# SUPPLEMENTAL METHODS

## Inference and Evaluation of the Predicted Interactions in the Predicted Arabidopsis Interactome Resource

### Overview

The interaction prediction methods used for an earlier version of the Predicted Arabidopsis Interactome Resource (PAIR, V2.3) have been previously described in Lin et al., (2009). The data quality and user interface of the latest version (V3.3) are discussed in Lin et al. (2011). Information on the detailed interaction prediction protocols, accuracy evaluation results, and user interface usages can also be found at the PAIR website (http://www.cls.zju.edu.cn/pair/, the trailing '/' is necessary). The PAIR project started as a simple attempt to infer *Arabidopsis* interactions by homology mapping. Within three years, it has gradually evolved into a dedicated effort aiming to provide the most accurate interactome predicted by the state-of-the-art machine learning approach. The last major release (V2) predicted *Arabidopsis* interactions based on four types of indirect evidence: gene co-expression, protein domain interaction, annotation similarity, and sub-cellular co-localization (Lin et al., 2009). In the current major release (V3), we have considerably improved the quality of indirect evidence. Two new types of indirect evidence, interolog and phylogenetic profile similarity, have been added. In addition, we changed the way in which co-expression features were calculated. To improve sensitivity to detect transient protein co-expression under specific conditions, the expression similarities were measured separately within the five well-organized AtGenExpress perturbation datasets (light, ecotypes, pathogen, development and abiotic stress) (Schmid et al., 2005; Kilian et al., 2007; Goda et al., 2008), as well as within all available microarray data in TAIR.

With careful engineering considerations from the lessons learned in previous prediction studies, 145,494 interactions involving 9,480 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 predicted interactions was estimated to be 43%. These predicted interactions had 1,584 (26.4%) overlap with the 5,990 experimentally reported interactions available in IntAct (Aranda et al., 2010), BioGRID (Stark et al., 2006), BIND (Alfarano et al., 2005) and TAIR (Swarbreck et al., 2008) as of July 23$^{rd}$, 2010, among which only 4,545 interactions reported before June 15$^{th}$, 2009 were used to train the prediction model and evaluate its accuracy.

The current release of PAIR (V3.3) contains a total of 149,900 interactions, including the 5,990 recently compiled experimentally reported interactions and the 145,494 predicted interactions. 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. Though not used in interaction prediction, co-publication (two proteins appearing in the same article) information is also shown in the interaction details page to facilitate mining of related literature on the predicted interactions.

## PAIR V3 Prediction methods

### *Gold standard data*

PAIR V3 used a support vector machine (SVM) model to predict interactions. Known pairs of interacting proteins (positive examples) and non-interacting proteins (negative examples) are required to train an SVM prediction model. To assemble an accurate positive example dataset, we retrieved and integrated experimentally determined *Arabidopsis* interactions from IntAct (Aranda et al., 2010), BioGRID (Stark et al., 2006), BIND (Alfarano et al., 2005) and TAIR (Swarbreck et al., 2008). As shown in Table SM1, this resulted in 4545 interactions involving 2276 proteins.

**Table SM1. Composition of gold-standard positive examples.**

| Database | Number of Interactions | Number of Proteins | Date |
|---|---|---|---|
| IntAct | 3,503 | 1,724 | June 8, 2009 |
| BioGRID | 945 | 756 | May 27, 2009 |
| TAIR | 1,709 | 1,326 | May 27, 2009 |
| BIND | 970 | 465 | June 9, 2009 |
| Total | 4,545 | 2,276 | June 15, 2009 |

Unlike interacting proteins, it is rare to find reported non-interacting proteins. Several studies have attempted to choose negative examples as pairs of proteins with different cellular localizations. However, this approach did not seem to improve model accuracy. Furthermore, it was argued that this would lead to biased estimation of accuracy, because the constraints placed on the cellular distribution of the negative examples could make the prediction task easier (Ben-Hur and Noble, 2006). Therefore, we followed the approach described in (Zhang et al., 2004) to select negative examples as random protein pairs that did not overlap with positive examples. The resulting negative examples may contain several potentially interacting protein pairs. Yet given the ratio of interacting protein pairs versus non-interacting protein pairs in yeast (1:775) (Yu et al., 2008), this level of contamination likely is acceptable. By this way, we generated 4545 negative examples.

A good example data set shall be able to well (homogenously) represent interactions among the entire proteome. Therefore, we compared the protein distribution in our example interactions to the protein distribution in the entire *Arabidopsis* proteome based on four categorizing systems. The TAIR database provided three classification systems, i.e. the molecular function based system, the cellular component based system, and the biological process based system, which are collectively known as the GO Slim classification systems. GO Slim organizes *Arabidopsis* genes according to broad GO ontology categories. In most cases, a GO slim category corresponds to an existing term in the GO ontology. Genes that are annotated to the term itself, or to any of the

children of the GO Slim term, are included in the corresponding GO Slim category. In some cases, the GO Slim category encompasses more than one term. In our analysis, genes annotated to the "unknown categories" (i.e. "molecular function unknown", "biological process unknown" and "cellular component unknown") were ignored. As shown in Figure SM1A, proteins in our interaction examples and proteins in the entire proteome showed similar histogram based on GO Slim biological process categories. The Pearson's correlation coefficient between these two distributions is 0.97 (p-value = $1.36 \times 10^{-8}$). Similarly, we observed a correlation coefficient of 0.63 (p-value = 0.0075) based on GO slim molecular function categories (Figure SM1B) and a correlation coefficient of 0.81 (p-value = 0.0001) based on GO slim cellular component categories (Figure SM1C). Besides the annotation based categories, sequence based protein families also serve as a good classification system to compare protein distributions. According to the Pfam database (Finn et al., 2010), all *Arabidopsis* proteins were classified into 2855 families. The frequency distribution of the proteins in our interaction examples and that in all proteins showed a correlation coefficient of 0.79 (p-value < $1 \times 10^{-10}$, Figure SM1D). Therefore, these analyses indicated that proteins in our interaction examples represented a good sample of the *Arabidopsis* proteome.

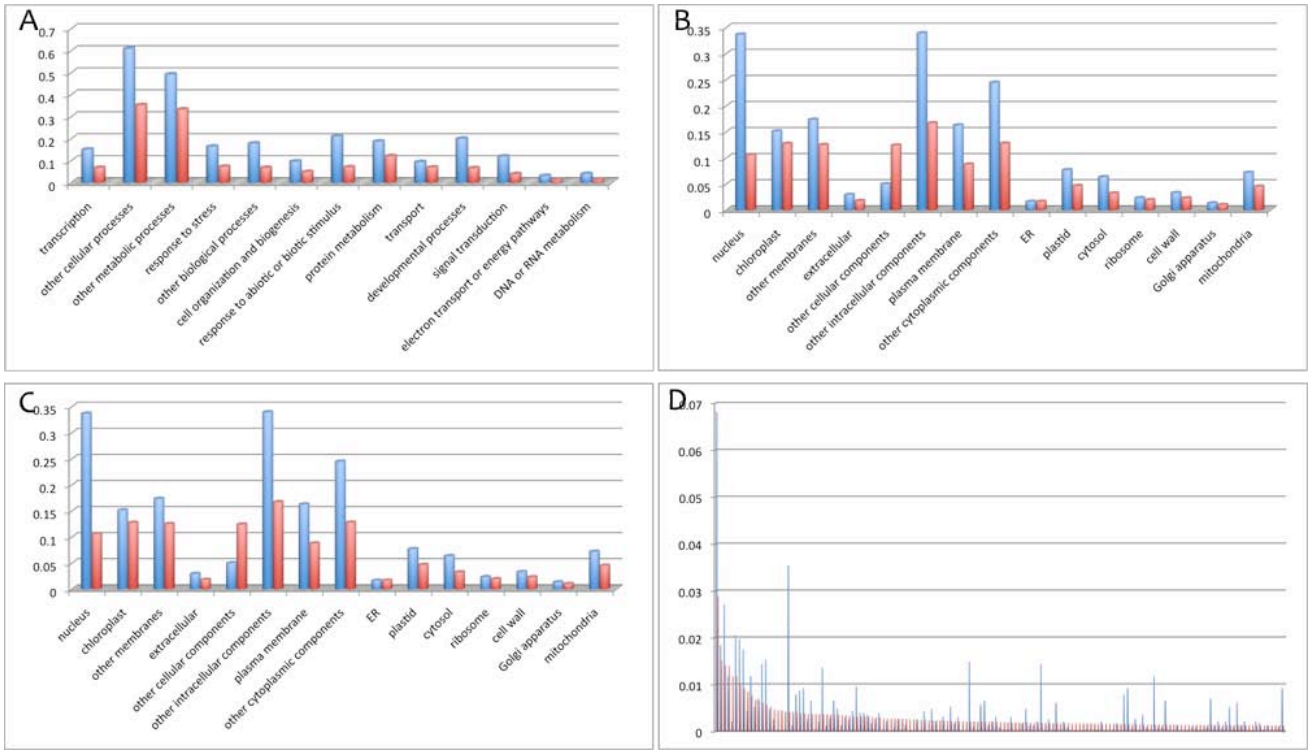


**Figure SM1: Protein distributions in our example dataset and those in the entire *Arabidopsis* proteome.** Blue bars show the fraction of example interacting proteins in each category. Red bars show the fraction of all *Arabidopsis* proteins in each category. **A)** Protein distributions based on the GO Slim biological process categories. The Pearson's correlation coefficient between two distributions is 0.97 (p-value = $1.36 \times 10^{-8}$). **B)** Protein distributions based on the GO Slim molecular function categories. The correlation coefficient is 0.63 (p-value = 0.0075). **C)** Protein distributions based on the GO Slim cellular component categories. The correlation coefficient is 0.81 (p-value = 0.0001). **D)** Protein distributions based on the Pfam protein family classifications. The correlation coefficient is 0.79 (p-value < $1 \times 10^{-10}$). In this diagram, only the largest 150 families are shown for clarity.

***Indirect evidence***

Six types of indirect evidence were chosen to be used in the PAIR prediction methods. This choice was based on earlier evaluations of the usefulness of many indirect lines of evidence (Rhodes et al., 2005; Shoemaker and Panchenko, 2007). Twenty four features were computed to represent these six indirect lines of evidence by different mathematical characterizations. Therefore, in computation, each protein pair (interactions or non-interactions) was represented by a 24 dimensional feature vector (Table SM2). The six types of indirect evidence are as follows.

**1. Gene co-expression:** Interacting proteins are often co-expressed. We computed co-expression features from six microarray expression data sets: the AtGenExpress light, AtGenExpress pathogen, AtGenExpress development, AtGenExpress abiotic stress, AtGenExpress ecotypes (Schmid et al., 2005; Kilian et al., 2007; Goda et al., 2008), and the TAIR data set 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 (Irizarry et al., 2003). Using these data sets, we calculated the Pearson's correlation coefficients for each pair of proteins, which produced six co-expression features.

**2. 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 methods, the domain composition of each protein was annotated according to the Pfam database (Finn et al., 2010), 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 (Raghavachari et al., 2008). DOMINE contains two domain interaction datasets inferred from PDB entries (i.e. iPfam and 3did), and 8 datasets predicted by different computational approaches (i.e. ME, RCDP, P-value, Fusion, LP+DPEA, RDFF, and DIMA). According to each dataset, we counted the number of interacting domains in a pair of proteins as its feature value. This resulted in 10 domain interaction features.

**3. Shared annotation (GO):** Interacting proteins were often expected to share similar molecular functions, to involve in the same biological processes, and to localize in the same cellular components. The Gene Ontology (GO), structured as a directed acyclic graph, provides a standardized way to annotate proteins from these aspects and offers a good basis to compare their annotations. In the GO annotation term graph, strongly related terms are expected to have a shared parent term close to the leaf terms and narrow in concept. The fraction of proteins annotated to this shared parent term and all its child terms reflects the unexpectedness that two proteins share this particular parent annotation term. Consequently, this fraction could be used as a score to measure the similarity between two annotation terms. The smaller this score is, the more similar the two terms are. Extending the concept of term similarity, we defined the protein annotation similarity score as the smallest term similarity score between their annotation terms. In TAIR, there were 151,568 annotations assigned to *Arabidopsis* genes. By focusing specifically on terms in one of the three aspects (i.e. molecular function, biological process, and cellular component), we computed three annotation similarity scores as three features.

**4. Co-localization:** Interacting proteins are usually co-localized. In addition to the GO cellular component annotations, the Arabidopsis Sub-cellular Database (SUBA) (Heazlewood et al., 2007) also provides high-quality information on protein sub-cellular localizations. Let $f(L)$

be the fraction of all proteins presented in a location $L$, and let $I(i,L)$ be an indicator function showing whether protein $i$ is observed to localize in $L$. The co-localization feature value was calculated as follows: $F(A,B) = \max\{-\log[f(L) \times I(A,L) \times I(B,L)]\}$; i.e., the co-localization feature value for a protein pair $(A,B)$ was computed as the negative logarithm of the fraction of all proteins presented in the most specific location where both proteins *A* and *B* were observed to localize.

**5. Phylogenetic profile:** The phylogenetic profile feature measures the co-presence or co-absence of a pair of non-homologous genes across genomes. It is assumed that pairs of non-homologous genes that are always present or absent together, are likely to have co-evolved. Co-evolving proteins are more likely to interact (Pellegrini et al., 1999; Goh et al., 2000; Pazos and Valencia, 2001). For each protein in *Arabidopsis*, its orthologs in other genomes were identified according to the RoundUp database (Wall et al., 2003). This generated the phylogenic profile of a protein. The degree of similarity between two profiles was measured with both the mutual information score (Wu et al., 2005) and the Pearson's correlation coefficient. Therefore two phylogenetic profile similarity features were computed.

**6. Homologous interactions (interologs):** Homologous proteins often retain similar functions. A pair of genes interacting in one species often indicates their homologs interacting in other species. This indirect evidence has been used to predict protein interactions in several previous studies (Geisler-Lee et al., 2007; Cui et al., 2008; De Bodt et al., 2009). Based on this idea, protein interactions in *Homo sapiens*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Saccharomyces cerevisiae* were retrieved from IntAct (Aranda et al., 2010), BioGRID (Stark et al., 2006), and HPRD (Keshava Prasad et al., 2009). The homologous relationships between *Arabidopsis* genes and genes in these four species were determined by PSI-blast. Two homologous interaction features were computed. The first one characterized the support of an *Arabidopsis* interaction using the below equation. $H(A,B) = \max\{I(i,j) \times \min[L(A,i),L(B,j)]\}$, where $A$ and $B$ are *Arabidopsis* proteins; $i$ and $j$ are proteins in another species; $I(i,j)$ is an indicator function showing whether protein $i$ and $j$ are found to interact in any other species; $L(A,i)$ is the negative log E-value of the blast match between $A$ and $i$. The second homologous interaction feature characterized the support of an *Arabidopsis* interaction using the below equation. $H(A,B) = \max\{I(i,j) \times \min[M(A,i),M(B,j)]\}$ where $M(A,i)$ is the ortholog mapping score between $A$ and $i$ provided by the InParanoid database (Ostlund et al., 2010).

**Table SM2. Composition of features.**

| Type of indirect evidence | Number of Features |
|---|---|
| Co-expression | 6 |
| Domain Interaction | 10 |
| Co-localization | 1 |
| Shared Annotation (GO) | 3 |
| Phylogenetic Profile | 2 |
| Homologous Interaction | 2 |
| **Total** | **24** |

***Predictive strength of indirect evidence***

It must be noted that all these types of indirect evidence are weak, and one line of indirect evidence alone cannot be used to accurately predict interactions. This is why many studies looking at any one line of evidence (Matthews et al., 2001; Kim and Subramaniam, 2006; Wu et al., 2006; Geisler-Lee et al., 2007; Obayashi et al., 2009) have concluded that indirect evidence is of limited value in interaction prediction. Nevertheless, most of these studies showed that the indirect evidence provided some value, although limited, for interaction prediction. For example in Matthews et al. (2001) the authors showed that, although interolog is not a very strong indicator of interaction, the presence of an interolog in another species still provides additional support for the validity of an interaction reported from a low-confidence high-throughput experiment. Consistently, our analysis also indicated that these indirect lines of evidence were weak. The feature strength diagrams demonstrated that individual features could only accurately predict a small fraction of interactions when these features reached extreme values. Therefore, their predictive powers should be combined to create an improved prediction model, which no longer requires one feature to take an extreme value to make an accurate prediction, and instead requires specific combinations of feature values.

One way to assess the discrimination power of a feature is to compare its value distribution in interaction examples against its value distribution in non-interaction examples. Informative features are expected to show apparent differences. Therefore, we divided the value range of each feature into 20 equal bins. The fraction of positive examples and the fraction of negative examples that fell into each bin were calculated to make the corresponding distribution histograms. We plotted these distribution histograms and the associated Likelihood Ratio (LR) curve in the feature strength diagrams. The LR for each bin is defined as the fraction of interaction examples above this bin divided by the fraction of non-interaction examples above this bin. Above a bin means that a feature value is expected to better indicate an interaction than all values in this bin. For features describing co-expression, co-localization, homologous interaction, phylogenetic profile and domain interaction, this means that a feature value is greater than the biggest value in this bin; but for features describing shared annotation, above a bin means that a feature value is lower than the smallest value in this bin. For convenience in comparison, it is desired to have all LR curves increasing with the horizontal axes. Therefore, for features describing shared annotation, the horizontal axes are plotted as (1 - feature value). Figure SM2 shows one typical feature strength diagram for each type of indirect evidence. A complete set of feature strength diagrams for all the 24 features used in interaction prediction can be found at http://www.cls.zju.edu.cn/pair/helpfaq/indirect_evidences.html.

It is clear that in these diagrams, no sharp differences exist between the value distributions in interaction examples and those in non-interaction examples. Pearson's correlation between the interaction and non-interaction distributions ranged from 0.324 to 1. Kolmogorov-Smirnov test also indicated that these interaction and non-interaction distributions were not significantly different, displaying distances from 0.1 to 0.4, corresponding to a p-value range of [1.0, 0.313] (no significance to reject the hypothesis that they were the same). However, it is also noted that there are always clear deviations between the interaction and non-interaction distributions for each feature when the feature takes an extreme value. In best-case scenarios (at the rightmost point), these features always lead to significant LRs (i.e. >10 folds). Hence, these features are all weak but nonetheless useful. A delicate scheme to exploit their combined discriminative power is therefore necessary for accurate prediction of interactions.


**Figure SM2. Feature strength diagrams.** For each feature, green columns and yellow columns show the fraction of interaction examples and the fraction of non-interaction examples that fall into each bin. Blue curves show the Likelihood Ratio (LR) that a protein pair is likely interacting when the observed feature value is no worse than the value shown on the horizontal axis. The horizontal axis shows the feature value. The left vertical axis shows the fraction of protein pairs. The right vertical axis shows the LR.

***Prediction of interactions***

Statistical learning algorithms can learn the patterns of feature value combinations in interacting proteins by contrasting the feature values of interacting proteins to those of non-interacting proteins. Some statistical learning algorithms known as the "white-box algorithms" are able to express the learned patterns as human interpretable rules. However, they are often less accurate in prediction than those "black-box algorithms" which cannot express the learned patterns as interpretable rules. The support vector machine (SVM) algorithm is a "black-box algorithm" famous for its high precision (Brown et al., 2000; Burbidge et al., 2001). PAIR used the SVM algorithm to predict interactions.

The SVM algorithm is well documented (Winters-Hilt et al., 2006). For implementation, we used the public software package LIBSVM version 2.88 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). The radial basis function kernel was chosen to be used for its best performance (data not shown). In this setup, two parameters, the regularization parameter $C$ and the kernel width parameter $\gamma$, were optimized by a grid search in an empirical range (0.001 to 1,000 for $C$ and $3 \times 10^{-5}$ to $4 \times 10^{4}$ for $\gamma$), maximizing the F1 measure as defined below. Model performance with each parameter set was estimated by 10-fold cross-validation.

Based on whether a real positive or negative example in the test dataset is predicted as positive or negative, the prediction result for each protein pair can be divided into four categories: true positive (TP), when a real positive example is predicted correctly; true negative (TN), when a real negative example is predicted correctly; false positive (FP), when a real negative example is predicted as positive; and false negative (FN), when a real positive example is predicted as negative. These four indicators are collectively known as the confusion matrix (Table SM3). Many accuracy measurements are derived from the confusion matrix, such as sensitivity [or recall; TP/ (TP + FN)], specificity [TN/(TN + FP)], and precision [TP/(TP + FP)]. In the application of predicting interactions, the ability to accurately and comprehensively recognize potential interactions is critical. Therefore, the measurements of precision and recall are most relevant. Recall conveys the idea of what proportion of the real interactions can be recognized by the model. Precision measures how confident we are when the model predicts an interaction. Note that the calculation of precision depends on the ratio of positive examples to negative examples in the test dataset. Therefore, the precision observed in testing does not represent the expected precision in real application, if the positive-to-negative ratio in the test dataset is not equal to the expected positive-to-negative ratio in real application. Giving equal weights to precision (p) and recall (r), van Rijsbergen (1979) introduced the "F1 measure" as the harmonic mean of precision and recall [2rp/(r+p)].

Accuracy of the resulting model was evaluated by a more precise 100-iteration bootstrap experiment. In each iteration, 1/10 of the example data was randomly left out from training and these 1/10 data was used to test the model and compute the confusion matrix. After 100 iterations, the mean and standard deviation (SD) of the confusion matrix (Table SM3), together with other derived accuracy measurements, were computed. The optimal model produced 81.36% ± 0.31% sensitivity, and 90.25% ± 0.28% specificity.

Although the above prediction model worked well on our example dataset consisting of equal number of positive and negative examples, its application in interactome prediction was

still limited. Assuming that the ratio of interacting protein pairs out of random pairs in *Arabidopsis* is similar to that in yeast, i.e. ~1/775 (Yu et al., 2008), its 90.25% specificity would translate to a large number of false positives. And the precision of its predicted interactions in interactome prediction would drop to ~0.95%, worthless for human inspection. Based on our past experiences (Cai et al., 2003; Xue et al., 2004; Xu et al., 2007), the SVM algorithm often displays an increased accuracy for the type of examples with either larger quantity or greater diversity. Thus one way to increase specificity without significantly compromising sensitivity is to expand the negative examples in training data. Therefore, 447,480 randomly generated negative examples were added to the training dataset, raising the positive to negative ratio to 1:100. With the same protocol, this enlarged example dataset produced an optimal SVM model showing 24.47% ± 0.15% sensitivity, and 99.96% ± 0.0015% specificity (Table SM3). Applying this model to all possible *Arabidopsis* protein pairs, 145,494 potential interactions were identified.

Comparing the above two models, the one trained with equal number of positive and negative examples had a high sensitivity (81.36%), or it would less likely to miss a real interaction. Therefore it was called the high-coverage model, or 1:1 model. However, this good coverage came at a cost. The confidence of its predicted interactions (precision) could be as low as 0.95%. In contrast, the model trained with 100 times more negative examples showed a high specificity (99.96%), or it would less likely to wrongly predict an interaction. It was thus called the high-confidence model, or 1:100 model. However, its predicted interactions were expected to cover only 24.47% of the entire interactome.

Obviously, the high-confidence predictions are more suitable for human inspection. They can be queried at the PAIR website together with experimentally reported interactions. The high-coverage predictions contain too many false positives and are therefore less useful for direct manual analysis, but they may still be useful to reconfirm interactions identified by other low-confidence interaction detection approaches. For example, the total number of *Arabidopsis* protein pairs is $2.29 \times 10^8$. Among these protein pairs, the high-coverage model predicted $2.64 \times 10^7$ interactions. This is to say that the high-coverage model predicted $2.03 \times 10^8$ (~89%) non-interacting protein pairs. On the other hand, the high-coverage model was able to recognize ~80% of all interactions. If another low-confidence interaction detection approach is independent from ours, and the reported interactions that are not predicted by our high-coverage model are considered false-positives, we will expect to remove 89% false-positive reports at the expense of 20% of true interactions.

**Table SM3. The confusion matrix of our prediction models.**

| | | **Actual** | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| **Predicted** | **Positive** | TP: 3697.03±14.01(1:1 model)<br>1112.12±5.02 (1:100 model) | FP: 443.14±8.45(1:1 model)<br>161.80±5.41 (1:100 model) |
| | **Negative** | FN: 847.97±14.01(1:1 model)<br>3432.88±5.02 (1:100 model) | TN: 4101.86±8.45(1:1 model)<br>454338.19±5.41 (1:100 model) |

## Evaluation of PAIR V3 prediction accuracy

Two issues need to be clarified before the accuracy assessment results are presented. First, PAIR is evaluated as a protein interaction network instead of as a functional linkage network. Although our prediction method might as well predict *bona fide* "non-physical" interactions, these predictions were regarded as false positives during evaluation. In other words, with the reported level of coverage and precision, PAIR is a collection of predicted physical protein interactions. It may also be regarded as a functional linkage network with a higher level of coverage and precision. However, at this time, we do not have data to show its accuracy as a functional linkage network. Second, the interactome dataset searchable at the PAIR website is essentially a compilation including the PAIR V3 predictions and the experimentally reported interactions deposited in the major interaction databases before July 23$^{rd}$, 2010. In the accuracy assessments described below, only the PAIR V3 predictions, without the additional experimentally reported interactions, were evaluated.

Before using newly reported interactions to evaluate prediction accuracy, it was noted that the high-coverage predictions included the entire set of high-confidence predictions, indicating the self-consistency of our prediction methods.

### *Rediscovery of the newly reported interactions*

Two external benchmark datasets were used to verify the accuracy of PAIR. The first dataset contains newly reported interactions that were not included in the major interaction databases at the time (June 15$^{th}$, 2009) when all the data used for interaction prediction were assembled. This independent evaluation dataset was retrieved from an update (as of December 27$^{th}$, 2009) of the BioGRID database (Stark et al., 2006), which includes 448 new interactions that had been double-checked to avoid any overlap with our example interactions. As shown in Supplemental Dataset 11, 115 (25.67%) of these new interactions could be successfully recognized by our high-confidence model. This sensitivity was comparable to the expected one (24.47%), indicating that this model is free of severe over-fitting problem. In contrast, the high-coverage model recognized 334 (74.55%) of these new interactions in BioGRID. This number was slightly lower than the expected sensitivity (81.36%), indicating that the high-coverage model slightly over-fitted the training examples.

The other external benchmark dataset was reported by a recent article (Apr. 2010) in The Plant Cell (Boruc et al., 2010), published two months after the latest version of our predicted interactome was released. This experiment tested 917 protein pairs among 58 proteins using two complementary interaction assays, bimolecular fluorescence complementation (BiFC) and high-confidence yeast-two-hybrid (Y2H), resulting in 357 interactions, of which 293 had not been reported previously. Among these 357 reported interactions, the PAIR high-confidence model predicted 137 of them (Supplemental Dataset 12). The coverage was 38.37%. Among the 293 newly reported interactions, the high-confidence model predicted 107 of them. The coverage was similarly high (36.52%). By contrast, this sensitivity was over five times higher than those of other available predicted interactomes (Supplemental Dataset 12). Altogether, PAIR predicted 291 interactions from the 917 experimentally tested protein pairs, of which 137 were confirmed. The precision of high-confidence predictions therefore reached 47.08%. The coverage and precision observed with this external benchmark dataset (38.4% / 47.08%) were higher than our

in-house estimations (24.5% / 43.5%, the estimation of precision, 43.5%, is discussed later). This finding might be due to the fact that 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 multiple lines of indirect evidence based on which our predictions were made. Therefore, the high coverage and precision observed with this dataset may not apply to the entire predicted interactome. However, these results certainly show that with an ever-growing body of protein characteristics data, PAIR has the potential to predict the *Arabidopsis* interactome at higher levels of coverage and precision. On the other hand, the PAIR high-coverage model predicted 305 of the 357 reported interactions (Supplemental Dataset 12). The coverage was 85.43%. Among the 293 newly reported interactions, the high-coverage model predicted 241 of them. The coverage was similarly high (82.25%).

***Estimation of the Arabidopsis interactome size***

The above accuracy evaluations using external benchmark datasets indicated that for the high-confidence model, the sensitivity estimated during model training can be safely generalized to interactome prediction. Assuming that the estimated specificity is also accurate, the number of interactions predicted by the high-confidence model, 145,494, can be used to estimate the size of *Arabidopsis* interactome.

The equation "True Positives + False Positives = All predicted positives" can be re-written as:

$$N_{\text{int}} \times Sen + (N_{all} - N_{\text{int}}) \times (1 - Spe) = N_{pred}$$

where $N_{\text{int}}$ represents the interactome size, $N_{all}$ is the number of all possible *Arabidopsis* protein pairs (2.29 x $10^8$), $N_{pred}$ is the number of predicted interactions (145,494), $Sen$ and $Spe$ are the sensitivity and specificity of our model. Solving this equation gave an estimated *Arabidopsis* interactome size of 2.58 x $10^5$. In other words, 1 out of 1014 random protein pairs is expected to interact. This ratio is similar to the one observed in yeast (1/775) (Yu et al., 2008). Considering that 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. This also indirectly supported our assumption that the estimated specificity for our high-confidence model is accurate.

As mentioned above, the estimation of sensitivity and specificity is independent from the positive-to-negative ratio in the test dataset, but the estimation of precision (confidence of the predicted interactions) is dependent on this ratio. Therefore, the precision scores estimated during our model training stage, using testing datasets with positive to negative ratio 1:1 or 1:100, do not reflect the expected precision scores in interactome prediction, where a positive to negative ratio of 1:1014 is expected. The prediction precision in interactome prediction can be calculated as $N_{\text{int}} \times Sen / N_{pred}$. Therefore, based on the above estimated *Arabidopsis* interactome size, the high-confidence predictions are expected to have a precision of 43.52%, and the high-coverage predictions are expected to have a precision of 0.64%.

### *Summary of accuracy evaluation results*

Two prediction models, the high-coverage model and the high-confidence model, were constructed to predict the *Arabidopsis* interactome. During training, the high-confidence model was expected to have a sensitivity of 24.47%, and a specificity of 99.96%; the high-coverage model was expected to have a sensitivity of 81.36%, and a specificity of 90.25%. Two external datasets containing only newly reported interactions were used to evaluate the sensitivities of both models. One dataset contains interactions between well-studied cell cycle proteins (Boruc et al., 2010). For this dataset, both models were able to predict interactions with the expected sensitivities. The other dataset contains new interactions between random proteins (the BioGRID database update). For this dataset, the high-coverage model showed slightly decreased sensitivity (~74.55% vs. 81.36%). To be conservative, we report the lowest sensitivity either observed during training or during external evaluation as the expected sensitivity for each model. Therefore, the coverage of high-confidence interactions is deemed as 24.47%, and the coverage of high-coverage interactions is deemed as 74.55%.

Because the high-confidence model showed consistent sensitivity during training and external evaluation, it is expected to be free of severe over-fitting problem. Assuming that the sensitivity and specificity of the high-confidence model estimated during training are accurate, we had an estimation of the size of *Arabidopsis* interactome, $2.58 \times 10^5$, which corresponds to a rate of protein interaction similar to that observed in yeast (Yu et al., 2008). Based on this estimated interactome size, the precision of model predictions (the reliability of the predicted interactions) in interactome prediction can be estimated. The high-confidence predictions are expected to have a reliability of 43.52%. The high-coverage predictions are expected to have a reliability of 0.64%.

Because the high-confidence dataset is more suitable for manual analysis, the PAIR website only provides query to this dataset, and all the discussions in the main text are based on the high-confidence dataset. The high-coverage dataset is available at the PAIR download page (http://www.cls.zju.edu.cn/pair/download.html).

## SUPPLEMENTAL REFERENCES

**Alfarano, C., Andrade, C.E., Anthony, K., Bahroos, N., Bajec, M., et al.** (2005). The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res **33:** D418-424.

**Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., et al.** (2010). The IntAct molecular interaction database in 2010. Nucleic Acids Res **38:** D525-531.

**Ben-Hur, A., and Noble, W.S.** (2006). Choosing negative examples for the prediction of protein-protein interactions. BMC Bioinformatics **7 Suppl 1:** S2.

**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.

**Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., Jr., and Haussler, D.** (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A **97:** 262-267.

**Burbidge, R., Trotter, M., Buxton, B., and Holden, S.** (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. Comput Chem **26:** 5-14.

**Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X., and Chen, Y.Z.** (2003). SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res **31:** 3692-3697.

**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:** D999-1008.

**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.

**Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R., and Bateman, A.** (2010). The Pfam protein families database. Nucleic Acids Res **38:** D211-222.

**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.

**Goda, H., Sasaki, E., Akiyama, K., Maruyama-Nakashita, A., Nakabayashi, K., et al.** (2008). The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. Plant J **55:** 526-542.

**Goh, C.S., Bogan, A.A., Joachimiak, M., Walther, D., and Cohen, F.E.** (2000). Co-evolution of proteins with their interaction partners. J Mol Biol **299:** 283-293.

**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:** D213-218.

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

**Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., et al.** (2009). Human Protein Reference Database--2009 update. Nucleic Acids Res **37:** D767-772.

**Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J., and Harter, K.** (2007). The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant J **50:** 347-363.

**Kim, Y., and Subramaniam, S.** (2006). Locally defined protein phylogenetic profiles reveal previously missed protein interactions and functional relationships. Proteins **62:** 1115-1124.

**Lin, M., Shen, X., and Chen, X.** (2010). PAIR: the predicted Arabidopsis interactome resource. Nucleic Acids Res. **39 (Database Issue),** D1134-1140.

**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.

**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.

**Obayashi, T., Hayashi, S., Saeki, M., Ohta, H., and Kinoshita, K.** (2009). ATTED-II provides coexpressed gene networks for Arabidopsis. Nucleic Acids Res **37:** D987-991.

**Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D.N., Roopra, S., Frings, O., and Sonnhammer, E.L.** (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res **38:** D196-203.

**Pazos, F., and Valencia, A.** (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng **14:** 609-614.

**Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O.** (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A **96:** 4285-4288.

**Raghavachari, B., Tasneem, A., Przytycka, T.M., and Jothi, R.** (2008). DOMINE: a database of protein domain interactions. Nucleic Acids Res **36:** D656-661.

**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.

**Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D., and Lohmann, J.U.** (2005). A gene expression map of Arabidopsis thaliana development. Nat Genet **37:** 501-506.

**Shoemaker, B.A., and Panchenko, A.R.** (2007). Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Comput Biol **3:** e43.

**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:** D535-539.

**Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., and Huala, E.** (2008). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res **36:** D1009-1014.

**Wall, D.P., Fraser, H.B., and Hirsh, A.E.** (2003). Detecting putative orthologs. Bioinformatics **19:** 1710-1711.

**Wu, J., Mellor, J.C., and DeLisi, C.** (2005). Deciphering protein network organization using phylogenetic profile groups. Genome Inform **16:** 142-149.

**Wu, X., Zhu, L., Guo, J., Zhang, D.Y., and Lin, K.** (2006). Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. Nucleic Acids Res **34:** 2137-2150.

**Xu, H., Lin, M., Wang, W., Li, Z., Huang, J., Chen, Y., and Chen, X.** (2007). Learning the drug target-likeness of a protein. Proteomics **7:** 4255-4263.

**Xue, Y., Li, Z.R., Yap, C.W., Sun, L.Z., Chen, X., and Chen, Y.Z.** (2004). Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. J Chem Inf Comput Sci **44:** 1630-1638.

**Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., et al.** (2008). High-quality binary protein interaction map of the yeast interactome network. Science **322:** 104-110.

**Zhang, L.V., Wong, S.L., King, O.D., and Roth, F.P.** (2004). Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinformatics **5:** 38.