we calculate the Tanimato Correlation Score of each PLV as the value of their col-localization feature.

Phylogenetic profile: The proteins which are co-evolved are more likely functional related [27,28], i.e., if a pair of proteins are co-presence and co-absence across different genomes, there might be functional interactions between them. The phylogenetic profiles are extracted from Roundup [29] as of the end of 2014. As in the feature of co-localization, we constructed the vector of n×0 or 1 to represent the presence or absence in one genome of any species. Then we calculated three feature values using different methods, including Mutual Information, Pearson's Correlation, Tanimato Correlation of each pair of protein vectors.

Homologous interactions: Homologous proteins may have more similar functions than other proteins, and therefore if there is an interaction proteins pair was found in a species, we can consider that the homologous proteins of the proteins in the interaction pair might have a higher probability to interact [17,30] and functionally associated. We collected the interactomes of four species, including Homo sapiens, Caenorhabditis elegans, Drosophila melanogaster, and Saccharomyces cerevisiae from following databases: IntAct, BioGRID, MINT and DIP. The data were updated to the end of 2014. We calculated the feature of interology using the equation as follows:

$$H(A,B) = \max\{I(i,j) \cdot \min[l(A,i), l(B,i)]\}$$

I this equation, A, B represent the protein-coding gene pair to be calculated whereas i and j represent the possible homologous interacting proteins in one of other four species. I(i, j) is an indicator tells if there is a homologous interaction pair in a certain species. l(A, i) or l(B, j) is an indicator reflects how much is the homology between two proteins(genes), and in this study, there are two different methods were used to generate this indicator: one is the e-Value from the BLAST, and the other one is the ortholog mapping score in the InParanoid database [31].

We chose the value of the most homologous proteins between (A, j) and (B, i) as the representation of the homologous of the homology between interaction pairs (A, B) and (i, j) if (i, j) exists. Then we identified the most non-homologous interaction pair through all interactomes from four species as the representation of their value of homologous interaction feature. Since we used two different kinds of l(A, i) or l(B, j), we finally computed two values of this feature.

B. Network Evaluation

We evaluated the newly inferred Arabidopsis interactome (PAIR 5.0) together with other four available interaction networks, i.e., Geisler-Lee, De Bodt, STRING and AtPID. Since PAIR v5.0 was inferred from data that were released up to the 12/22/2014, we selected 1313 genes from the Gene Ontology database with new annotations that were added after the time of our data collection. These genes have a total of 19178 annotations and 4930 of them were newly updated since 2015. These genes and their annotations were used to evaluate the quality of interaction networks.

For each of the 1313 target genes, we first identified all first-degree neighbors of a target gene in the PAIR interaction network to create a gene set. This gene set was analyzed by PANTHER to find enriched annotation terms.

For each p-value cutoff (as in PANTHER), we counted a) how many terms are predicted (N); b) how many of the N terms are correct (consistent with the known 19178 gene annotations, Y); and c) how many of the 4930 new annotations were covered by the N terms. Precision and Recall were calculated as follows.

$$\text{Recall} = \frac{X}{4930}$$

$$\text{Precision} = \frac{Y}{N}$$

ACKNOWLEDGMENT

This work is supported by Zhejiang Provincial Natural Science Foundation of China grant (LR13C020001) and National Natural Science Foundation of China grant (31571356).

REFERENCE

[1] B. Berger, J. Peng, M. Singh, Computational solutions for omics data, Nat. Publ. Gr. 14 (2013).

[2] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, J. Tegnér, Data integration in the era of omics: current and future challenges, BMC Syst. Biol. 8 (2014) I1.

[3] R. Saha, A. Chowdhury, C.D. Maranas, Recent advances in the reconstruction of metabolic models and integration of omics data, Curr. Opin. Biotechnol. 29 (2014) 39–45.

[4] S. Winters-Hilt, A. Yelundur, C. McChesney, M. Landry, Support Vector Machine Implementations for Classification & Clustering, BMC Bioinformatics. 7 (2006) S4.

[5] C.-C. Chang, C.-J. Lin, LIBSVM : A Library for Support Vector Machines, ACM Trans. Intell. Syst. Technol. 2 (2013) 1–39.

[6] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, R. Apweiler, IntAct: an open source molecular interaction database., Nucleic Acids Res. 32 (2004) D452-5.

[7] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, M. Tyers, BioGRID: a general repository for interaction datasets., Nucleic Acids Res. 34 (2006) D535-9.

[8] G.D. Bader, D. Betel, C.W. V Hogue, BIND: the Biomolecular Interaction Network Database., Nucleic Acids Res. 31 (2003) 248–50.

[9] P. Lamesch, T.Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D.L. Alexander, M. Garcia-Hernandez, A.S. Karthikeyan, C.H. Lee, W.D. Nelson, L. Ploetz, S. Singh, A. Wensel, E. Huala, The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools., Nucleic Acids Res. 40 (2012) D1202-10.

[10] H. Yu, P. Braun, M.A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R.R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. de Smet, A. Motyl, M.E. Hudson, J. Park, X. Xin, M.E. Cusick, T. Moore, C. Boone, M. Snyder, F.P. Roth, A.-L. Barabási, J. Tavernier, D.E. Hill, M. Vidal, High-quality binary protein interaction map of the yeast interactome network., Science. 322 (2008) 104–10.

[11] H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, P.D. Thomas, PANTHER version 11: expanded annotation data from Gene

Ontology and Reactome pathways, and data analysis tool enhancements, Nucleic Acids Res. 45 (2017) D183–D189.

[12] J.Z. Wang, Z. Du, R. Payattakool, P.S. Yu, C.-F. Chen, A new method to measure the semantic similarity of GO terms, Bioinformatics. 23 (2007) 1274–1281.

[13] J. Geisler-Lee, N. O'Toole, R. Ammar, N.J. Provart, A.H. Millar, M. Geisler, A Predicted Interactome for Arabidopsis, Plant Physiol. 145 (2007) 317–329.

[14] S. De Bodt, S. Proost, K. Vandepoele, P. Rouzé, Y. Van de Peer, Predicting protein-protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-expression, BMC Genomics. 10 (2009) 288.

[15] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K.P. Tsafou, M. Kuhn, P. Bork, L.J. Jensen, C. von Mering, STRING v10: protein-protein interaction networks, integrated over the tree of life, Nucleic Acids Res. 43 (2015) D447–D452.

[16] J. Cui, P. Li, G. Li, F. Xu, C. Zhao, Y. Li, Z. Yang, G. Wang, Q. Yu, Y. Li, T. Shi, AtPID: Arabidopsis thaliana protein interactome database an integrative platform for plant systems biology, Nucleic Acids Res. 36 (2007) D999–D1008.

[17] D. Rhodes, S. Tomlins, S. Varambally, Probabilistic model of the human protein-protein interaction network, Nature. (2005).

[18] B. Shoemaker, A. Panchenko, Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners, PLoS Comput. Biol. (2007).

[19] M. Schmid, T. Davison, S. Henz, U. Pape, M. Demar, A gene expression map of Arabidopsis thaliana development, Nature. (2005).

[20] J. Kilian, D. Whitehead, J. Horak, D. Wanke, The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV‐B light, drought and cold stress responses, The Plant. (2007).

[21] H. Goda, E. Sasaki, K. Akiyama, The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access, The Plant. (2008).

[22] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, T.P. Speed, Exploration, normalization, and summaries of high density oligonucleotide array probe level data, Biostatistics. 4 (2003) 249–264.

[23] M. Punta, P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E.L.L. Sonnhammer, S.R. Eddy, A. Bateman, R.D. Finn, The Pfam protein families database, Nucleic Acids Res. 40 (2012) D290–D301.

[24] S. Yellaboina, A. Tasneem, D. V. Zaykin, B. Raghavachari, R. Jothi, DOMINE: a comprehensive collection of known and predicted domain-domain interactions, Nucleic Acids Res. 39 (2011) D730–D735.

[25] S.K. Tanz, I. Castleden, C.M. Hooper, M. Vacher, I. Small, H.A. Millar, SUBA3: a database for integrating experimentation and prediction to define the SUBcellular location of proteins in Arabidopsis, Nucleic Acids Res. 41 (2013) D1185–D1191.

[26] C.M. Hooper, S.K. Tanz, I.R. Castleden, M.A. Vacher, I.D. Small, A.H. Millar, SUBAcon: a consensus algorithm for unifying the subcellular localization data of the Arabidopsis proteome, Bioinformatics. 30 (2014) 3356–3364.

[27] C.-S. Goh, A.A. Bogan, M. Joachimiak, D. Walther, F.E. Cohen, Co-evolution of proteins with their interaction partners 1 1Edited by B. Honig, J. Mol. Biol. 299 (2000) 283–293.

[28] F. Pazos, A. Valencia, Similarity of phylogenetic trees as indicator of protein-protein interaction, Protein Eng. Des. Sel. 14 (2001) 609–614.

[29] T.F. DeLuca, J. Cui, J.-Y. Jung, K.C. St. Gabriel, D.P. Wall, Roundup 2.0: enabling comparative genomics for over 1800 genomes, Bioinformatics. 28 (2012) 715–716.

[30] M.S. Scott, G.J. Barton, Probabilistic prediction and ranking of human protein-protein interactions., BMC Bioinformatics. 8 (2007) 239.

[31] E.L.L. Sonnhammer, G. Ostlund, InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic, Nucleic Acids Res. 43 (2015) D234–D239.