Computational identification of protein-protein interactions in model plant proteomes

Ziyun Ding & Daisuke Kihara

Protein-protein interactions (PPIs) play essential roles in many biological processes. A PPI network provides crucial information on how biological pathways are structured and coordinated from individual protein functions. In the past two decades, large-scale PPI networks of a handful of organisms were determined by experimental techniques. However, these experimental methods are time-consuming, expensive, and are not easy to perform on new target organisms. Large-scale PPI data is particularly sparse in plant organisms. Here, we developed a computational approach for detecting PPIs trained and tested on known PPIs of Arabidopsis thaliana and applied to three plants, Arabidopsis thaliana, Glycine max (soybean), and Zea mays (maize) to discover new PPIs on a genome-scale. Our method considers a variety of features including protein sequences, gene co-expression, functional association, and phylogenetic profiles. This is the first work where a PPI prediction method was developed for is the first PPI prediction method applied on benchmark datasets of Arabidopsis. The method showed a high prediction accuracy of over 90% and very high precision of close to 1.0. We predicted 50,220 PPIs in Arabidopsis thaliana, 13,175,414 PPIs in corn, and 13,527,834 PPIs in soybean. Newly predicted PPIs were classified into three confidence levels according to the availability of existing supporting evidence and discussed. Predicted PPIs in the three plant genomes are made available for future reference.

Identification of protein-protein interactions (PPIs) is important for understanding how proteins work together in a coordinated fashion in a cell to perform cellular functions. PPIs are essential for individual protein functions, forming various cellular pathways, and are also involved in the development of diseases. PPI data is directly useful for identifying protein multimeric complexes1,2, identifying biological pathways as well as predicting protein function3–5. For more on the application side, PPIs are also important targets for drug design6 and artificial design of protein complexes7.

There are experimental methods for determining individual PPIs, such as co-immunoprecipitation8, fluorescence resonance energy transfer9, and surface plasmon resonance10. Ultimately, biophysical methods such as nuclear magnetic resonance spectroscopy (NMR)11,12, X-ray crystallography13, and electron microscopy14, can be used to determine the tertiary structure of protein complexes to obtain detailed atomic or molecular level information about how the proteins interact. Moreover, from late 1990’s, PPIs have been determined in a large-scale using yeast-two hybrids15–18 and affinity chromatography combined with mass spectrometry19–22. However, experimental methods have several shortcomings for detecting PPIs. First, these experimental methods are time-consuming and labor-intensive. Second, the applicability of experimental methods depends on how well assay protocols are established in target organisms. Also, a method may not work on some classes of proteins23,24. Third, it is known that experimental methods often have difficulty in identifying weak interactions, which leaves out many transient interactions25. Fourth, it is discussed that the results of large-scale methods often have a substantial disagreement with each other, which may be partly due to false positives and false negatives26–28.

Consequently, PPIs have been identified only for a limited number of organisms; moreover, the coverage of PPI networks is very small for the majority of organisms. This is particularly true for plant species. Table 1 shows the statistics of experimentally identified PPIs in representative plant species taken from the BioGRID database29. Surprisingly, except for Arabidopsis thaliana, virtually no other plant species have experimentally determined PPI data available. Even for Arabidopsis, known PPI data cover interactions with only about 34.55% of proteins.

1Department of Biological Sciences, Purdue University, West Lafayette, IN, 47907, USA. 2Department of Computer Science, Purdue University, West Lafayette, IN, 47907, USA. 3Department of Pediatrics, University of Cincinnati, Cincinnati, OH, 45229, USA. Correspondence and requests for materials should be addressed to Z.D. (email: ding48@purdue.edu) or D.K. (email: dkihara@purdue.edu)
do not interact. We construct negative sets in two different ways as mentioned by Guo67. One is to pair proteins by genetic experimental systems. We reduced the number of PPIs after removal of short proteins of less than 50 amino acid residues and PPI identified in Arabidopsis and surface plasmon resonance. The downloaded known spectrometry, co-immunoprecipitation, fluorescent resonance energy transfer, isothermal titration calorimetry, and protein pairs using SVM. The SVM protocol was named as SVMloc and SVMrand corresponding to the PPIloc and PPIrand data sets, respectively, and a rigorous nested cross-validation evaluation was performed. These methods typically use a machine learning framework and consider various features of proteins as input. Protein features used for PPI prediction include occurrence of functional domains31–33, short sequence patterns (e.g. n-grams, auto-covariation)34–38, interlog (interaction inferred from homology)39–43, function44, similarity in phylogenetic trees47–49, phylogenetic profiles50, gene expression51, and protein tertiary structures52–57. Although many approaches were explored, there are not many works that applied developed methods to provide new proteomics-scale PPI predictions. Existing works are mainly limited to microbial genomes and eukaryotes40,43,56,58–65, and only applications to the plant domain are for Arabidopsis thaliana41,63,64 and Oryza sativa (rice)63,68.

In this work, we developed a computational method for PPI prediction, named PPIP (PPI prediction for Plant genomes) and applied to three major plant proteomes, Arabidopsis thaliana, Zea mays (corn), and Glycine max (soybean). To capture different aspects of proteins that are relevant to PPIs, we used a combination of four features for predicting PPIs, i.e. protein sequence properties, protein functional similarity, co-expression patterns, and phylogenetic profile similarity. To provide a confidence level of predictions, two machine learning methods, support vector machine (SVM) and random forest (RF), were separately trained on different features, and commonly predicted PPIs by the two methods were considered to have high confidence. The machine learning methods were trained on known PPIs from Arabidopsis. The accuracy on the testing dataset of Arabidopsis achieved a high accuracy of over 90%. PPIP predicted 50,220, 13,175,414, and 13,527,834 confident PPIs in Arabidopsis, corn, and soybean, respectively.

Examples of predicted novel PPIs with high confidence are discussed. All confident predictions are provided on our lab website (http://kiharalab.org/PPIP_results/) so that they can be referenced by plant biologists.

Table 1. Statistics of the number of experimentally determined PPIs in representative plant species. The statistics of PPIs were taken from the BioGRID database. The numbers of protein genes were taken from the KEGG database.

| Organism          | Common Name | Number of Protein genes | Identified unique PPIs | Unique Proteins in PPIs | Fraction of proteins involved in known PPIs (%) |
|------------------|-------------|-------------------------|-----------------------|-------------------------|-----------------------------------------------|
| Arabidopsis thaliana | mouse cereals | 27,636                  | 35,908                | 9,574                   | 34.55                                         |
| Zea mays         | corn        | 37,376                  | 13                    | 21                      | 0.056                                         |
| Glycine max      | soybean     | 46,993                  | 39                    | 43                      | 0.092                                         |
| Oryza sativa     | rice        | 35,679                  | 90                    | 72                      | 0.202                                         |
| Solanum lycopersicum | tomato      | 25,613                  | 107                   | 44                      | 0.172                                         |
| Solanum tuberosum | potato      | 28,463                  | 2                     | 3                       | 0.011                                         |

Other representative plant species cover even less protein involved in known PPIs. Thus, it is apparent that plants are largely lagged behind from PPI studies. In this omics era when various types of large-scale data are combined and used for formulating hypotheses and to interpret experimental data, PPI networks are fundamental reference data to have for studying an organism.

To complement experimental methods for identifying PPIs, several computational methods have been developed50. These methods typically use a machine learning framework and consider various features of proteins as input. Protein features used for PPI prediction include occurrence of functional domains31–33, short sequence patterns (e.g. n-grams, auto-covariation)34–38, interlog (interaction inferred from homology)39–43, function44, similarity in phylogenetic trees47–49, phylogenetic profiles50, gene expression51, and protein tertiary structures52–57. Although many approaches were explored, there are not many works that applied developed methods to provide new proteomics-scale PPI predictions. Existing works are mainly limited to microbial genomes and eukaryotes40,43,56,58–65, and only applications to the plant domain are for Arabidopsis thaliana41,63,64 and Oryza sativa (rice)63,68.

In this work, we developed a computational method for PPI prediction, named PPIP (PPI prediction for Plant genomes) and applied to three major plant proteomes, Arabidopsis thaliana, Zea mays (corn), and Glycine max (soybean). To capture different aspects of proteins that are relevant to PPIs, we used a combination of four features for predicting PPIs, i.e. protein sequence properties, protein functional similarity, co-expression patterns, and phylogenetic profile similarity. To provide a confidence level of predictions, two machine learning methods, support vector machine (SVM) and random forest (RF), were separately trained on different features, and commonly predicted PPIs by the two methods were considered to have high confidence. The machine learning methods were trained on known PPIs from Arabidopsis. The accuracy on the testing dataset of Arabidopsis achieved a high accuracy of over 90%. PPIP predicted 50,220, 13,175,414, and 13,527,834 confident PPIs in Arabidopsis, corn, and soybean, respectively.

Examples of predicted novel PPIs with high confidence are discussed. All confident predictions are provided on our lab website (http://kiharalab.org/PPIP_results/) so that they can be referenced by plant biologists.

**Results**

**Constructing a benchmark dataset of known Arabidopsis PPIs.** First, we tested two machine learning prediction algorithms in our prediction method, PPIP, namely, support vector machine (SVM) and random forest (RF) on the dataset of known Arabidopsis PPIs obtained from the TAIR database69 (Additional File 1: Table S1). These PPIs were determined by experiments including X-ray crystallography, affinity-capture mass spectrometry, co-immunoprecipitation, fluorescent resonance energy transfer, isothermal titration calorimetry, and surface plasmon resonance. The downloaded known Arabidopsis PPI dataset contained 4,908 PPIs, which were reduced to 4,759 PPIs after removal of short proteins of less than 50 amino acid residues and PPI identified in Arabidopsis and surface plasmon resonance. The downloaded known Arabidopsis PPI dataset contained 4,908 PPIs, which were reduced to 4,759 PPIs after removal of short proteins of less than 50 amino acid residues and PPI identified in Arabidopsis and surface plasmon resonance.

To train and test a machine learning method, we also need a negative dataset, i.e. a dataset of protein pairs that do not interact. We construct negative sets in two different ways as mentioned by Guo67. One is to pair proteins from different cellular localizations and thus highly unlikely to interact with each other. The cellular localization information was downloaded from the TAIR database. The dataset with the positive set and the negative set constructed in this way is named as PPIloc. Another one is to randomly pair proteins in the positive set and then exclude pairs that are already in the list of positive interacting pairs. The dataset with the positive set and the negative set constructed in this way is named as PPIrand. Both PPIloc and PPIrand included the equal number of interacting and non-interacting pairs; thus there are 9,518 pairs in total.

For training and testing RF, protein pairs that lack co-expression information needed to be removed. This reduced the number of interacting pairs to 3,427. By adding the equal number of non-interacting protein pairs either from PPIloc or PPIrand, the total number of the dataset for RF has become 6,854. The dataset was then split into a training, a validation, and a testing set, respectively, and a rigorous nested cross-validation evaluation was performed to evaluate the prediction accuracy of PPI.

**The design of PPIP.** PPIP predicts if a pair of proteins is likely to have physical interaction or not in physiological condition from the proteins’ sequence and proteomic features. As illustrated in Fig. 1, for a query protein pair, their physical interaction is predicted from physicochemical property features of amino acid sequences of the protein pairs using SVM. The SVM protocol was named as SVMloc and SVMrand corresponding to the PPIloc and PPIrand data sets, respectively.
and PPIrand dataset used, respectively. The features used were hydrophobicity, hydrophilicity, side-chain volumes, polarity, polarizability, solvent-accessible surface area, and net charge index (NCI) of side-chains. In parallel, the complementary features of gene co-expression, functional similarity, and the phylogenetic profile were used to make another independent prediction by RF. The RF protocol was named as RFloc and RFrand corresponding to refer to the PPIloc and PPIrand dataset used, respectively. Predictions with RF were performed in two settings, one with all the features and the other without gene expression features (thus three functional similarity features and the phylogenetic profile) because gene expression data is currently not available for corn and soybean. See Methods for more details about the features.

**Prediction performance on the known Arabidopsis PPIs.** On the PPIloc and PPIrand datasets of known PPIs and non-interacting protein pairs of Arabidopsis, parameters of SVM and RF were trained and tested using six-fold nested cross-validation. In this rigorous validation procedure, the dataset is split into six subsets, and prediction accuracy was measured on each of the subsets using parameters optimized on the rest of the five subsets. See Methods and Supplementary Tables S2 and S3 for more details of this procedure.

Using SVM, the overall prediction accuracies for the PPIloc and PPIrand datasets were 91.9% and 70.8%, respectively (Supplementary Table S4). Thus, SVM performed better on the negative set with protein pairs from different cellular locations than on negative protein pairs that were randomly combined from the interacting pairs (Supplementary Table S4). This is consistent with the conclusion in the paper by Guo, who performed a similar comparison of negative datasets.

On the other hand, RFrand trained on PPIrand performed better than RFloc, which was trained on PPIloc. This order was consistently observed when the eight and the four features were used in RF. The accuracy of RF8rand and RF4rand were 92.0% and 92.6% accuracy, respectively. On the PPIloc, the accuracies were lower, 80.0% and 79.6% for RF8loc and RF4loc, respectively (Supplementary Table S5).

An advantage of random forest is that it can provide the importance of each feature in making correct classification using two metrics, the mean decrease of accuracy (MDA) and the mean decrease of Gini importance (MDGI) (see Methods and Supplementary Note). As shown in Supplementary Table S6, two functional association scores, IAS, PAS, were found to be the two most important features for both RF8loc and RF4loc models. The IAS score was calculated based on the frequency of two GO terms annotating interacting proteins while the PAS score was calculated from the co-occurrence of two GO terms in PubMed abstracts. Therefore, it is reasonable that these two scores contribute largely to classifying interacting and non-interacting protein pairs, because they evaluate biological contexts of GO terms. The results in the table also showed that the phylogenetic profile was more informative than gene expression features.

**Comparison with STRING confidence scores.** We compared our prediction with the confidence score of protein-protein interactions provided in the STRING database. Since SVM used in the PPIP pipeline perform binary classifications and do not provide probability values that are comparable to the STRING scores, we used SVM regression for this comparison. Two types of SVM regression models were used, SVM-ε and SVM-υ.

SVM-ε controls the error tolerance in the training set whereas SVM-υ uses an additional parameter υ to control the proportion of the data points to be used as support vectors. We trained these two SVM regression models using the same hyperparameters as SVMloc and SVMrand. As for RF, we used the fraction of decision trees in RF that vote for protein-protein interaction as the probability.

First, we checked how consistent the predictions by SVM-ε and SVM-υ were with the SVM models in PPIP if a probability value of 0.5 in the SVM regressions was used to convert a probability value of SVM-ε and SVM-υ to binary prediction. On the PPIloc dataset, the prediction by SVMloc-ε was consistent with SVMloc on 99.2% (9445 among 9518) of the cases while SVMloc-υ was consistent with SVMloc on 97.8% of the cases (9312 among 9518).
RFrand recorded a small false positive rate of 4.36%. Finally, as designed in the PPIP pipeline, we took the overlap data (so that the RF model can run with gene expression features), which remained 1,048,575 pairs in the dataset.

The recall values of SVM were somewhat lower than what was observed on the TAIR-based dataset (0.8880 for SVMloc and 0.823, respectively, the combination of the two methods showed a higher value of 0.980. Similarly, on PPI rand, RFrand performed the best with AUC of 0.98 and STRING came the second (AUC: 0.86) (Fig. 2B). Thus, on the both datasets, at least one of our methods showed substantially better AUC than STRING.

Based on these results (Fig. 2), we combine SVMloc and RF rand in the PPIP pipeline (Fig. 1) since they achieved a very high AUC, substantially better than STRING, in the genome-scale prediction in Arabidopsis, corn, and soybean.

**PPI detection by taking overlap between SVM and RF.** In Supplementary Table S7, we examined PPI detection performance by combining SVM and RF on the two datasets, PPIloc and PPI rand. As shown in the table, precision improved by taking consensus: On PPIloc while the precision of SVMloc and RF loc were 0.947 and 0.823, respectively, the combination of the two methods showed a higher value of 0.980. Similarly, On PPI rand, the combination of the two methods showed the perfect precision of 1.0. Note that the improvement of precision was a tradeoff with the accuracy, which decreased because apparently PPIs are missed if they are not detected by both SVM and RF. However, we consider that maintaining a high precision is more important when it comes to a genome-scale prediction because producing many false positive would be a serious concern.

We have further tested the performance of SVMloc and RF rand on a large negative dataset of 2,038,222 protein pairs that are assembled by exhaustively pairing proteins from different cellular locations. The false positive rate of SVMloc on this dataset was 63.7%. We also ran RF rand after excluding pairs that do not have gene expression data (so that the RF model can run with gene expression features), which remained 1,048,575 pairs in the dataset. RF rand recorded a small false positive rate of 4.36%. Finally, as designed in the PPIP pipeline, we took the overlap between SVMloc and RF rand, which yielded a very small false positive rate of 2.68%. The results confirm that the design of PPIP was effective in reducing false positives and that PPIP is suitable for a genome-scale prediction.

**Prediction performance on the BioGRID Arabidopsis Dataset.** We further tested the prediction ability of our prediction method PPIP on a different Arabidopsis dataset, which was obtained from the BioGRID database. BioGRID contained PPI data that were not included in the TAIR database, partly because it was updated more recently. In total, 3,280 Arabidopsis physical PPIs which have been verified by at least two experiments and not included in the TAIR-based dataset were found in BioGRID. The data were further pruned by the sequence identity cutoffs, 80%, 50%, and 30% (Table 2). The overlaps between the training dataset (PPI loc and PPI rand) were excluded. The sequence identity is a measurement of the protein sequence similarity from BLAST. Prediction by RF was applied for a smaller fraction of PPIs, only to those which have co-expression information. The data sets used by SVM and RF with different sequence identity cutoffs are provided in Supplementary Table S8. While an agreement of 1–6 test sets among the total of positives, i.e. the sum of “True Positives” and “False Negatives” in Supplementary Table S4. On the other hand, the recall values of RF were in the same range as the value observed on the TAIR-based dataset (0.7642 for RF loc, 0.7610 for RF rand, 0.8984 for RF loc, 0.9050 for RF rand), which can be computed in the same way). RF’s two predictions with different feature sets, the full features and the feature set without gene expression features, yielded almost identical recall.
the percentage of protein pairs among all the possible protein pairs that satisfied the CC FunSim score criterion. SVMloc and RFrand, were 0.0081%, 7.19%, and 3.77% out of all the possible protein pairs for

The Summary of the number of predicted PPIs in the three plant genomes. CC > 0.4, the number and the percentage of protein pairs among all the possible protein pairs that satisfied the CC FunSim score criterion of over 0.4; %all; SVMloc and RFrand predicted PPIs among pairs that satisfied the CC > 0.4 criterion by SVMloc and RFrand, respectively; Common, commonly predicted PPIs by SVMloc and RFrand. Degree exponent, the parameter value of the power-law distribution of PPIs (Fig. 3). %all is the percentage relative to the all possible protein pairs of the organism while %cc is the percentage relative to the protein pairs that satisfied CC > 0.4.

Table 3. The prediction accuracy (recall) on the BioGRID PPI dataset. The PPIs were clustered by the sequence identity cutoffs of 80%, 50%, and 30% to reduce similar sequences. The sequence identity of protein pairs was computed with the Needleman-Wunsch (global sequence alignment) algorithm implemented in the nwalign python library. On this dataset, we evaluated recall, i.e. the fraction of PPIs in the datasets that were correctly predicted as interacting protein pairs. SVM trained by PPIloc or PPIrand were named as SVMloc or SVMrand; RF trained by PPIloc or PPIrand were named as RFloc or RFrand. (a) RF with the eight features was able to be applied only for PPIs that have gene co-expression data available. The numbers in the parentheses count such PPIs with expression data available. (b) Recall is the fraction of PPIs that are correctly predicted. In the parentheses, SVM recall values measured on the PPIs with co-expression data i.e. the same dataset as used for prediction by RF with the eight features, are shown. (c) The values show the recall of RF using the eight features including the gene expression features. Results with the four feature combinations that only use the functional association scores and the phylogenetic profile are provided in the parentheses.

Among the two SVM models, SVMloc showed a higher recall. When the two RF models were compared, RFrand achieved a higher recall than the counterpart, RFloc. These results are consistent with what we observed on the TAIR-based dataset (Supplementary Tables S4 and S5), which would justify our earlier choice of combining SVMloc and RFrand for the genome-scale PPI predictions to be discussed in the subsequent sections.

PPI prediction for three plant genomes. Next, we applied PPIP to the three plant genomes, Arabidopsis, Zea mays (corn), and Glycine max (soybean). The genome sequences of the three plants were downloaded from the UniProt database\(^ \text{a} \). Since the number of all possible protein pairs in the whole genome is too large, we applied PPIP only for protein pairs that are likely to co-locate in a cell, having a sufficient similarity in their Cellular Component (CC) Gene Ontology (GO) category terms\(^ b \), which describe the sub-cellular locations of proteins. Since many protein genes in corn and soybean do not have GO term annotations in UniProt, we used a function prediction method, PFP\(^ 73,74 \) to predict GO terms to supplement annotations to proteins. PFP is one of the top performing function prediction methods, which performs better than conventional methods, e.g. BLAST\(^ 75 \), as was also demonstrated in a community-wide function annotation assessment, CAFA\(^ 76,77 \). From PFP, only high confidence GO predictions with a score of over 10,000 were used\(^ 73 \). The similarity of CC terms of two proteins was mined from the distribution of the CC-FunSim score of predicted PPIs in three previous papers that\(^ a \) showed an average pairwise similarity of CC GO terms\(^ b \). Protein pairs with a FunSim score of CC terms over 0.4 were subject to the prediction with PPIP. This cutoff was determined from the distribution of the CC-FunSim score of predicted Arabidopsis PPIs in three previous papers that made predictions based on the assumption that PPIs are conserved across species\(^ 2,6,4 \) (Supplementary Fig. S1). Proteins that do not have CC annotations even with PFP prediction were also discarded from the PPI prediction. Applying this pre-screening reduced the number of protein pairs to 21.36% (133,074,361 pairs), 13.54% (24,814,793 pairs), and 15.27% (54,814,995 pairs) of all possible protein pairs for Arabidopsis, corn, and soybean, respectively. This pre-screening process would likely miss some true PPIs that do not satisfy the CC similarity criteria, nevertheless, we decided to apply the process because having a common subcellular co-localization can serve as additional supporting evidence of PPIs.

The numbers of predicted PPIs in the three genomes by PPIP are summarized in Table 3. For protein pairs that satisfied the CC GO term similarity, SVMloc and RFrand were independently applied, and commonly predicted PPIs by SVMloc and RFrand were selected. Among protein pairs with CC-FunSim > 0.4, SVMloc selected about 10%, 56% and 56% as interacting pairs while RFrand predicted 1.83%, 17.45%, and 18.06% as interacting, for Arabidopsis, corn, and soybean, respectively (Table 3). The final PPI predictions, which are the PPIs predicted commonly by SVMloc and RFrand were 0.0081%, 7.19%, and 3.77% out of all the possible protein pairs for Arabidopsis, corn, and soybean, respectively. Compared to the fraction of known PPIs in several other well-studied organisms in
Supplementary Table S9, the fraction of *Arabidopsis*, corn, and soybean PPIs from the current study is at the same level. Particularly, the fraction of predicted PPIs for *Arabidopsis* (0.0081%) seems relatively small, but this fraction is consistent in Table S9. All the predicted PPIs for the three plant genomes are available as Supplementary Data on our lab website (http://kiharalab.org/PIPI_results/).

It is known that the degree distribution of a PPI network of an organism follows a power-law distribution, i.e. the histogram of the number of interactions (called the degree) for each protein is well approximated with a power-law, $p(k) \sim k^{-\gamma}$, where $k$ is the fraction of proteins with a certain number of interactions, and $\gamma$ is a parameter called the degree exponent, which determines the slope of the distribution\(^{78,91,42}\). Figure 3 shows that the PPIs of the three plants detected in the current work follow the power-law, with $\gamma$ being 1.362 in *Arabidopsis*, 0.204 in corn, and 0.401 in soybean. Smaller degree exponents for corn and soybean indicate that these two plants have more hub proteins that interact with many proteins.

Next, we compared our PPI prediction on *Arabidopsis* with three existing genome-scale prediction results. These three works used a very different approach for prediction, the interlog concept\(^{79}\), which assumes that interactions of orthologous proteins across different species are conserved. Geisler-Lee et al.\(^{61}\) and De Bodt et al.\(^{64}\) used the same reference organisms, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *H. sapiens*, whereas Dutkowski et al.\(^{64}\) used the same four organisms with two more organisms, *M. musculus*, and *R. norvegicus* (Supplementary Table S10). Supplementary Table S11 shows the number of PPIs predicted by the three works. All the three works predicted about the same number of PPIs, 14,009 to 19,974. The works by Geisler-Lee et al. and De Bodt et al. made a very similar number of predictions, which is probably due to them using the same set of reference organisms. In Supplementary Table S11, we compared commonly predicted PPIs by pairs of works including our method, PPIP. Geisler-Lee et al. and De Bodt et al. had the largest number of common predictions, although the common predictions would be small considering the same approach and the reference organisms they used. Common predictions by other pairs including pairs with PPIP are roughly about the same numbers. Thus, PPIP has a similar level of agreement with the three previous works, although they took a very different approach from ours.

**Examples of predicted PPIs in *Arabidopsis*.** From the predicted PPIs for the three plant genomes (http://kiharalab.org/PIPI_results/), here we discuss examples with three different levels of confidence. Supplementary Table 12–21 provide examples from *Arabidopsis*. All the listed PPIs in the tables were predicted consistently by the SVM\(_{log}\) and RF\(_{rand}\) and the probability score of RF\(_{rand}\) was over 0.95. The difference of the confidence levels is based on the availability of additional supporting data.

The predictions are separated into three classes for each genome according to the availability of other evidence that supports the predictions. Supplementary Table S12 lists predicted PPIs with two more supporting evidence: a very high score of over 900 in the STRING database\(^{86}\) and also satisfy at least one of the following three conditions: (a) the two proteins are known to locate in the same pathway in the KEGG database\(^{85}\), (b) the two proteins are co-mentioned in a literature. The predicted PPIs with correlated elution profile in PPI detection using mass spectrometry (MS)\(^{84,92}\) are also included in Supplementary Table S12. STRING collects several different types of evidence for the functional association of protein pairs and provides a score that ranges from 0 to 1000 with 1000 as the most confident score. In the works by Aryan et al.\(^{84,85}\), proteins in *Arabidopsis* were fractionated using size exclusion chromatography, and abundance profiles across the column fractions were quantified using label-free precursor ion (MS1) intensity. Proteins with correlated profiles and clustered together among all detected proteins were more likely to form complexes (see the original papers for more details). Supplementary Table S13 lists predicted PPIs with at least one piece of additional evidence, either a common KEGG pathway or literature that co-mention the two proteins. Protein pairs listed in the Supplementary Table S14 do not have additional evidence because the proteins were not much studied before but were predicted with the highest score (1.0) by RF\(_{rand}\).

The first half of Supplementary Table S12 lists predicted *Arabidopsis* PPIs with literature that describes evidence of their interaction. This list selected the most confident prediction in the *Arabidopsis*. Most of the interacting proteins are ribosomal proteins. Besides ribosomal proteins, several pairs of Sm-like proteins are predicted to interact, which are known as subunits of the heteroheptameric complexes and function in mRNA splicing and degradation\(^{86,87}\). Another predicted PPI is plastid division protein PDV2 (AT2G16070) and protein accumulation and replication of chloroplasts 6 ARC6 (AT5G42480), which has been shown to interact in the intermembrane...
space to coordinate the division machinery of chloroplast membrane. Our predicted PPIs also included some subunits of known complexes such as coatomer, RNA polymerase, DNA directed RNA polymerase, augmin, and adaptor complex 1 (AP-1).

The latter half of the table provides PPIs with similar MS elution profiles. For example, V-type proton ATPase subunit C (AT1G12840) and V-type proton ATPase subunit G2 (AT4G23710) are predicted to interact by RF and SVM. Additionally, they have highly correlated protein elution profile and are involved in the two common KEGG pathway including oxidative phosphorylation (ath00190), and phagosome (ath04145).

Supplementary Table S13 lists predicted PPIs in Arabidopsis with the next level of confidence, which has extra evidence but no information available in STRING or with a low STRING score (<0.4). An interesting example is asymmetric leaves 2 (AS2) (AT1G65620) and histone deacetylase 6 (HDA6) (AT5G63110). Although they have a very low STRING score a previous study showed that HDA6 functions with AS2 to regulate the leaf development and suggested that HDA6 may be the part of the AS1-AS2 repression complex to repress KNOX gene expression in Arabidopsis.

Another interesting example is protection of telomeres protein 1α (POT1a) (AT2G605210) and CST complex subunit TEN1 (AT1G56260). It is known that CTC1, STN1, and TEN1 consist of telomere complex and POT1a interplay with CST components to regulate the telomerase enzyme activity.

The last table for Arabidopsis, Supplementary Table S14, shows a list of PPIs with the highest RF probability score yet have no other known evidence. An interesting example in this table is pentatricopeptide repeat-containing protein (AT1G05600 and AT5G27270). They may function in RNA editing in chloroplast.

**Examples of predicted PPIs in corn.** Predictions for corn are shown in Supplementary Tables S15–S17, and Fig. 4A. PPIs shown in these tables are predicted both by the SVM loc and RF and with a high RF probability score of 0.9 or higher. As in the tables for Arabidopsis, Supplementary Table S15 lists predicted PPIs with two additional supporting pieces of evidence, and Supplementary Table S16 is for PPIs with a single existing piece of additional evidence, in this case having a common KEGG pathway. While Supplementary Table 18 includes the predicted PPIs with no existing evidence for interaction.

In Supplementary Table S15, 12 protein pairs out of 18 listed turned out to be NAD(P)H-quinone oxidoreductase subunits. This list also includes interaction between hydroxymethylglutaryl-CoA synthase (UniProt ID: B6U9M4) and acetyl-CoA acetyltransferase (UniProt ID: B4F9B2), both of which are involved in the four same pathways including terpenoid backbone biosynthesis (zma00900), valine leucine and isoleucine degradation (zma00280), butanoate metabolism (zma00650), and synthesis and degradation of ketone bodies (zma00070), and they have a very high database score in STRING.

Supplementary Table S16 shows predicted PPIs with the highest RF score (1.0) locating at least in one common KEGG pathway. For these PPIs, a STRING score was not available. As shown, most of the proteins are kinases in the same pathways, the plant hormone signal transduction (KEGG: zma04075) or the plant-pathogen interaction pathway (KEGG: zma04626).

Thus, it is reasonable to conclude that they are interacting. In the table, we also provided protein functional association score (FunSim) with IAS scores. As mentioned in the Methods, IAS directly indicates the likelihood that proteins with the GO annotations interact. Since the IAS scores listed in the table are very high relative to the background distribution (within top 1% for IAS and PAS for proteins in Arabidopsis protein pairs), these scores also support the PPI predictions.

The third list, Supplementary Table S17, includes PPIs that were predicted with a high confidence score (RF probability = 1) and high functional correlations (IAS > 200 and PAS > 20 and phylogenetic similarity > 0.9), but do not have other existing supporting information or protein annotations. When the 226 proteins involved in these PPIs were represented into a network by connecting protein pairs in the PPIs using Cytoscape, 224 of them (99.1%) are clustered into four subnetworks (Fig. 4A). Since they have no functional annotations, we used PFP to predict their GO terms and performed GO enrichment analysis using NaviGO as shown in Supplementary Table S18. It turned out that each of the subnetworks has enriched GO terms that are common in the proteins in the network: The largest subnetwork involved in 101 proteins were predicted by PFP to be involved in flavonoid glucuronidation, flavonoid biosynthetic process, cellular glucuronidation, and quercetin 3-O-glucosyltransferase activity.

In the second largest subnetwork, 57 out of 59 proteins were predicted to be involved in the RNA metabolic process and RNA secondary structure unwinding. All proteins in the third subnetwork were predicted to function in proteolysis and protein catabolic process while proteins in the last subnetwork were predicted to be involved in the regulation of the metabolic process, regulation in gene expression, and regulation of transcription DNA-templated. Thus, proteins in the predicted PPI networks have coherent biological functions.

**Examples of predicted PPIs in soybean.** The last plant genome we analyzed was soybean. Soybean has much less available functional information in databases comparing with Arabidopsis and corn. Supplementary Table S19 selected a list of predicted PPIs using the same standard as Supplementary Table S15 for corn, i.e. PPIs supported by two additional evidence. This list includes subunits from known complexes, including ATP synthase, chalcone synthase, cytochrome, and NADH dehydrogenase.

The next table, Supplementary Table S20 shows PPIs without conclusive information in STRING. However, two proteins in each pair are found in the same KEGG pathway, and most of them have a similar function, judging from the name in KEGG or UniProt annotation. The predicted PPIs in this list include protein interaction between glycosyltransferase and protein kinase involved in plant hormone signal transduction, plant-pathogen interaction, zeatin biosynthesis, and carotenoid biosynthesis signaling pathway.

Supplementary Table S21 lists high confident PPIs (RF probability = 1, IAS > 200, PAS > 20 and phylogenetic similarity > 0.9), which do not have other existing supporting evidence and functional annotation in UniProt. As we did for the predicted PPIs in corn, in Fig. 4B we constructed networks with 224 proteins that are involved in the PPIs in Table S21 and performed the functional enrichment analysis using predicted GO terms by PFP.
Figure 4. The networks constructed with predicted PPIs with the highest RF confidence scores (1.0) but do not have documented other supporting evidence. Connected PPIs are predicted by both SVM loc and RF rand. PPIs were further selected by high functional similarity scores: IAS-FunSim (>200), PAS-FunSim (>20), and the phylogenetic profile similarity (>0.9). (A) Predicted PPIs for corn. 224 out of 226 proteins in the PPIs qualified for the criteria are included in four subnetworks shown. The number of proteins in each network is 101, 59, 41, and 23 for the subnetwork 1 to 4, respectively. Supplementary Table S17 provides the subnetwork index of each predicted PPI. (B) Predicted PPIs for soybean. 215 out of 224 proteins in the PPIs qualified for these criteria are included in four subnetworks shown. The number of proteins in each network is 145, 41, 20, and 9 for the subnetwork 1 to 4, respectively. Supplementary Table S21 provides the subnetwork index of each predicted PPI. See Supplementary Table S18 for the results of the functional enrichment analysis of the subnetworks.

Discussion

A PPI network is fundamental for understanding an organism’s functional and structural units. For example, PPIs are very useful for predicting the function of individual proteins as well as pathways of protein groups. Although large-scale PPIs of several model organisms have been revealed by experimental methods and by computational methods, the works for plant PPIs were sparse. This work is intended to fill the gap for plant PPIs by providing PPI predictions with the method that was calibrated on known PPIs in Arabidopsis.
It is inevitable that a computational method often makes wrong PPI predictions. However, as discussed in the introduction, the situation would be similar in experimental methods, as it has been reported that independent experiments have substantial disagreements26–28. To reduce errors of a method, either computational or experimental, it would be useful to compare outputs from multiple methods. Having this idea in mind, we designed PPIP such that it combines two independent predictions, one using sequence-based features and the other with a combination of orthogonal features (Fig. 1). This architecture sacrificed the recall rate but in return achieved a very low false positive rate, which we consider as a higher priority since we provide all predicted PPIs for reference information for biologists. Also, in the genome-scale predictions for the three genomes, we provided additional evidence from other sources whenever available. In the analysis, we highlighted predicted PPIs with three levels of confidence. All the PPIs in these three levels were predicted not only with high scores but with additional supporting evidence. PPIs in the first (best) confidence have direct literature information or multiple supporting data, including a high score in STRING and co-existence in the same pathway. In the second level, PPIs have at least one evidence including the co-existence in the same pathway. PPIs of the third level confidence have are between proteins with functional coherence. While it is true that functional similarity between proteins does not necessarily indicate physical interaction between them, it is certainly highly related with each other as it is a common practice to verify experimentally detected PPIs by checking their functional similarity7,18,96 and functional similarity is an informative feature of proteins for predicting PPIs45,46,99–101. PPI predictions made for the three plants are made available on our lab website (http://kiharalab.org/PPIP_results/). We hope they are used as a reference and found informative.

Methods
Below we describe details of features and the machine learning methods used to predict PPIs.

Sequence-based prediction. We explain protein sequence features used in PPIP. It is the left branch of the flowchart in Fig. 1. To capture physicochemical properties of interacting proteins the following seven features are assigned to each amino acid of query protein sequences102–108: hydrophobicity, hydrophilicity, side-chain volumes, polarity, polarizability, solvent-accessible surface area, and net charge index (NCI) of side-chains. Then each query protein sequence is represented with auto-covariance (AC) using strings of the seven features as follows:

$$AC(lag, j) = \frac{\sum_{i=1}^{L-\text{lag}} P_i (l) \frac{1}{\sum_{i=1}^{2\text{lag}} P_i (l)} - \frac{1}{\sum_{i=1}^{2\text{lag}} P_i (l)} \times \left( \frac{P_{i+(\text{lag})} (l) - \frac{1}{\sum_{i=1}^{2\text{lag}} P_i (l)}}{L - \text{lag}} \right)$$

where lag is the distance between covariant residues to consider, which ranges from 1 to 30, $j$ is the $j$-th physicochemical feature, $i$ is the position in the sequence, $P_{ij}$ is the value of the physicochemical feature $j$ of amino acid position $i$, and $L$ is the length of the sequence. Thus, AC of a physicochemical feature with a certain lag length will be large if amino acid with a large (or small) property value appears periodically with an interval of lag. AC is computed for each protein sequence, and thus a query protein pair is represented as a 2 (sequences) * 7 (features) * 30 (lag intervals) = 420-dimensional vector. The vector representation was used as input of SVM for predicting if protein pairs are interacting or not interacting. We took this approach because it was reported to be successful in a previous paper87.

Gene expression features. On the other branch of PPIP (Fig. 1), we used three features, gene co-expression, functional similarity, and phylogenetic profile similarity in the framework of RF to predicted PPI of a query protein pair. We explain the three features in the following three subsections.

If two proteins are upregulated or down-regulated simultaneously under various conditions, it is highly likely that the two proteins are involved in the same pathway and have a higher chance that they physically interact with each other. Thus, co-expression patterns can provide indirect evidence for predicting PPIs. We obtained gene co-expression information derived from microarray experiments and RNA-seq experiments from the ATTED-II database (http://atted.jp/)100. It is a database of pre-calculated Pearson’s correlation coefficients (PCC) and the mutual rank (MR) of co-expressed genes. MR is defined as the geometric mean of the rank of the correlation gene A to gene B among proteins in the genome and the rank of gene B to gene A. The smaller MR is, the stronger the genes are co-expressed. Since gene expression data are provided in two sources, microarray and RNA-seq, four features (microarray MR, microarray PCC, RNA-seq MR, and RNA-seq pCC) were used to represent the co-expression profile of protein pairs.

Protein function features. The second feature used is protein functional similarity. Proteins with the same or similar biological functions are likely to physically interact because they may form permanent complexes or take part in the same pathway. Functional similarity of proteins was quantified by established similarity scores of Gene Ontology (GO) terms72. GO annotations were obtained from UniProt72 and TAIR (for Arabidopsis). We used three GO similarity/relevance scores, Interaction association score (IAS)45, Co-Occurrence Association Score (CAS), and PubMed Association Score (PAS)93,110. IAS, CAS, and PAS quantify how significantly a pair of GO terms appear in physically interacting proteins, annotations of individual genes, and PubMed abstracts. Thus, they evaluate co-occurrence of GO terms in biological contexts and shown to be effective in identifying proteins that physically interact45 or in the same pathways111.

Phylogenetic profile similarity. It has been observed that interacting proteins tend to coevolve129. We use the phylogenetic profile to exploit the evolutionary co-occurrence patterns of interacting proteins. The basic assumption of the phylogenetic profile method is that interacting proteins either co-present or co-absent across
organisms. The original phylogenetic profile is a binary pattern of presence or absence of homologs in a set of reference genomes, but we used a modified version of the profile that used BLAST bit score instead of binary values as follows:

\[
sim(i, j) = \frac{\sum_{k=1}^{n} R_{ik} \times R_{jk}}{\left(\sum_{k=1}^{n} R_{ik}^2 \times \left(\sum_{k=1}^{n} R_{jk}^2\right)^{1/2}\right)}
\]

where

\[
R_{ik} = \frac{B_{ik}}{B_{ij}}
\]

The similarity of protein \(i\) and \(j\) are defined in Eq. 2. \(k\) is the \(k\)-th reference genome, \(R_{ik}\) is the BLAST search bit score of homolog of protein \(i\) in the \(k\)-th genome divided by the BLAST search bit score of \(i\) in the query genome (Eq. 3). We used 100 reference genomes \((n = 100)\) (Supplementary Fig. S2). These genomes were selected in the following steps: we ran BLAST searches from all Arabidopsis protein sequences against the UniProt database using the default E-value cutoff of 10. Then, we constructed a phylogenetic tree for the genomes and manually selected the genomes from each branch of the tree so that the selected genomes are well distributed and represent the tree.

**Machine learning methods.** We used two machine learning methods in PPIP, SVM for making predictions from sequence features and RF for predicting from a combination of four other features (Fig. 1). For SVM, we used the software package libsvm 2.84. SVM uses a kernel function to transform input features and two hyper-parameters that need to be determined, a regularization parameter \(\gamma\), which defines how far each training data influences the model and \(C\), which controls the tradeoff of misclassification on training examples. For our kernel function, we used a radial kernel following previous works that predict PPI prediction from protein sequence features. The two hyper-parameter values, \(C\) and \(\gamma\), were determined to be \(\log C = 5\) and \(\log \gamma = -1\) for SVMRF and \(\log C = 1\) and \(\log \gamma = 1\) for SVMMDGI by performing nested cross-validation as shown in Supplementary Table 2.

In parallel to the sequence-based prediction with SVM, RF was used to make an independent prediction from three features, functional similarity, gene co-expression, and phylogenetic profile similarity. RF is an ensemble learning method, which combines predictions made by a number of decision trees by a majority vote. RF can also determine important variables that contributed most in classification by calculating two metrics, the mean decrease of accuracy (MDA) and the mean decrease of Gini importance (MDGI) (refer to Additional file 8: Note S1 for more details). MDA is the difference of the error rate of classification caused by permuting feature values with values of other data points in a dataset. MDGI tells how much less a particular feature is selected as a node in the random forest after permuting this feature. The larger MDA and MDGI of a certain feature are, the more important that feature is. Similarly to SVM, we performed nested cross-validation to determine three hyper-parameter values used in RF (Supplementary Table S3).

**Data Availability**

All data generated or analyzed during this study are included in this article and its supplementary information files. The genome-scale prediction results are available on our lab website http://kiharalab.org/PPIP_results/.

**References**

1. Habibi, M., Eslahchi, C. & Wong, L. Protein complex prediction based on k-connected subgraphs in protein interaction network. *BMC systems biology* 4, 129, https://doi.org/10.1186/1752-0509-4-129 (2010).

2. King, A. D., Przulj, N. & Jurisica, I. Protein complex prediction via cost-based clustering. *Bioinformatics* 20, 3013–3020, https://doi.org/10.1093/bioinformatics/btn351 (2004).

3. Hawkins, T. & Kihara, D. Function prediction of uncharacterized proteins. *Mol Biosyst* 4, 223–231 (2008).

4. Khan, I. K. & Kihara, D. Genome-scale prediction of moonlighting proteins using diverse protein association information. *Bioinformatics* 32, 2281–2288, https://doi.org/10.1093/bioinformatics/btw166 (2016).

5. Shinn, W. H., Christoffer, C. W. & Kihara, D. In silico structure-based approaches to discover protein-protein interaction-targeting drugs. *Methods* 131, 22–32, https://doi.org/10.1016/j.ymeth.2017.08.006 (2017).

6. King, N. P. et al. Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* 336, 1171–1174 (2012).

7. Sambrook, J. & David W. R. “Identification of associated proteins by coimmunoprecipitation.” Cold Spring Harbor Protocols 2006i, pdb-prot3898 (2006).

8. Keroworthy, A. K. Imaging protein-protein interactions using fluorescence resonance energy transfer microscopy. *Methods* 24, 289–296 (2001).

9. Nikolovska-Coleska, Z. Studying protein-protein interactions using surface plasmon resonance. *Protein-Protein Interactions: Methods and Applications*, 109–138 (2015).

10. Vinogradova, O. & Qin, J. In NMR of Proteins and Small Biomolecules 35–45 (Springer, 2011).

11. Zuidweg, E. R. Mapping protein—protein interactions in solution by NMR spectroscopy. *Biochemistry* 41, 1–7 (2002).

12. Kobe, B. et al. (Portland Press Limited, 2008).

13. Dudkina, N. V., Koulil, R., Bulterma, J. B. & Boekema, E. J. Imaging of organelles by electron microscopy reveals protein–protein interactions in mitochondria and chloroplasts. *FEBS letters* 584, 2510–2515 (2010).

14. Fields, S. & Stengel, R. The two-hybrid system: an assay for protein–protein interactions. *Trends in Genetics* 10, 286–292 (1994).

15. Walhout, A. J., Boulton, S. J. & Vidal, M. Yeast two-hybrid system and protein interaction mapping projects for yeast and worm. *Yeast* 17, 88–94 (2000).
17. Rual, J.-F. et al. Towards a proteome-scale map of the human protein–protein interaction network. Nature 437, 1173–1178 (2005).
18. Rajagopal, S. V. et al. The binary protein–protein interaction landscape of Escherichia coli. Nature biotechnology 32, 285–290 (2014).
19. Boeri Erba, E. & Petosa, C. The emerging role of native mass spectrometry in characterizing the structure and dynamics of macromolecular complexes. Protein Science 24, 1176–1192 (2015).
20. Dunham, W. H., Mullin, M. & Ginsgas, A. C. Affinity-purification coupled to mass spectrometry: Basic principles and strategies. Proteomics 12, 1576–1590 (2012).
21. Morris, J. H. et al. Affinity purification–mass spectrometry and network analysis to understand protein–protein interactions. Nature protocols 9, 2359–2354 (2014).
22. Guruharsha, K. et al. A protein complex network of Drosophila melanogaster. Cell 147, 690–703 (2011).
23. Rao, V. S., Srinivas, K., Sujini, G. & Kumar, G. Protein–protein interaction detection: methods and analysis. International journal of proteomics 2014 (2014).
24. Pihler, J. New methodologies for measuring protein interactions in vivo and in vitro. Current opinion in structural biology 15, 4–14 (2005).
25. Wette, N. et al. Investigation of stable and transient protein–protein interactions: past, present, and future. Proteomics 13, 538–557 (2013).
26. Huang, H. & Bader, J. S. Precision and recall estimates for two-hybrid screens. Bioinformatics 25, 372–378 (2009).
27. Serebriiskii, I. G. & Golemis, E. A. Two-Hybrid System and False Positives: Approaches to Detection and Elimination. Two-Hybrid Systems: Methods and Protocols, 123–134 (2001).
28. Ginsgros, A.-C., Gstaiger, M., Raught, B. & Aebersold, R. Analysis of protein complexes using mass spectrometry. Nature reviews Molecular cell biology 8, 645–654 (2007).
29. Chait, A. et al. The BioGRID interaction database: 2017 update. Nucleic Acids Res 45, D369–D379, https://doi.org/10.1093/nar/gkw1102 (2017).
30. Ding, Z. & Kihara, D. Computational Methods for Predicting Protein–Protein Interactions Using Various Protein Features. Current Protocols in Protein Science, e62 (2018).
31. Chen, X. W. & Liu, M. Prediction of protein–protein interactions using random decision forest framework. Bioinformatics 21, 4394–4400, https://doi.org/10.1093/bioinformatics/btt721 (2005).
32. Sprinzak, E. & Margalit, H. Correlated sequence-signatures as markers of protein–protein interaction. Journal of molecular biology 311, 681–692, https://doi.org/10.1016/j.jmb.2001.4920 (2001).
33. Pitre, S. et al. PIPE: a protein–protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. BMC bioinformatics 7, 365, https://doi.org/10.1186/1471-2105-7-365 (2006).
34. Shen, J. et al. Predicting protein–protein interactions based only on sequences information. Proceedings of the National Academy of Sciences of the United States of America 104, 4337–4341, https://doi.org/10.1073/pnas.0607879104 (2007).
35. Nanni, L. & Lumini, A. An ensemble of K-local hyperplanes for predicting protein–protein interactions. Bioinformatics 22, 1207–1210, https://doi.org/10.1093/bioinformatics/btl055 (2006).
36. Ding, Y., Tang, J. & Guo, F. Predicting protein–protein interactions via multivariate mutual information of protein sequences. BMC bioinformatics 17, 398, https://doi.org/10.1186/s12859-016-1253-9 (2016).
37. You, Z. H., Chan, K. C. & Hu, P. Predicting protein–protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. PLoS one 10, e0125811, https://doi.org/10.1371/journal.pone.0125811 (2015).
38. Guo, Y., Yu, L., Wen, Z. & Li, M. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic acids research 36, 3025–3030, https://doi.org/10.1093/nar/gkn139 (2008).
39. Wallhout, A. J. et al. Protein interaction mapping in C. elegans using proteins involved in vulval development. Science 287, 116 (2000).
40. Huang, T. W. et al. POINT: A database for the prediction of protein–protein interactions based on the orthologous interactome. Bioinformatics 20, 3273–3276, https://doi.org/10.1095/obioinformatics/bfh366 (2004).
41. Lee, S. A. et al. Ortholog-based protein–protein interaction prediction and its application to inter-species interactions. BMC bioinformatics 9(Suppl 12), S11, https://doi.org/10.1186/1471-2105-9-S12-S11 (2008).
42. De Bodt, S., Proost, S., Vandepoele, K., Rouze, P. & Van de Peer, Y. Predicting protein–protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-expression. BMC genomics 10, 288, https://doi.org/10.1186/1471-2164-10-288 (2009).
43. Gu, H., Zhu, P., Jiao, Y., Meng, Y. & Chen, M. PRIN: a predicted rice interactome network. BMC bioinformatics 12, 161, https://doi.org/10.1186/1471-2165-12-161 (2011).
44. Najafabadi, H. S. & Salavati, R. Sequence-based prediction of protein–protein interactions by means of codon usage. Genome biology 9, R87, https://doi.org/10.1186/gb-2008-9-5-r87 (2008).
45. Yermen, S., Khan, I. K., Wei, Q. & Kihara, D. IAS: Interaction Specific GO Term Associations for Predicting Protein–Protein Interaction. Networks. IEEE/ACM transactions on computational biology and bioinformatics/IEEE, ACM 15, 1247–1258, https://doi.org/10.1109/TCBB.2015.2476809 (2018).
46. Zhang, S. B. & Tang, Q. R. Protein–protein interaction inference based on semantic similarity of Gene Ontology terms. J Theor Biol 401, 30–37, https://doi.org/10.1016/j.jtbi.2016.04.020 (2016).
47. Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein engineering 14, 609 (2001).
48. Juan, D., Pazos, F. & Valencia, A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proceedings of the National Academy of Sciences of the United States of America 105, 934–939, https://doi.org/10.1073/pnas.0709671105 (2008).
49. Sato, T., Yamanishi, Y., Horimoto, K., Kanehisa, M. & Toh, H. Partial correlation coefficient between distance matrices as a new indicator of protein–protein interactions. Bioinformatics 22, 2488–2492, https://doi.org/10.1093/bioinformatics/btl219 (2006).
50. Szklarczyk, D. et al. The STRING database in 2017: quality-controlled protein interaction association networks, made broadly accessible. Nucleic Acids Res 45, D362–D368, https://doi.org/10.1093/nar/gkw937 (2017).
51. Soong, T. T., Wrzeszczynski, K. O. & Rost, B. Physical protein–protein interactions predicted from microarrays. Bioinformatics 24, 2608–2614, https://doi.org/10.1093/bioinformatics/btn498 (2008).
52. Wass, M. N., Puente, G., Pons, C., Pazos, F. & Valencia, A. Toward the prediction of protein interaction partners using physical docking. Molecular systems biology 7, 469, https://doi.org/10.1038/msb.2011.3 (2011).
53. Ohue, M., Matsuzaki, Y., Uchikoga, N., Ishida, T. & Akiyama, Y. MEGADOCK: an all-to-all protein–protein interaction prediction system using tertiary structure data. Protein and peptide letters 18, 766–774 (2014).
54. Tuncbag, N., Gursoy, A., Nussinov, R. & Keskin, O. Predicting protein–protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. Nature protocols 6, 1341–1354, https://doi.org/10.1038/nprot.2011.367 (2011).
55. Mirabello, C. & Wallner, B. InterPred: A pipeline to identify and model protein–protein interactions. Proteins 85, 1159–1170, https://doi.org/10.1002/prot.25280 (2017).
56. Zhang, Q. C. et al. Structure-based prediction of protein–protein interactions on a genome-wide scale. Nature 490, 556–560 (2012).
57. Garzón, J. I. et al. A computational interaction and functional annotation for the human proteome. Elife 5, e18715 (2016).
58. Betel, D. et al. Structure-taped predictions of novel protein interactions from sequence information. PLoS computational biology 3, 1783–1789, https://doi.org/10.1371/journal.pcbi.0030182 (2007).
59. Matthews, L. R. et al. Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or “interologs”. Genome Res. 11, 2120 (2001).
60. Kuchaiev, O., Rasajski, M., Higham, D. J. & Przulj, N. Geometric de-noising of protein–protein interaction networks. PLoS computational biology 5, e1000454, https://doi.org/10.1371/journal.pcbi.1000454 (2009).
61. Geisler-Lee, J. et al. A predicted interactome for Arabidopsis. Plant Physiol 145, 317–329, https://doi.org/10.1104/pp.107.103465 (2007).
62. Wang, F. et al. Prediction and characterization of protein–protein interaction networks in swine. Proteome science 10, 2, https://doi.org/10.1186/1477-5596-10-2 (2012).
63. Lee, S. A. et al. PROINeT: protein interactome with sub-network analysis and hub prioritization. BMC bioinformatics 10, 114, https://doi.org/10.1186/1471-2105-10-114 (2009).
64. Dukowski, J. & Tiuryn, J. Phylogeny-guided interaction mapping in seven eukaryotes. BMC bioinformatics 10, 393, https://doi.org/10.1186/1471-2105-10-393 (2009).
65. Hosur, R. et al. A computational framework for boosting confidence in high-throughput protein–protein interaction datasets. Genome biology 13, R76, https://doi.org/10.1186/gb-2012-13-8-r76 (2012).
66. Garcia-Hernandez, M. et al. TAIR: a resource for integrated Arabidopsis data. Funct Integr Genomics 2, 239–253, https://doi.org/10.1007/s10242-002-0077-z (2002).
67. Guo, Y., Yu, L., Wen, Z. & Li, M. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic acids research 36, 3025–3030 (2008).
68. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 96, 4285–4288 (1999).
69. Louppe, Gilles, et al. “Understanding variable importances in forests of randomized trees.” Advances in neural information processing systems. (2013).
70. Chang, C.-C. & Lin, C.-J. Training v-support vector regression: theory and algorithms. Neural computation 14, 1959–1977 (2002).
71. Pundir, S., Martin, M. J. & O’Donovan, C. UniProt Protein Knowledgebase. Methods Mol Biol 1558, 41–55, https://doi.org/10.1007/978-1-4939-6783-4_2 (2017).
72. Consortium, G. O. Gene Ontology Consortium: going forward. Nucleic acids research 31, D1049–1056, https://doi.org/10.1093/nar/gki179 (2015).
73. Khan, I. K., Qing, W. & Kihara, D. PFP/ESG: automated protein function prediction servers enhanced with gene ontology visualization tool. Bioinformatics 31, https://doi.org/10.1093/bioinformatics/btu646 (2015).
74. Hawkins, T., Chitale, M., Luban, S. & Kihara, D. PFP: automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. Proteins: Struct, Funct, Bioinf 74, https://doi.org/10.1002/prot.22172 (2009).
75. Hawkins, T. & Kihara, D. PPF:Automatic annotation of protein function by relative GO association in multiple functional contexts. The 13th Annual International Conference on Intelligent Systems for Molecular Biology, 117 (2005).
76. Alsoschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. Journal of molecular biology 215, 403–410 (1990).
77. Radivojac, P. et al. A large-scale evaluation of computational protein function prediction. Nature methods 10, 221–227, https://doi.org/10.1038/nmeth.2340 (2013).
78. Jiang, T. et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome biology 17, 184, https://doi.org/10.1186/s13059-016-1037-6 (2016).
79. Hawkins, T., Chitale, M. & Kihara, D. Functional enrichment analyses and construction of functional similarity networks with high confidence function prediction by PFP. BMC bioinformatics 11, 265, https://doi.org/10.1186/1471-2105-11-265 (2010).
80. Wei, Q., Khan, I. K., Ding, Z., Yermen, S. & Kihara, D. NaviGO: interactive tool for visualization and functional similarity and coherence analysis with gene ontology. BMC bioinformatics 18, 177, https://doi.org/10.1186/s12859-017-1600-5 (2017).
81. Barabasi, A.-L. & Oltvai, Z. N. Network biology: understanding the cells functional organization. Nature reviews. Genetics 5, 101 (2004).
82. Clauset, A., Shalizi, C. R. & Newman, M. E. Power-law distributions in empirical data. SIAM review 51, 661–703 (2009).
83. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic acids research 45, D353–D361, https://doi.org/10.1093/nar/gkw1092 (2017).
84. Aryal, U. K. et al. A proteomic strategy for global analysis of plant protein complexes. Plant Cell 26, 3867–3882, https://doi.org/10.1105/tpc.114.127563 (2014).
85. Aryal, U. K., McBride, Z., Chen, D., Xie, J. & Szymanski, D. B. Analysis of protein complexes in Arabidopsis leaves using size exclusion chromatography and label-free protein correlation profiling. Journal of Proteomics (2017).
86. Perea-Rosa, C., Hernández-Verdeja, T., López-Cobollo, R., del Mar Castellanos, M. & Salinas, J. L. SM proteins provide accurate splicing and decay of selected transcripts to ensure normal Arabidopsis development. The Plant Cell, 112, 103697 (2012).
87. Golisz, A., Sikorski, P. J., Krużska, K. & Kufel, J. Arabidopsis thaliana DSM proteins function in mRNA splicing and degradation. Nucleic acids research 41, 6232–6249 (2013).
88. Glynn, J. M., Froehlich, J. E. & Osteryoung, K. W. Arabidopsis ARC6 coordinates the division machineries of the inner and outer chloroplast membranes through interaction with PDV2 in the intermembrane space. The Plant Cell 20, 2460–2470 (2008).
89. Luo, M. et al. Histone deacetylase HDAC6 is functionally associated with AS1 in repression of KNOX genes in Arabidopsis. PLoS genetics 8, e1003114 (2012).
90. Renfrew, K. B., Song, X., Lee, J. R., Arora, A. & Shippen, D. E. POT1a and components of CST engage telomerase and regulate its activity in Arabidopsis. PLoS genetics 10, e1004738 (2014).
91. Kotera, E., Tasaoka, M. & Shikanai, T. A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. Nature 433, 326 (2003).
92. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27, 431–432 (2010).
93. Ding, Z., Wei, Q. & Kihara, D. In Data Mining for Systems Biology 113–130 (Springer, 2018).
94. Khan, I. K. et al. Prediction of protein group function by iterative classification on functional relevance network. Bioinformatics (2018).
95. Afrozuzzaman, M. et al. Large-scale identification of protein–protein interaction of Escherichia coli K-12. Genome research 16, 686–691 (2006).
96. Miller, J. P. et al. Large-scale identification of yeast integral membrane protein interactions. P Natl Acad Sci USA 102, 12123–12128 (2005).
97. Sato, S. et al. A large-scale protein–protein interaction analysis in Synecocystis sp. PCC6803. DNA research 14, 207–216 (2007).
98. Li, Z. et al. Large-scale identification of human protein function using topological features of interaction network. Scientific Reports 6, 37179 (2016).
99. Qi, Y., Bar-Joseph, Z. & Klein-Seetharaman, J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins: Structure, Function, and Bioinformatics 63, 490–500 (2006).
100. Zhang, J., Jia, K., Jia, J. & Qian, Y. An improved approach to infer protein-protein interaction based on a hierarchical vector space model. *BMC bioinformatics* **19**, 161 (2018).
101. Bandyopadhyay, S. & Mallick, K. A new feature vector based on gene ontology terms for protein-protein interaction prediction. *IEEE/ACM transactions on computational biology and bioinformatics* **14**, 762–770 (2017).
102. Tanford, C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *Journal of the American Chemical Society* **84**, 4240–4247 (1962).
103. Hopp, T. P. & Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences* **78**, 3824–3828 (1981).
104. Breiman, L. Random forests.
105. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974).
106. Charton, M. & Charton, B. I. The structural dependence of amino acid hydrophobicity parameters. *Journal of theoretical biology* **99**, 629–644 (1982).
107. Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. Hydrophobicity of amino acid residues in globular proteins. *Science* **229**, 834–838 (1985).
108. Zhou, P., Tian, F., Li, B., Wu, S. & Li, Z. Genetic algorithm-based virtual screening of combinative mode for peptide/protein. *Acta Chimica Sinica-Chinese Edition* **64**, 691 (2006).
109. Aoki, Y., Okamura, Y., Tadaka, S., Kinoshita, K. & Obayashi, T. ATTED-II in 2016: A Plant Coexpression Database Towards Lineage-Specific Coexpression. *Plant Cell Physiol* **57**, e5, https://doi.org/10.1093/pcp/pcv165 (2016).
110. Chitale, M., Khan, I. K. & Kihara, D. Missing gene identification using functional coherence scores. *Scientific reports* **6**, 31725 (2016).
111. Chitale, M., Khan, I. K. & Kihara, D. Quantification of protein group coherence and pathway assignement using functional association. *BMC bioinformatics* **12**, 373, https://doi.org/10.1186/1471-2105-12-373 (2011).
112. Tanford, C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins.
113. Grantham, R. Amino acid difference formula to help explain protein evolution.
114. Chang, C.-C. & Lin, C.-J. LIBSVM: a library for support vector machines.
115. An, J. Y. et al. Improving protein-protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model. *Proceeding of the Protein Society* **25**, 1825–1833, https://doi.org/10.1002/pr.2991 (2016).
116. Huang, P. & Ko, P. Linear discriminant analysis for gene selection in cancer classification with replications. *Proceeding of the World Congress in Computer Science, Computer Engineering and Applied Computing* (2010).
117. You, Y. Transmission of the information in the protein's molecule. *Journal of the American Chemical Society* **73**, 3632–3633 (2010).
118. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *Machine Learning* **40**, 105–131 (2000).
119. Aliferis, C. F., Statnikov, A. & Tsamardinos, I. Challenges in the analysis of mass-throughput data: a technical commentary from the statistical machine learning perspective. *Cancer Informatics* **7**, 117693510600200004 (2006).
120. Breiman, L. Random forests. *Machine learning* **45**, 5–32 (2001).

**Acknowledgements**

This work was partly supported by the National Institute of General Medical Sciences of the NIH (R01GM123055) and the National Science Foundation (DMS1614777, CMMI1825941). ZD was supported by the Purdue Research Foundation. The authors thank Lyman Monroe for proofreading the manuscript.

**Author Contributions**

D.K. conceived the study. Z.D. performed the experiments. Z.D. and D.K. analyzed the data. Z.D. drafted the manuscript. D.K. supervised and completed the writing. D.K. agrees to serve as the author responsible for contact and ensures communication.

**Additional Information**

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-45072-8.

**Competing Interests:** The authors declare no competing interests.

**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2019.