Briefings in Bioinformatics, 2017, 1–14

doi: 10.1093/bib/bbx123
Paper

Critical assessment and performance improvement of plant–pathogen protein–protein interaction prediction methods

Shiping Yang, Hong Li, Huaqin He, Yuan Zhou and Ziding Zhang

Corresponding authors: Yuan Zhou, State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China. E-mail: soontide6825@163.com; Ziding Zhang, State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China. E-mail: zidingzhang@cau.edu.cn

Abstract

The identification of plant–pathogen protein–protein interactions (PPIs) is an attractive and challenging research topic for deciphering the complex molecular mechanism of plant immunity and pathogen infection. Considering that the experimental identification of plant–pathogen PPIs is time-consuming and labor-intensive, computational methods are emerging as an important strategy to complement the experimental methods. In this work, we first evaluated the performance of traditional computational methods such as interolog, domain–domain interaction and domain–motif interaction in predicting known plant–pathogen PPIs. Owing to the low sensitivity of the traditional methods, we utilized Random Forest to build an interspecies PPI prediction model based on multiple sequence encodings and novel network attributes in the established plant PPI network. Critical assessment of the features demonstrated that the integration of sequence information and network attributes resulted in significant and robust performance improvement. Additionally, we also discussed the influence of Gene Ontology and gene expression information on the prediction performance. The Web server implementing the integrated prediction method, named InterSPPI, has been made freely available at http://systbio.cau.edu.cn/intersppi/index.php. InterSPPI could achieve a reasonably high accuracy with a precision of 73.8% and a recall of 76.6% in the independent test. To examine the applicability of InterSPPI, we also conducted cross-species and proteome-wide plant–pathogen PPI prediction tests. Taken together, we hope this work can provide a comprehensive understanding of the current status of plant–pathogen PPI predictions, and the proposed InterSPPI can become a useful tool to accelerate the exploration of plant–pathogen interactions.

Key words: plant–pathogen interaction; protein–protein interaction; machine learning; feature integration; prediction

Introduction

Plants face a battery of pathogens such as bacteria, oomycetes, fungi and viruses during their lifetime. It was reported that the loss in crop production caused by pathogen infections ranged from 20% to 40% and the direct economic loss was up to 40 billion dollars yearly in the United States alone [1, 2]. In China, one plant pathogenic fungus named *Rhizoctonia solani* could affect

Shiping Yang is a PhD student at State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University. His research interests include protein bioinformatics and machine learning.

Hong Li is a PhD student at State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University. Her research interests include protein bioinformatics and protein–protein interaction network.

Huaqin He is a Professor of Bioinformatics at the College of Life Sciences, Fujian Agriculture and Forestry University. His research interests are protein bioinformatics and genomics.

Yuan Zhou received his PhD in Bioinformatics at State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University in 2015 and now works at Department of Biomedical Informatics, Peking University. His research interests include machine learning, protein bioinformatics, transcriptome and epitranscriptome.

Ziding Zhang is a Professor at State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University. His research interests are protein bioinformatics and systems biology.

Submitted: 26 July 2017; **Received (in revised form):** 24 August 2017

Table 1. Existing databases and resources related to plant–pathogen PPIs

Name	Description	Method	PPIs[a]	URLs
PHI-base	It stores functional interactions between pathogens and plants, among which few are physical PPIs.	Experimentally verified	18	http://www.phi-base.org/index.jsp
HPIDB 2.0	It stores PPIs from existing interaction resources and manual curation of published literature. Only about 1% of all collected PPIs are between plant species and their pathogens.	Experimentally verified	531	http://www.agbase.msstate.edu/hpi/main.html
PPIN1	It contains PPIs between 36 *Psy*, 60 *Hpa* effectors and 165 *Arabidopsis* proteins.	Experimentally verified	342	http://signal.salk.edu/interactome/PPIN1.html
XooNET	It stores predicted PPIs between *Xanthomonas Oryza pv. oryzae* membrane proteins and rice proteins.	Interolog and DDI	3407	http://www.inetbio.org/xoonet/downloadnetwork.php
PPIRA	It contains predicted PPIs between *Ralstonia solanacearum* proteins and *Arabidopsis* proteins.	Interolog and DDI	3074	http://protein.cau.edu.cn/ppira/
PCPPI	It contains predicted PPIs between *Penicillium expansum* and seven crops.	Interolog and DDI	439 904	http://bdg.hfut.edu.cn/pcppi/index.html
UVPID	It stores predicted PPIs between *Ustilaginoidea virens* and rice.	Interolog and DDI	3597	http://sunlab.cau.edu.cn/uvpid/
Sahu *et al.*	It provides a proteome-wide prediction of PPIs between *Psy* and *Arabidopsis*.	Interolog and DDI	0.79 million	NA
Kshirsagar *et al.*	It uses ML to predict PPIs between *Salmonella* and *Arabidopsis*. Features such as protein sequence information (e.g. n-mer or n-gram), GO similarity and gene expression patterns were used as input.	Transfer learning	NA	http://www.cs.cmu.edu/~mkshirsa/

[a] The PPIs are either experimentally identified or computational predicted by interolog, domain-domain interaction (DDI) and machine-learning (ML)-based methods. Only physical PPIs between plant and pathogen proteins are counted here.

approximately 15–20 million hectares of rice growing area, causing 6 million tons of rice grains loss per year [3]. Therefore, in-depth understanding of plant–pathogen interaction is critical for the breeding of disease-resistant crops and agricultural production improvement.

The plant–pathogen interaction is a two-way biological communication process. On the one hand, plants attempt to recognize the molecules secreted by pathogens to avoid being infected, but on the other hand, pathogens manipulate plants as much as possible to make the host environment more beneficial to them [4]. Such a complicated relationship is often vividly regarded as the 'arms race' between plants and pathogens. So far, two levels of plant immune responses to pathogens have been well established. Briefly, pattern recognition receptors located on the plant cell surface first recognize pathogen-associated molecular patterns (PAMPs) from pathogens and activate the first tier of plant immunity called PAMP-triggered immunity (PTI). To sabotage the PTI response, pathogens secrete virulence molecules called effectors into plant cells. In response, plants use intracellular resistance proteins (R-proteins) to specifically recognize effectors and trigger the second tier of immune response named effector-triggered immunity (ETI) [5, 6]. The interaction partners of effectors in plants are defined as targets. The effector recognition process includes direct and indirect recognition, which means that targets could be either R-proteins or other accessory proteins. Therefore, the ETI process particularly depends on protein–protein interactions (PPIs) between pathogen effector proteins and their host targets.

Currently, experimental determination methods, such as yeast two-hybrid [7] and tandem affinity purification-mass spectroscopy [8], have been used to identify host–pathogen PPIs. In the meantime, a series of experimentally validated host–pathogen PPI databases have been constructed, including VirHostNet [9], HPIDB [10], PHISTO [11], PATRIC [12], VirusMentha [13] and HIV-1 Human Interaction Database [14]. Most of these databases focus on the PPIs between human and pathogens (especially viruses). By contrast, the plant–pathogen PPI data are quite limited in the existing host–pathogen PPI databases, and there is no plant-specific host–pathogen PPI database. To have a global view of current plant–pathogen PPIs, we survey and summarize existing plant–pathogen PPI resources [10, 15–23] in Table 1. Among these data resources, HPIDB is probably the most comprehensive database containing plant–pathogen PPIs. It collects 569 plant–pathogen PPIs, most of which are related to the model plant *Arabidopsis thaliana* (*Arabidopsis*) and the corresponding pathogens. Owing to the fact that experimental methods are still time-consuming and labor-intensive, the number of plant–pathogen PPIs is still limited, which is still not sufficient to keep pace with the rapid development of functional genomics studies in the field of plant pathology. To complement experimental methods, there is an urgent need to develop computational methods to accelerate the identification of new plant–pathogen PPIs.

Traditional PPI prediction methods such as interolog mapping [24], domain-based inference [25], gene fusion [26], phylogenetic similarity [27], gene co-expression [28] and structure-based method [29, 30] are initially designed to predict intra-species PPIs. With the advance of inter-species PPI studies, some of these methods, such as interolog and domain-based inference, have also been used to predict inter-species PPIs (e.g. human–pathogen PPIs [31–33] and plant–pathogen PPIs [18, 20]). In recent years, machine learning (ML) has been widely applied to solve diverse bioinformatics classification tasks including PPI prediction [34]. A series of ML algorithms, such as Naïve Bayes (NB) [35], Support Vector Machine (SVM) [36], Random Forest (RF) [37] and multitask learning [38], have been used to predict human–hepatitis C virus, human–papillomaviruses, human–*Plasmodium falciparum* PPIs. In contrast, there are few reports about ML-based plant–pathogen PPI prediction methods. *Arabidopsis* as an important model plant has been widely used to study

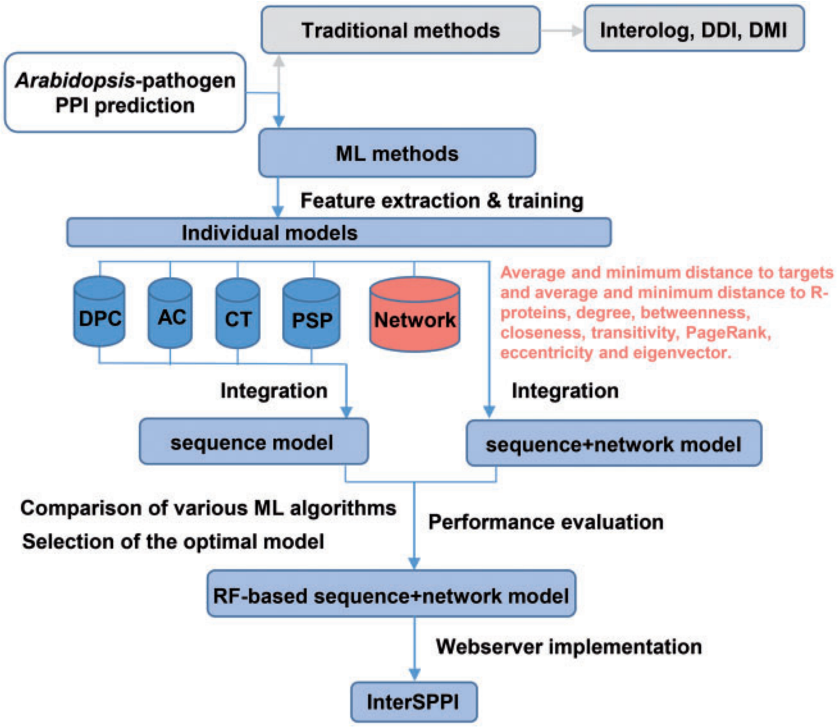
Traditional methods
Interolog, DDI, DMI
Arabidopsis-pathogen PPI prediction
ML methods
Feature extraction & training
Individual models
DPC
AC
CT
PSP
Network
Average and minimum distance to targets and average and minimum distance to R-proteins, degree, betweenness, closeness, transitivity, PageRank, eccentricity and eigenvector.
Integration
Integration
sequence model
sequence+network model
Comparison of various ML algorithms
Selection of the optimal model
Performance evaluation
RF-based sequence+network model
Webserver implementation
InterSPPI

Figure 1. The workflow of this work. The assessed approaches are mainly divided into two categories. One is traditional methods including interolog, DDI and DMI. The other is ML-based methods. To achieve a good predictive performance, we designed and compared sequence-based and network attributes-based encodings. Then, we used various ML algorithms to train prediction models. After comparing different predictors, the RF-based classification model that utilizes sequence and network attributes was finally chosen for webserver implementation because of its superior prediction performance.

plant–pathogen PPIs. Therefore, the studies of *Arabidopsis*–pathogen PPIs will largely promote the understanding of plant pathology.

To develop a ML-based PPI predictor, designing appropriate encoding schemes of the interacting protein pair is a prerequisite, which transforms an interacting protein pair into a feature vector that is further used as the input of ML algorithms. Previous studies had used a series of sequence-based encodings such as di-peptide composition (DPC) [39], auto covariance (AC) [40] and conjoint triad (CT) [41] to extract features from a protein pair. However, for inter-species PPI prediction, the prediction task is much more challenging owing to the need of describing the relationship between two species. To solve this problem, the computational framework by integrating heterogeneous biological information was proposed to predict human–pathogen PPIs [42]. The result showed that the integrated features outperformed pure sequence-based encodings.

Rather than predicting all possible plant–pathogen PPIs, here we focused on the prediction of PPIs between pathogen effectors and their host targets, which is based on the following reasons. First, the effector–target PPIs constitute the vast majority of the inter-species PPIs between plants and pathogens. Second, hundreds of PPIs between pathogen effectors and *Arabidopsis* targets have been identified through high-throughput interactomics studies [16, 43], which provides essential data to assess/develop plant–pathogen PPI predictors. Third, a plethora of well-performed effector prediction tools have been developed [44, 45], which provides a solid basis for the prediction of PPIs between effectors and their host targets. Last but not the least, previous studies conducted systematic analyses about the network topology of effector targets in the host PPI network. For instance, it has been established that effectors tend to attack hubs in the host PPI network [16], and we have recently revealed that the known targets are closer to each other [46]. These findings also provide important clues to design some effective encodings to predict effector–target interactions.

In this work, we conducted a comprehensive assessment of current plant–pathogen PPI prediction methods and proposed an improved ML-based predictor. The pipeline is illustrated in Figure 1. First, we evaluated the interolog and domain-based inference methods on all experimentally validated *Arabidopsis*–pathogen PPIs. It turned out that the traditional methods failed to effectively detect PPIs. Then, we systematically benchmarked different encoding schemes such as sequence- and network-based features in predicting *Arabidopsis*–pathogen PPIs. To develop a new predictor with improved performance, we used RF to train the classification models based on integrative feature encoding design, and the independent test showed that the sequence + network encoding outperformed sequence-based encoding alone. Finally, we implemented the RF-based method on a webserver termed as InterSPPI, and we tested it in cross-species prediction and proteome-wide plant–pathogen PPI identification. We anticipated that the current work could also provide inspiration to design more powerful inter-species PPI prediction tools.

Materials and methods

Data collection and data set construction

In general, experimentally validated PPIs are referred as positive samples, which were 459 *Arabidopsis*–pathogen PPIs collected from two recently published literature [16, 43]. These experimentally verified PPIs cover three representative pathogens, including *Pseudomonas syringae* (*Psy*), *Hpaloperonospora arabidopsis* (*Hpa*) and *Golovinomyces orontii* (*Gor*). *Pseudomonas syringae* is a

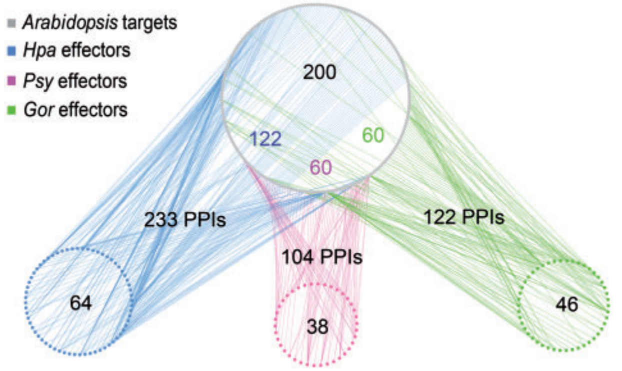

Figure 2. Summary of known inter-species PPI data between effectors of three pathogens and their host targets in *Arabidopsis*. In total, there are 459 PPIs between *Arabidopsis* targets and effectors of three pathogens, including 233 *Ara-Hpa*, 104 *Ara-Psy* and 122 *Ara-Gor* PPIs. The total number of the three pathogens' host targets is 200. Proteins from *Arabidopsis* and three pathogens are represented by four circles with different colors.

bacterial pathogen, and the number of *Arabidopsis–Psy* (*Ara-Psy*) PPIs is 104, which involve 60 *Arabidopsis* proteins and 38 *Psy* effectors. *Hpaloperonospora arabidopsis* belongs to oomycetes, and the corresponding number of *Arabidopsis–Hpa* (*Ara-Hpa*) PPIs is 233, involving 122 *Arabidopsis* proteins and 64 *Hpa* effectors. *Golovinomyces orontii* is a fungal pathogen, and the corresponding *Arabidopsis-Gor* (*Ara-Gor*) PPI number is 122, involving 60 *Arabidopsis* proteins and 46 *Gor* effectors. Details about the interaction data set are also illustrated in Figure 2. Considering the difference between pathogens, three pathogen-specific models and a general model covering all pathogens were built simultaneously.

To compile negative samples, we first collected *Arabidopsis* PPIs from public PPI databases including TAIR (Version of 2016.03.02) [47], BioGRID (Version of 2016.03.02) [48] and IntAct (Version of 2016.03.02) [49]. After removing redundant PPIs, we obtained 28 110 PPIs containing 7437 *Arabidopsis* proteins. These PPIs constituted the primary *Arabidopsis* intra-species PPI network (AraPPI). Because the prerequisite of network-based encoding is the presence of host target proteins in AraPPI, we only selected the proteins in AraPPI when building negative PPIs. Although several negative sampling schemes have been proposed previously, there is still no well-established 'gold standard' for non-interactions. The most widely used method is to randomly select PPIs from the set of all possible protein pairs except those already reported to interact. Regarding inter-species PPIs, the whole negative set is the combination of protein pairs between the host proteins in AraPPI and pathogen effectors. The possible combinations are $38 \times 7437 - 104 = 282\,502$, $64 \times 7437 - 223 = 475\,745$, $46 \times 7437 - 122 = 341\,980$ and $148 \times 7437 - 459 = 1\,100\,217$ for *Ara-Psy*, *Ara-Hpa*, *Ara-Gor* and *Ara-all_pathogens*, respectively.

After we obtained the initial positive and negative samples, the final step is to construct appropriate data sets for training a prediction model and assessing the performance. To better describe the data set construction, we took *Ara-Psy* as an example, and the generation of other data sets is similar to *Ara-Psy*. First, we randomly selected approximately one-fifth PPIs (i.e. 21) from the positive samples as the positive samples of the independent test set, and the remaining four-fifth PPIs (i.e. 83) were used as the positive samples of the training set. Second, we randomly selected negative samples from the whole negative sets to keep a 1:10 positive-to-negative ratio in both sets. To reduce the bias of negative samples, the negative sampling in the training set was repeated 10 times. The details are also illustrated in Supplementary Figure S1. Note that all samples in the independent test were not used during the training.

The interolog and domain-based methods for PPI prediction

Based on the compiled inter-species PPI data, we assessed the performance of three conventional computational methods [i.e. interolog, domain–domain interaction (DDI), and domain–motif interaction (DMI)] in predicting *Arabidopsis*–pathogen PPIs. The interolog method is based on the conservation of interacting protein pairs across different species. Briefly, if protein *A* interacts with protein *B* in one organism, the corresponding homologs in another organism (protein *A′* and protein *B′*) should also interact. Here protein pair *A* and *B* are regarded as the template to infer the predicted interaction pair *A′* and *B′*. To implement the interolog method, we first compiled a comprehensive template library by collecting experimentally verified PPIs from public interaction databases, including BioGRID (Version of 2016.03.02), IntAct (Version of 2016.03.02), DIP (Version of 2016.03.02) and HPIDB2.0. Moreover, PPIs supported by genetic interactions were removed to ensure the retaining PPIs are all physical interactions. As a result, we obtained 765 774 PPIs. We used BLAST [50] to identify the homologs between the query protein and all proteins in the template library, and the cutoffs of BLAST were set as follows. In a query PPI, one protein is from a pathogen; the other is from *Arabidopsis*. For both pathogen and *Arabidopsis* proteins, the E-value thresholds were defined as 0.01. Besides, both of the sequence identity and coverage cutoffs were set as 40% for each *Arabidopsis* protein. Considering it is relatively difficult for a pathogen protein to find homologs in the library, both of the sequence identity and alignment coverage cutoffs were set as 30% for each pathogen protein.

The central idea of DDI is that a protein pair should interact if they contain an interacting domain pair. Similarly, the DMI method is based on the observation that many PPIs are mediated by the interactions between domains and short linear motifs. For example, it has been known that DMI is a frequent interaction mode for viruses to attack their hosts [51]. To implement the DDI and DMI methods, the known DDI and DMI information was obtained from the 3did database [52]. The domain annotation was assigned through the PfamScan tool (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/) against the Pfam database (Pfam 30.0) [53] with the default parameter setting, and the motif scanning was conducted through the regular expression search based on the specific patterns curated by 3did.

ML algorithms

As an ensemble learning algorithm, RF creates a series of decision trees from randomly sampled subspaces of the input feature vectors. In this work, we used RF to train the classification models. Besides, we also compared RF with other popular ML algorithms, such as SVM, NB and Adaptive Boosting (AdaBoost) [54], K-Nearest Neighbors (KNN) and Logistic regression (LR). We used scikit-learn [55], a Python-based ML library, to implement these algorithms. For RF, the number of trees in the forest was set as 1000, and the number of features when seeking for the best split was set as the square root of the total number of features. Other parameters were set as default. SVM performs the classification by mapping a low-dimensional space to a highdimensional space through the kernel trick. Here, the radial basis function was chosen as the kernel, and parameters C and γ were optimized through grid search, where the ranges of C and γ were set as $[2^{-5}, 2^{11}]$ and $[2^{-13}, 2^{3}]$, respectively. AdaBoost is an algorithm for constructing a strong classifier from a series of weak classifiers and focuses more on the harder-to-classify cases in the subsequent classifiers. For the AdaBoost algorithm, the maximum number of trees at which boosting is terminated was also set as 1000. The NB classifier is based on Bayes theorem with the independence assumption among features. Here, we used GaussianNB which allows training on non-integer feature values. KNN is non-parametric and neighbors-based classification method. The number of neighbors was set as 5. LR is a special type of regression that estimates the probability of a binary outcome by analyzing the relationship between one or more independent variables. The linear model was adopted and the L2-norm penalty function was assigned.

Performance evaluation

We used the 10-fold cross-validation test to compare the performance of different models on the training data sets. As we randomly selected negative samples for 10 times, the final result is the average performance of the 10 replicates. To make a more stringent comparison, the independent test was also conducted. In this work, four parameters such as Precision, Recall (i.e. Sensitivity), Specificity and Matthew correlation coefficient (MCC) were used to evaluate the prediction performance. These measures are defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

where TP, TN, FP and FN represent the numbers of true positives, true negatives, false positives and false negatives, respectively. To provide a more comprehensive assessment of the models, the Precision–Recall (PR) curve, which is suitable for the cases when the positive and negative samples are not balanced [56], was used. The area under the PR curve is called auPRC. The closer the auPRC value is to 1, the better the performance of a prediction method is. All the PR curves were prepared using the pROC [57] package in R.

Sequence-based encoding schemes

One of the major challenges to build an inter-species PPI predictor is how to present a protein pair through a fixed-dimension feature vector. The vector usually contains features inferred from the primary protein sequence, although the biological meaning of most sequence-based features is not straightforward. Here, we mainly assessed four sequence-based encoding schemes, including DPC, AC, CT and predicted structural properties (PSP). The brief introduction of each encoding scheme is provided as follows (see also Supplementary Table S1).

DPC: DPC represents the percentage of two consecutive amino acids in the protein sequence [39], which can be calculated by:

$$S_{DPC}(A_iA_j) = \frac{N_{A_iA_j}}{L-1}, i, j \in (1, 2, \ldots, 20)$$

Where A_i and A_j stand for two of the 20 amino acids,$N_{A_iA_j}$is the total number of the di-peptide A_iA_j in the sequence. L is the sequence length, and $S_{DPC}(A_iA_j)$ is the percentage of A_iA_j in the sequence. For an individual protein, the feature dimension of DPC is $20 \times 20 = 400$. But for a protein pair, the final feature vector is characterized by concatenating the feature vectors of two proteins. Therefore, an 800-dimensional vector is constructed to represent each protein pair for DPC. Similar strategy is used to describe a PPI pair for other sequence-based encoding schemes.

CT: The CT encoding considered the properties of three consecutive amino acids in the sequence [41]. The 20 amino acids are clustered into seven groups based on the dipoles and volumes of the residue side chains. The equation to infer the CT encoding is defined as:

$$S_{CT}(G_iG_jG_k) = \frac{N_{GiGjGk}}{L-2}, i, j, k \in (1, 2, \ldots, 7)$$

Where G_i, G_j and G_k stand for three of the seven residue groups, N_{GiGjGk} is the total number of the CT $G_iG_jG_k$ in the sequence. $S_{CT}(G_iG_jG_k)$ is the final composition of $G_iG_jG_k$. The dimension of CT is $7 \times 7 \times 7 \times 2 = 686$.

AC: The AC encoding considers the neighboring effects through describing the interaction effects of residues with a certain distance [40]. Seven standardized physicochemical properties of amino acids were used to represent the interaction modes. They are hydrophobicity, hydrophilicity, polarity, polarizability, side chain volumes of amino acids, solvent-accessible surface area and net charge index of residue side chains. The final equation to calculate the score is as follows:

$$S_{AC}(lag, j) = \frac{1}{L - lag}\sum_{i=1}^{L-lag}\left(R_{i,j} - \frac{1}{L}\sum_{k=1}^{L}R_{k,j}\right) \times \left(R_{(i+lag),j} - \frac{1}{L}\sum_{k=1}^{L}R_{k,j}\right), j \in (1, 2, \ldots, 7)$$

Where i, k denotes the ith,kth residue in the sequence, and j represents one of the seven properties, $R_{i,j}$ and $R_{k,j}$ stands for the corresponding jth physicochemical property for the ith and kth residue, respectively. Here, lag is the sequence distance between the ith residue and its neighbors, which ranges from 1 to 30 in this work. Finally, the feature vector AC consists of $30 \times 7 \times 2 = 420$ values.

PSP: Previous studies have discovered that protein secondary structure composition [58] and protein disorder information [59, 60] have an impact on PPIs. To assess the structure-based encoding, we constructed a feature vector called PSP. Briefly, we considered three regions in the protein sequence: N-terminal (one-third of the full length of sequence at N-terminal), C-terminal (one-third of the full length of sequence at C-terminal) and the full sequence. For each region, we calculated the fraction of three different secondary structure elements (α-helix, β-strand and coil), and the percentage of disordered residues. The secondary structure and disorder content were predicted by PSSpred [61] and IUPred [62], respectively. As the structural features were predicted from protein sequences, PSP was also essentially sequence-based.

Integration of different prediction models

We first built a single prediction model for each encoding type (i.e. DPC, CT, AC or PSP). Then, these predictors were integrated into a comprehensive sequence-based prediction model. The model integration is based on Sun *et al.*'s work [63], which is defined as follows:

$$W_j = e^{(-k/auPRC)}, k \in (1, 2 \ldots, 30)$$

$$\hat{S} = 1 - \sum_{j=1}^{N} (1 - W_j \times S_j)$$

Where k is a constant ranging from 1 to 30, and the optimal value is chosen when the auPRC of the integrated model reaches the maximum value. W_j is the weight of the jth encoding type base on 10-fold cross-validation. N is the number of individual models, S_j denotes the prediction score of the jth individual model and $\hat{S}$ is the integrated prediction score.

Network-based feature vector

To develop a more accurate predictor, a series of network attributes of host proteins were also used as feature vector. These network attributes include degree, betweenness, closeness, transitivity, PageRank, eccentricity and eigenvector. Moreover, six novel plant–pathogen interaction-specific features, including the minimum and average network distances to known targets, the minimum and average network distances to the experimentally verified *Arabidopsis* R-proteins and the minimum and average network distances to the predicted *Arabidopsis* R-proteins were also incorporated into the network-based feature vector. The *Arabidopsis* R-protein information was retrieved from the PRGdb database, which contains experimentally verified and predicted *Arabidopsis* R-proteins [64]. The definition of known targets is the set of current *Arabidopsis* proteins involving in experimentally identified PPIs with a pathogen protein. Apparently, if the query protein itself is a known target, the minimum distance to known targets will be zero and thus yield biased results to some extent. Therefore, when calculating the distances, we excluded the query protein from the known target set in such case. All network attributes were calculated through igraph [65]. As a result, the network-based feature vector with a dimension of 13 was obtained (see also Supplementary Table S1). Note that only *Arabidopsis* proteins have the network-based feature vector. We trained a prediction model based on the network attributes, and integrated it with sequence encoding-based prediction models by adopting the aforementioned model integration strategy.

Other popular encoding schemes

Gene Ontology semantic similarity

Gene Ontology (GO) is a comprehensive resource to unify the functional description of gene and gene products across species [66]. The GO annotation includes three categories: molecular function (MF), cellular compartment (CC) and biological process (BP). The GO semantic similarity between two proteins has been proposed as an important feature in the prediction of host–pathogen PPIs [42]. Here, supposing that pathogen protein *A* can interact with host protein *B* and protein *B* has three endogenous interaction partners, we could first calculate the MF, BP and CC similarities between *A* and *B*. The GO semantic similarities between *A* and the three endogenous interaction partners of *B* were also computed. The maximal similarity score among *A* and three endogenous interaction partners of *B* was considered as the final similarity between *A* and endogenous partners of *B*. Thus, the GO semantic similarity-based encoding scheme contains a six-dimensional feature vector, including three effector–target GO similarities and three effector–partner GO similarities. We took *Ara-Psy* as an example. First, we conducted GO annotation for *Psy* and *Arabidopsis* proteins. The GO terms of *Arabidopsis* proteins are available at ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/, and the GO information for *Psy* is obtained through Blast2GO [67]. Then, we used GOSemSim [68] to calculate the corresponding GO similarity values. Finally, the GO semantic similarity-based feature vector was constructed with a dimension of 6. We built an RF-based predictor using the GO semantic similarity alone, and then integrated it with the sequence + network predictor to assess the effectiveness of the GO semantic similarity.

Gene expression pattern

Owing to the fact that interacting proteins tend to be co-expressed, expression correlation of two proteins is an indicator of interaction [22]. However, for inter-species PPI prediction, obtaining expression data that simultaneously detect the gene expression values of the host and the pathogen genes is quite difficult. An alternative way is to calculate the fold change value of the host gene because those differentially expressed genes are more likely to be involved in inter-species interaction. Inspired by this idea, we supposed that the fold changes of different time point comparing the pathogen infection condition and mock control condition could more comprehensively reflect the possibility that a host gene is involved in the inter-species interaction. To further verify our hypothesis, we conducted the test on the microarray data GSE56094, which is a set of gene expression data describing the dynamic transcriptional changes of *Arabidopsis* leaves during the infection by *Psy*. The expression data used in this work contain 13 time points with four replicates for each time point. First, we calculated the fold changes of *Arabidopsis* gene expression values under two conditions (infection and mock treatment) at each time point. Then, we merged these fold changes into a vector to represent a gene expression pattern. Similar to the encoding scheme of network attributes, only *Arabidopsis* proteins have the gene expression information-based feature vector. We built the gene expression information-based predictor, and investigated the combinational effect when the predictor was integrated with the sequence + network predictor.

Results and discussion

Traditional PPI prediction methods failed to identify *Arabidopsis*–pathogen PPIs

We used 459 experimentally validated *Arabidopsis*–pathogen PPIs and 4590 non-interaction protein pairs to assess the performance of traditional methods. The assessment is mainly based on two measures (i.e. Sensitivity and Precision). Among traditional PPI prediction approaches, interolog is one of the most widely used methods for intra-species and inter-species PPI prediction. However, only one PPI was successfully inferred among 459 experimentally validated *Arabidopsis*–pathogen PPIs by the interolog searching (Sensitivity = 0.2%, Precision = 50%; Table 2). There are two main reasons for such a low sensitivity. First, the size of the current PPI template library is still not sufficient. Many pathogen

Table 2. The performance of traditional methods in detecting Arabidopsis–pathogen PPIs

Traditional methods[a]	Reference databases	Sensitivity	Precision
Interolog search	Biogrid, IntAct, DIP and HPIDB	1/459 = 0.2%	1/2 = 50.0%
Domain–domain interaction	3did, Pfam	1/459 = 0.2%	1/5 = 20.0%
Domain–motif interaction	3did, Pfam	10/459 = 2.2%	10/161 = 6.2%
PSOPIA	Human PPIs	2/459 = 0.4%	2/45 = 4.4%
DXECPPI	Human PPIs	67/459 = 14.6%	67/638 = 10.5%

[a] PSOPIA and DXECPPI are ML-based predictors trained on human PPIs. The default prediction threshold values of PSOPIA and DXECPPI defined by the corresponding servers are used to decide whether submitted sequence pairs interact or not.

effector proteins failed to identify any homolog in the PPI template database. Although 95.5% (191 of 200) *Arabidopsis* proteins have homologous sequences, only 49.3% (71 of 144) pathogen proteins have homologous sequences. Second, the current PPI template library is mainly from intra-species PPIs (717 444/765 744 = 93.7%), resulting in handful straightforward inter-species PPI templates. Indeed, the strategy of inferring inter-species PPIs from intra-species PPIs is still disputable, although it has been applied to infer human–virus PPIs [51].

To verify the feasibility of domain-based method in our *Arabidopsis*–pathogen system, we conducted the domain annotation of every protein sequence against the Pfam database and then matched the paired domain information to the 3did database. Similarly, only one out of the 459 PPIs was successfully inferred through DDI. This result indicated that most of the pathogen effectors did not accommodate a domain involved in known DDIs. Therefore, the DDI method failed to predict *Arabidopsis*–pathogen PPIs in most cases. Comparatively, DMI achieved a much higher sensitivity. As a result, 10 of 459 known *Arabidopsis*–pathogen PPIs were identified through DMI. This result is consistent with previous observation that the inter-species interactions are likely mediated by DMI [69]. Generally, the length of a motif is much shorter than a domain, and a protein is easier to be annotated with multiple motifs. Thus, DMI is prone to yield false positives, resulting in a low precision (Precision = 6.2%). To solve this issue, a reliable DMI inference should adopt more stringent standards to identify motifs.

Moreover, we also notice that many ML-based PPI predictors trained on general PPIs (i.e. intra-species PPIs) have been developed. It is also interesting to evaluate these traditional ML-based predictors' performance on our *Arabidopsis*–pathogen system. To do so, we submitted our data set to two online predictors named PSOPIA (http://mizuguchilab.org/PSOPIA/) [70] and DXECPPI (http://ailab.ahu.edu.cn:8087/DXECPPI/) [71]. PSOPIA was trained on human PPIs by using averaged one-dependence estimator (a variant of the NB classifier) with features derived from known homologous PPIs, while DXECPPI was trained on human PPIs by using RF with the ensemble sequence encoding. The default prediction threshold values defined by the corresponding servers were adopted to decide whether submitted sequence pairs interact or not. The testing results indicated that both PSOPIA (Sensitivity = 0.4%, Precision = 4.4%; Table 2) and DXECPPI (Sensitivity = 14.6%, Precision = 10.5%; Table 2) could not obtain satisfying performance on the prediction of *Arabidopsis*–pathogen PPIs, further suggesting the large difference between plant–pathogen PPIs and general PPIs. Therefore, the existing ML models based on general PPIs are indeed not suitable for predicting inter-species PPIs between *Arabidopsis* and pathogens. Collectively, the current assessment experiments clearly showed that the traditional methods failed to effectively identify inter-species PPIs between *Arabidopsis* and pathogens, and the development of new predictors is therefore urgently required.

Performance assessment of RF models using sequence-based encodings

To evaluate the feasibility of ML-based method in *Arabidopsis*–pathogen PPI prediction, we used RF to train the classification model. Because the sequence information is easy to obtain, most feature encodings are sequence-based. We critically assessed four commonly used sequence-based encoding schemes (DPC, AC, CT and PSP). Each encoding scheme was applied to predict the interactions between *Arabidopsis* proteins and the effectors from *Psy*, *Hpa*, *Gor* or all of the three pathogens. Here we mainly use the auPRC values to quantify the performance. As MCC is also a comprehensive evaluation metric, the maximum MCC of each PR curve was also reported. Moreover, the other assessment parameters corresponding to the maximum MCC were also recorded. The average performance of each sequence-based encoding scheme in the 10-fold cross-validation test is listed in Supplementary Table S2. In general, the performances of DPC, CT and PSP are comparable, which outperform the AC encoding. Taking the prediction of *Ara-all_pathogens* as an example, the auPRC values are 0.759, 0.749 and 0.746 for DPC, CT and PSP, while the corresponding value is 0.706 for AC. Similar results were also observed in the independent test (Supplementary Table S3). Although the dimension of the PSP encoding is much lower than the other three encodings, it still obtained a relatively good performance either in the 10-fold cross-validation test (Supplementary Table S2) or the independent test (Supplementary Table S3), indicating that predicted structural features can capture effective information that distinguishes inter-species PPIs from non-PPIs.

We further evaluated the performance of the overall sequence encoding that integrates DPC, CT, AC and PSP. The performance of the 10-fold cross-validation test is summarized in Figure 3 and Supplementary Table S2. The auPRC values are 0.634, 0.690, 0.734 and 0.776 for *Ara-Psy*, *Ara-Hpa*, *Ara-Gor* and *Ara-all_pathogens*, respectively. Compared with the optimal performance achieved by single encoding, the overall sequence encoding could improve the performance in both cross-validation test (Supplementary Table S2) and independent test (Supplementary Table S3), indicating the four sequence-based encoding schemes are partially complementary to each other. Nevertheless, the performance improvement after integrating all sequence-based encodings was still not fully satisfactory. Therefore, heterogonous features must be taken into account to develop an accurate inter-species PPI predictor.

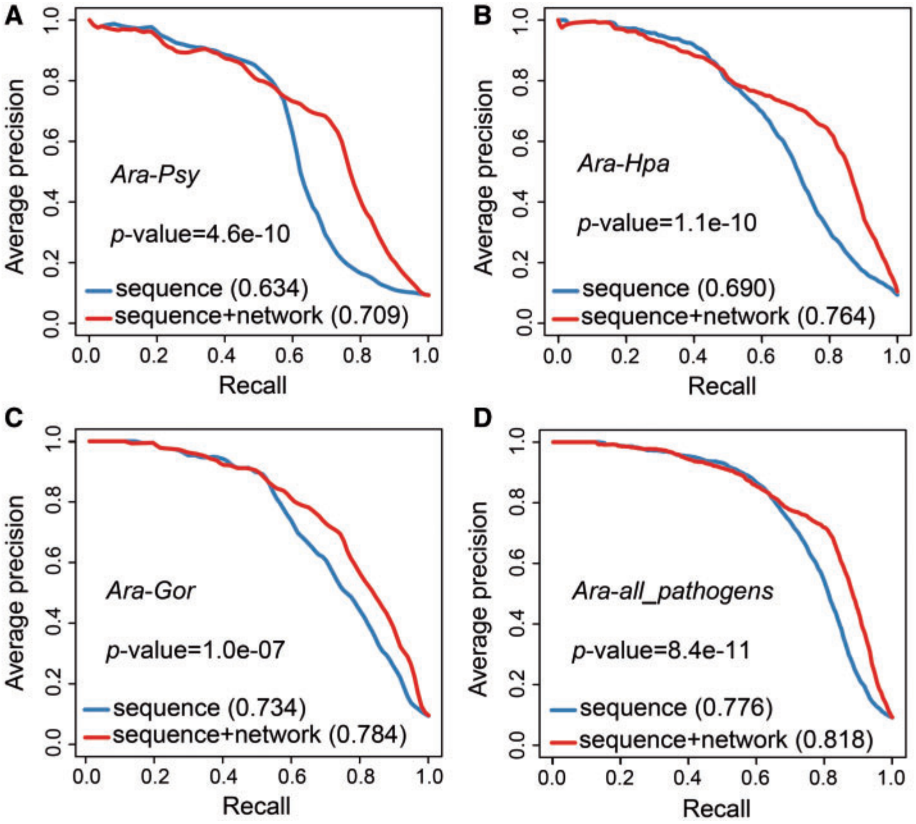

Figure 3. PR curves illustrating the performance of different models on the 10-fold cross-validation test. Panels A, B, C and D represent the results from *Ara-Psy*, *Ara-Hpa*, *Ara-Gor* and *Ara-all_pathogens*, respectively. The corresponding P-values were calculated through one-tailed *t*-test.

Network attributes significantly improved the performance of plant–pathogen PPI prediction

Owing to the complexity of *Arabidopsis*–pathogen PPIs, simple sequence encoding alone is hard to cover the complete biologically meaningful features. We attempted to integrate novel network topology attributes into the sequence-based prediction model, as previous studies have shown that network attributes could distinguish host targets from other proteins [46]. In addition to incorporating several routine network attributes (e.g. degree and betweenness), we would like to emphasize that some novel network distance-based features, which have never been used in predicting inter-species PPIs, were adopted in this work. These new features include the distance to known targets, distance to known R-proteins and distance to predicted R-proteins. To investigate the influence of network attributes on the performance, we compared sequence + network with sequence encoding alone based on the RF algorithm. As shown in Figure 3, Supplementary Tables S2 and S3, the auPRC values of sequence + network encoding in the 10-fold cross-validation test are 0.709, 0.764, 0.784 and 0.818 for *Ara-Psy*, *Ara-Hpa*, *Ara-Gor* and *Ara-all_pathogens*, while the corresponding values of sequence-based encoding are 0.634, 0.690, 0.734 and 0.776. In general, the performance of sequence + network encoding was significantly higher than that of the sequence-based encoding, and the corresponding *P*-values inferred from one-tailed *t*-test are 4.6e-10, 1.1e-10, 1.0e-07 and 8.4e-11 for *Ara-Psy*, *Ara-Hpa*, *Ara-Gor* and *Ara-all_pathogens*, respectively. Regarding the independent test, the auRPC values of sequence + network were also consistently better than the sequence-based encoding (Figure 4). These results clearly show that the network attributes could substantially improve the *Arabidopsis*–pathogen PPI prediction performance.

To explain the effectiveness of network attributes, we analyzed all the network features on the known inter-species PPIs between *Arabidopsis* and the three pathogens. First, the raw value of each feature was normalized by Z-score (Z-score = (x-μ)/σ, where *x* is the raw value, μ is the mean value of the feature and σ is the corresponding standard deviation.). Then, the mean normalized values of each feature in positive, and negative samples were used to plot a radar diagram (Supplementary Figure S2). Indeed, these network attributes clearly differ between positive and negative samples. Network attributes such as degree, betweenness, closeness, transitivity and PageRank in positive samples are larger than those in negative samples, while the distances to known targets and R-proteins show opposite trend. We further examined four features (i.e. degree, betweenness, average distance to known targets and average distance to known R-proteins) on pathogen-specific PPI data sets (Figure 5). The difference between positive and negative samples is consistent on *Ara-Psy*, *Ara-Hpa*, *Ara-Gor* and *Ara-all_pathogens* data sets. Based on the above analyses, several potential patterns in *Arabidopsis*–pathogen PPIs could be summarized. First, the higher degree of known targets in positive samples means that effector proteins tend to attack hubs in host PPI network. It is efficient for pathogens to attack hubs which often play important biological functions. Second, the distances to known targets of an *Arabidopsis* protein in positive samples is shorter than those of an *Arabidopsis* protein in negative samples, indicating effector targets tend to be clustered together in

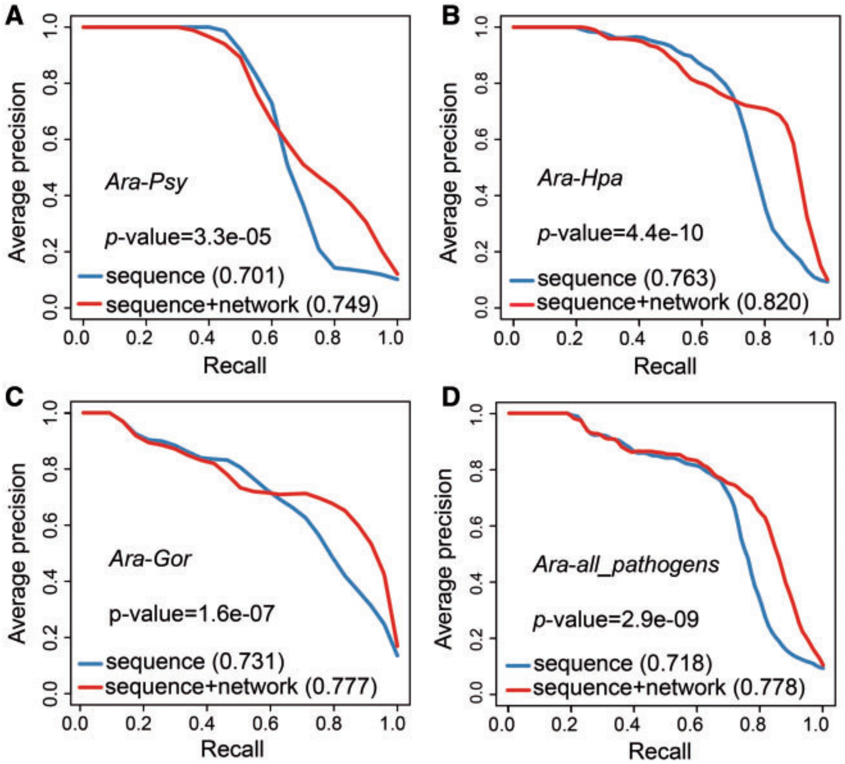

Figure 4. PR curves illustrating the performance of different models on the independent test. Panels A, B, C and D represent the results from *Ara-Psy*, *Ara-Hpa*, *Ara-Gor* and *Ara-all_pathogens*, respectively. The corresponding P-values were calculated through one-tailed *t*-test.

the network. This observation is also in accordance with our previous study [46]. Third, the distance to R-proteins can be used as a good indicator for distinguishing real targets from other *Arabidopsis* proteins because R-proteins can bind effectors directly or monitor the interaction between effectors and targets.

RF outperformed other popular ML algorithms in predicting *Arabidopsis*–pathogen PPIs

To validate that RF is an appropriate ML algorithm for building the prediction models in this work, we compared RF with other ML algorithms on *Ara-all_pathogens* data set. Here, we first tested the performance of different ML algorithms based on the sequence + network encoding in the 10-fold cross-validation test (Supplementary Figure S3A). We found that RF (auPRC = 0.818) achieved the best performance, followed by AdaBoost (auPRC = 0.774), SVM (auPRC = 0.767), KNN (auPRC = 0.760), LR (auPRC = 0.677) and NB (auPRC = 0.643). Similar performance ranking was observed in the independent test (Supplementary Figure S3B). We also compared the performance of different ML algorithms trained with sequence-based encodings alone. Again, RF was suggested to be the best algorithm in both the 10-fold cross-validation test (Supplementary Figure S3C) and independent test (Supplementary Figure S3D). Altogether, RF was the best ML algorithm in predicting *Arabidopsis*–pathogens PPIs, and it was therefore used for the final model construction. Regarding the future development, the ensemble strategy by integrating different ML algorithms may be used to build a better predictor. Moreover, the application of deep learning algorithms [72] could also boost the prediction of *Arabidopsis*–pathogen PPIs.

The implementation of InterSPPI

To facilitate the research community, we developed an online web server named InterSPPI (**Inter**-**S**pecies **P**rotein–**P**rotein **I**nteraction predictor, http://systbio.cau.edu.cn/intersppi/index.php) to predict *Arabidopsis*–pathogen PPIs. The standalone version of InterSPPI is also downloadable at the same Web address. The prediction model of InterSPPI was built based on all *Arabidopsis*–pathogen PPIs. The prediction cutoff was chosen when the precision was 70% and the corresponding specificity was 97%. Users could submit pathogen protein sequences and *Arabidopsis* TAIR IDs to InterSPPI, and then InterSPPI will automatically calculate the possibility of interaction between two query proteins. We hope the web server could help to predict and prioritize *Arabidopsis*–pathogen PPIs for experimental scientists and therefore could further enhance the understanding of the biological mechanisms of pathogen infection and plant immunity.

Cross-species prediction showing the extrapolation of the models based on individual pathogens

To systematically assess our method, we further examined whether a model trained on one *Arabidopsis*–pathogen system could predict inter-species PPIs between *Arabidopsis* and other pathogens. By procedures, we arbitrarily selected the PPI predictor between *Arabidopsis* and one of three pathogens, and assigned all of the inter-species PPIs related to the two remaining pathogens as the test set. As shown in Supplementary Table S4, the auRPC of the sequence + network encoding is always larger than 0.6 for each cross-species test, indicating its robust cross-species prediction performance. Moreover, the sequence + network models also

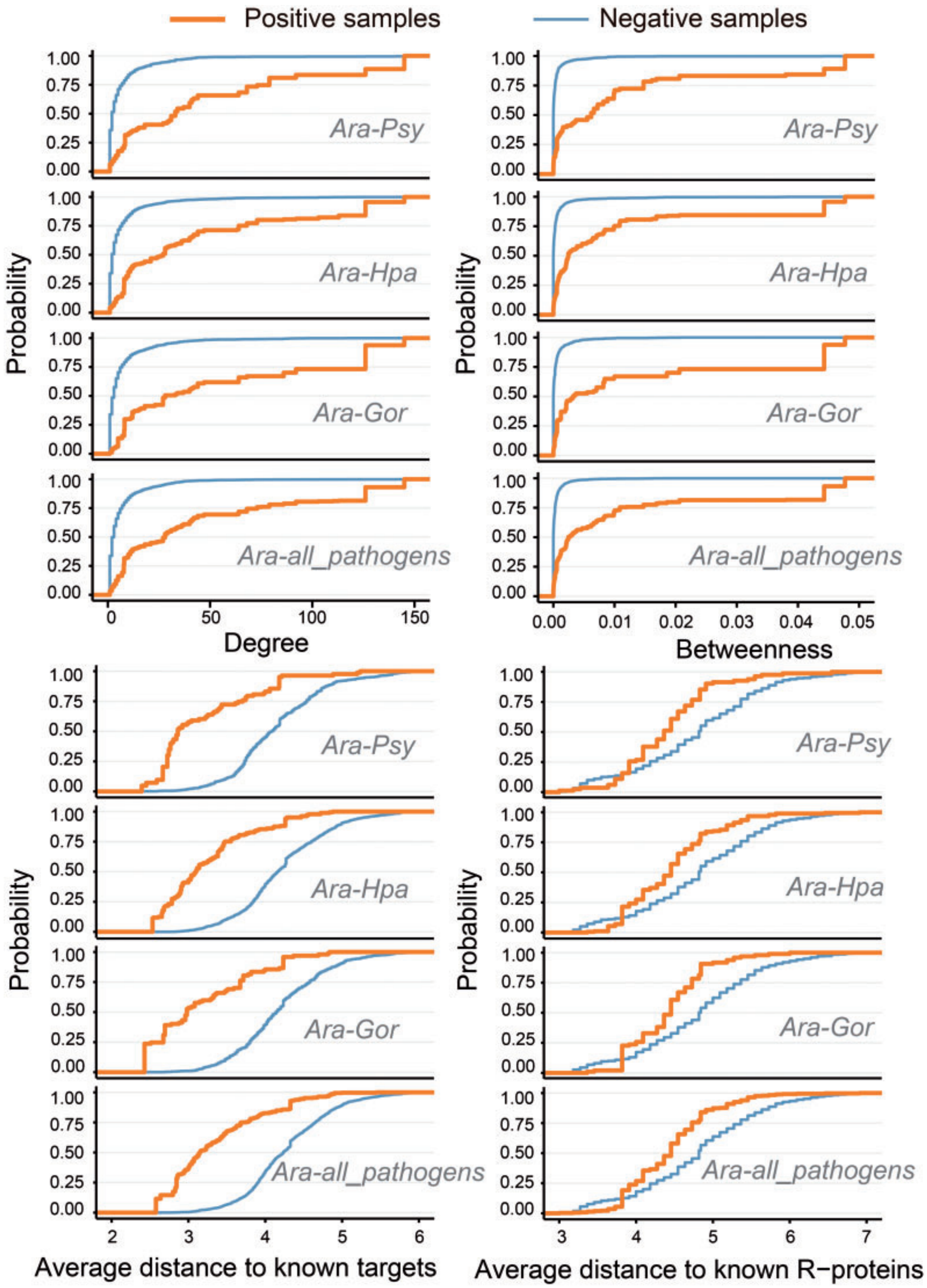

Figure 5. The difference in representative network attributes between positive and negative samples. The cumulative distributions for different network attributes were plotted for comparison.

consistently outperform sequence-based models in these cross-species tests (Supplementary Table S4). The auPRC of sequence + network is increased by an average of 0.16 compared with sequence-based encoding. These results indicate that the sequence + network encoding used by the InterSPPI server is suitable for predicting plant–pathogen PPIs across various pathogen species.

Proteome-wide prediction of *Arabidopsis*–pathogen PPIs

To have a global view of inter-species PPIs for *Ara-Psy*, *Ara-Hpa* and *Ara-Gor*, the pathogen-specific prediction model was used to scan all potential PPIs between *Arabidopsis* and the corresponding pathogen. The cutoff of the RF prediction score was set when the precision is 70%. As listed in Supplementary Table S5, there are 2318, 2260 and 9520 PPIs predicted for *Ara-Psy*, *Ara-Hpa* and *Ara-Gor*, respectively. To further test the reliability of our identified PPIs, we conducted GO enrichment analysis through g:Profiler [73] on all newly identified *Arabidopsis* target proteins. As shown in Supplementary Figure S4, defense-related BPs, such as 'regulation of defense response (GO:0031347)' and 'respond to chitin (GO:0010200)', and hormonal regulation-related terms, including 'respond to salicylic acid (GO:0009751)', 'response to gibberellin (GO:0009739)', 'response to ethylene (GO:0009723)' and 'response to chitin (GO:0010200)', are enriched among these newly identified targets. Previous studies have already reported that these defense processes and hormones are important for plants to defend themselves against pathogens [74, 75]. Therefore, these newly identified PPIs deserve further experimental validation.

GO similarity information for *Arabidopsis*–pathogen PPI prediction

It has been well accepted that two proteins sharing similar function or participating in the same process are more likely to interact [76], and this concept has been successfully applied in PPI prediction. It has also been established that an effector may mimic the interaction way between the host targets and their endogenous interaction partners. Therefore, the GO similarity between an effector and the corresponding host target, and that between the effector and the target's endogenous interaction partners were evaluated. Because the GO information of *Hpa* and *Gor* proteins was missing to a large extent, we only investigated the performance of the GO similarity information on *Ara-Psy*. The performance of the 10-fold cross-validation test was shown in Supplementary Figure S5. As expected, the auPRC value of the individual GO similarity model is only 0.314, indicating that the GO similarity model alone is far from practical application. By integrating GO similarity information with the existing sequence + network encoding, the auPRC value only marginally increases by 0.008, suggesting that the current GO annotation encoding scheme is not useful for improving the prediction performance. Moreover, we also note that the missing rate of GO annotation in *Psy* is still about 60%. A higher coverage of GO annotation information should be necessary to further explore the real application of GO similarity-based features. Taking the above limitations into account, the GO similarity information was not integrated into the final prediction model.

Gene expression features for *Arabidopsis*–pathogen PPI prediction

Transcriptional reprogramming is heavily involved in plant defense responses to pathogens. Therefore, the differential gene expression patterns of plant proteins between the infection and control conditions can be used as an indicator of plant–pathogen PPIs. In this work, we took the *Ara-Psy* system, where the most comprehensive time-series gene expression data (GSE56094) was available, as the example to test whether the gene expression information could improve the prediction performance of *Ara-Psy* PPIs or not. As we can see from Supplementary Figure S5, the auPRC value of the gene expression-based model is 0.521, which is much lower than those of individual sequence-based encoding schemes or network attribute-based encoding. Moreover, when we integrated the gene expression model to the sequence + network model, it could not improve the prediction performance. We note that current gene expression data detect the expression changes of *Arabidopsis* proteins only, while the expression data for pathogen proteins are missing. Thus, the direct co-expression pattern between pathogen proteins and host targets cannot be established. To obtain a comprehensive understanding of plant–pathogen PPIs, the gene expression changes in the corresponding plant and pathogen during infection should be monitored simultaneously. With the rapid development of the next-generation sequencing technologies, dual RNA-Seq can be used for such task, and its application in predicting PPIs from other host–pathogen system has been initiated [77]. We anticipate the advance of dual RNA-Seq will eventually provide important data for predicting plant–pathogen PPIs in the future.

Conclusions

Plant–pathogen PPI prediction is an important research topic, which will significantly promote methodological innovation for studying the PPIs between two different organisms and strengthen the understanding of plant pathology and plant immunity mechanisms. Here, we summarize several key points regarding current investigation on plant–pathogen PPI prediction. (1) Compared with the widely investigated human–pathogen (e.g. human–virus) PPI prediction, the plant–pathogen PPI prediction is rarely addressed. Recently, the accumulation of experimentally verified plant–pathogen PPI data has provided an unprecedented opportunity for building plant–pathogen PPI predictor. (2) Probably owing to the dissimilarity between plant–pathogen inter-species PPIs and intra-species PPIs, traditional PPI prediction methods such as interolog, DDI and DMI failed to effectively infer plant–pathogen PPIs. (3) Our results showed that the combination of sequence and network encoding schemes could lead to an improved ML-based predictor with reasonable performance. (4) Our survey also showed some popular features (e.g. GO similarity and gene expression) are not as useful as they worked in human–pathogen PPI predictions, indicating more *ad hoc* features for predicting plant–pathogen PPIs are continuously required. Regarding the future method development of plant–pathogen PPI predictions, several technical advances would be beneficial to establish a better predictor. First, with the rapid development of interactomics studies, more and more experimentally verified inter-species PPIs will be available in the near future, which will provide more templates for conventional PPI prediction methods as well as provide more training data to develop ML-based predictors. Second, dual RNA-Seq can detect the gene expression changes simultaneously in both the host and pathogen, and therefore this technique enables the measurement of the co-expression between an effector and its host targets, which would be helpful for improving inter-species PPI prediction. Last but not the least, deep learning technique has demonstrated excellent performance in various bioinformatics tasks including gene expression regulation and protein classification, and this technique may also be used to predict inter-species PPIs. Taken together, our results indicate that integration of sequence-based features and network-derived features could result in the improved prediction of *Arabidopsis*–pathogen PPIs, signifying the importance of the integrative feature design that combines heterogeneous biological information to predict inter-species PPIs. In the future, both the accumulation of experimental data and novel prediction methodology would significantly contribute to the improved prediction of inter-species PPIs between plants and pathogens.

Key Points

- Compared with the widely investigated human–pathogen PPI predictions, the plant–pathogen PPI prediction is rarely addressed. The accumulation of experimentally verified plant–pathogen PPI data has provided an unprecedented opportunity for building plant–pathogen PPI predictor.
- Probably owing to the dissimilarity between plant–pathogen inter-species PPIs and intra-species PPIs, traditional PPI prediction methods such as interolog, DDI and DMI failed to effectively infer plant–pathogen PPIs.
- The integration of sequence and network encoding schemes could lead to an improved ML-based predictor with reasonable performance. We have implemented the proposed integration method on a webserver termed as InterSPPI.
- Some popular features (e.g. GO similarity and gene expression) are not as useful as they worked in human-pathogen PPI predictions, indicating more *ad hoc* features for predicting plant–pathogen PPIs are continuously required.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

National Natural Science Foundation of China (31471249).

References

1. Savary S, Ficke A, Aubertot JN, *et al*. Crop losses due to diseases and their implications for global food production losses and food security. *Food Secur* 2012;**4**:519–37.
2. Fang Y, Ramasamy RP. Current and prospective methods for plant disease detection. *Biosensors* 2015;**5**:537–61.
3. Bernardes-de-Assis J, Storari M, Zala M, *et al*. Genetic structure of populations of the rice-infecting pathogen Rhizoctonia solani AG-1 IA from China. *Phytopathology* 2009; **99**:1090–9.
4. Boyd LA, Ridout C, O'Sullivan DM, *et al*. Plant-pathogen interactions: disease resistance in modern agriculture. *Trends Genet* 2013;**29**:233–40.
5. Jones JD, Dangl JL. The plant immune system. *Nature* 2006; **444**:323–9.
6. Dodds PN, Rathjen JP. Plant immunity: towards an integrated view of plant-pathogen interactions. *Nat Rev Genet* 2010;**11**: 539–48.
7. Ito T, Chiba T, Ozawa R, *et al*. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* 2001;**98**:4569–74.
8. Gavin AC, Bösche M, Krause R, *et al*. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002;**415**:141–7.
9. Guirimand T, Delmotte S, Navratil V. VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic Acids Res* 2015;**43**:D583–7.
10. Ammari MG, Gresham CR, McCarthy FM, *et al*. HPIDB 2.0: a curated database for host–pathogen interactions. *Database* 2016;**2016**:baw103.
11. Durmuş Tekir S, Çakir T, Ardiç E, *et al*. PHISTO: pathogen-host interaction search tool. *Bioinformatics* 2013;**29**:1357–8.
12. Wattam AR, Abraham D, Dalay O, *et al*. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* 2014;**42**:581–91.
13. Calderone A, Licata L, Cesareni G. VirusMentha: a new resource for virus-host protein interactions. *Nucleic Acids Res* 2015;**43**:D588–92.
14. Ako-Adjei D, Fu W, Wallin C, *et al*. HIV-1, Human Interaction database: current status and new features. *Nucleic Acids Res* 2015;**43**:D566–70.
15. Urban M, Cuzick A, Rutherford K, *et al*. PHI-base: a new interface and further additions for the multi-species pathogen-host interactions database. *Nucleic Acids Res* 2017;**45**:D604–10.
16. Mukhtar MS, Carvunis A-r, Dreze M, *et al*. Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science* 2011;**333**:596–601.
17. Kim J-G, Park D, Kim B-C, *et al*. Predicting the interactome of Xanthomonas oryzae pathovar oryzae for target selection and DB service. *BMC Bioinformatics* 2008;**9**:41.
18. Li Z-G, He F, Zhang Z, *et al*. Prediction of protein–protein interactions between Ralstonia solanacearum and Arabidopsis thaliana. *Amino Acids* 2012;**42**:2363–71.
19. Yue J, Zhang D, Ban R, *et al*. PCPPI: a comprehensive database for the prediction of Penicillium-crop protein-protein interactions. *Database* 2017;**2017**:baw170.
20. Zhang K, Li Y, Li T, *et al*. Pathogenicity genes in Ustilaginoidea virens revealed by a predicted protein-protein interaction network. *J Proteome Res* 2017;**16**:1193–206.
21. Sahu SS, Weirick T, Kaundal R. Predicting genome-scale *Arabidopsis-Pseudomonas syringae* interactome using domain and interolog-based approaches. *BMC Bioinformatics* 2014;**15**: S13.
22. Kshirsagar M, Schleker S, Carbonell J, *et al*. Techniques for transferring host-pathogen protein interactions knowledge to new tasks. *Front Microbiol* 2015;**6**:36.
23. Schleker S, Kshirsagar M, Klein-Seetharaman J. Comparing human-Salmonella with plant-Salmonella protein-protein interaction predictions. *Front Microbiol* 2015;**6**:45.
24. Matthews LR, Vaglio P, Reboul J, *et al*. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res* 2001;**11**:2120–6.
25. Deng M, Mehta S, Sun F, *et al*. Inferring domain-domain interactions from protein-protein interactions. *Genome Res* 2002; **12**:1540–8.
26. Dandekar T, Snel B, Huynen M, *et al*. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 1998;**23**:324–8.
27. Pazos F, Valencia A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng* 2001;**14**:609–14.
28. Ge H, Liu Z, Church GM, *et al*. Correlation between transcriptome mapping data from Saccharomyces cerevisiae. *Nat Genet* 2001;**29**:482–6.
29. Zhang QC, Petrey D, Deng L, *et al*. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* 2012;**490**:1–6.
30. Sinha R, Kundrotas PJ, Vakser IA. Docking by structural similarity at protein-protein interfaces. *Proteins* 2010;**78**:3235–41.

31. Dyer MD, Murali TM, Sobral BW. Computational prediction of host-pathogen protein-protein interactions. *Bioinformatics* 2007;**23**:i159–66.
32. Lee S-A, Chan C-h, Tsai C-H, *et al*. Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics* 2008;**9**:S11.
33. Evans P, Dampier W, Ungar L, *et al*. Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC Med Genomics* 2009;**2**):27.
34. Nourani E, Khunjush F, Durmus S. Computational approaches for prediction of pathogen-host protein-protein interactions. *Front Microbiol* 2015;**6**:94.
35. Emamjomeh A, Goliaei B, Zahiri J, *et al*. Predicting protein-protein interactions between human and hepatitis C virus via an ensemble learning method. *Mol Biosyst* 2014;**10**: 3147–54.
36. Cui G, Fang C, Han K. Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinformatics* 2012;**13**:S5.
37. Wuchty S. Computational prediction of host-parasite protein interactions between *P. falciparum* and H. sapiens. *PLoS One* 2011;**6**:e26960.
38. Kshirsagar M, Carbonell J, Klein-Seetharaman J. Multitask learning for host-pathogen protein interactions. *Bioinformatics* 2013;**29**:217–26.
39. Zhou Y, Zhou YS, He F, *et al*. Can simple codon pair usage predict protein-protein interaction?. *Mol Biosyst* 2012;**8**: 1396–404.
40. Guo Y, Yu L, Wen Z, *et al*. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res* 2008;**36**: 3025–30.
41. Shen J, Zhang J, Luo X, *et al*. Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci USA* 2007;**104**:4337–41.
42. Tastan O, Qi Y, Carbonell JG, *et al*. Prediction of interactions between HIV-1 and human proteins by information integration. *Pac Symp Biocomput* 2009;**14**:516–27.
43. Weßling R, Epple P, Altmann S, *et al*. Convergent targeting of a common host protein-network by pathogen effectors from three kingdoms of life. *Cell Host Microbe* 2014;**16**:364–75.
44. Dong X, Lu X, Zhang Z. BEAN 2.0: an integrated web resource for the identification and functional analysis of type III secreted effectors. *Database* 2015;**2015**:bav064.
45. An Y, Wang J, Li C, *et al*. Comprehensive assessment and performance improvement of effector protein predictors for bacterial secretion systems III, IV and VI. *Brief Bioinform* 2016, in press. doi:10.1093/bib/bbw100.
46. Li H, Zhou Y, Zhang Z. Network analysis reveals a common host-pathogen interaction pattern in arabidopsis immune responses. *Front Plant Sci* 2017;**8**:893.
47. Rhee SY, Beavis W, Berardini TZ, *et al*. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res* 2003;**31**:224–8.
48. Stark C, Breitkreutz B-J, Reguly T, *et al*. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;**34**: D535–9.
49. Kerrien S, Aranda B, Breuza L, *et al*. The IntAct molecular interaction database in 2012. *Nucleic Acids Res* 2012;**40**:841–6.
50. Altschul SF, Madden TL, Schäffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
51. Zhou H, Gao S, Nguyen NN, *et al*. Stringent homology-based prediction of *H.* sapiens-M. tuberculosis H37Rv protein-protein interactions. *Biol Direct* 2014;**9**:5.
52. Mosca R, Ceol A, Stein A, *et al*. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res* 2014;**42**:D374–79.
53. Finn RD, Bateman A, Clements J, *et al*. Pfam: the protein families database. *Nucleic Acids Res* 2014;**42**:222–30.
54. Müller G, Onoda T, Muller KT. Soft margins for AdaBoost. *Mach Learn* 2001;**42**:287–320.
55. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2012;**12**:2825–30.
56. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, 233–40.
57. Robin X, Turck N, Hainard A, *et al*. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;**12**:77.
58. Hoskins J, Lovell S, Blundell TL. An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. *Protein Sci* 2006;**15**:1017–29.
59. Meng F, Uversky VN, Kurgan L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cell Mol Life Sci* 2017;**74**:3069–90.
60. Hsu W-L, Oldfield C, Meng J, *et al*. Intrinsic protein disorder and protein-protein interactions. *Pac Symp Biocomput* 2012; 116–27.
61. Yan R, Xu D, Yang J, *et al*. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Sci Rep* 2013;**3**:2619.
62. Dosztányi ZCV, Tompa P, Simon I. IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 2005;**21**:3433–4.
63. Sun J, Sun Y, Ding G, *et al*. InPrePPI: an integrated evaluation method based on genomic context for predicting protein-protein interactions in prokaryotic genomes. *BMC Bioinformatics* 2007;**8**:414.
64. Sanseverino W, Hermoso A, D'Alessandro R, *et al*. PRGdb 2.0: Towards a community-based database model for the analysis of R-genes in plants. *Nucleic Acids Res* 2013;**41**:1167–71.
65. Pemberton JR. The igraph software package for complex network research. *InterJ Complex Syst* 2006;**2**:549–51.
66. Harris MA, Clark J, Ireland A, *et al*. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004;**32**: D258–61.
67. Conesa A, Götz S, García-Gómez JM, *et al*. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005;**21**: 3674–6.
68. Yu G, Li F, Qin Y, *et al*. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 2010;**26**:976–8.
69. Franzosa EA, Xia Y. Structural principles within the human-virus protein-protein interaction network. *Proc Natl Acad Sci USA* 2011;**108**:10538–43.
70. Murakami Y, Mizuguchi K. Homology-based prediction of interactions between proteins using Averaged One-Dependence Estimators. *BMC Bioinformatics* 2014;**15**:213.
71. Du X, Cheng J, Zheng T, *et al*. A novel feature extraction scheme with ensemble coding for protein-protein interaction prediction. *Int J Mol Sci* 2014;**15**:12731–49.
72. Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Brief Bioinform* 2016, in press. doi: 10.1093/bib/bbw068.

73. Reimand J, Arak T, Adler P, *et al.* g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res* 2016;**44**:W83–9.
74. Zhang Y, Cheng YT, Qu N, *et al*. Negative regulation of defense responses in Arabidopsis by two NPR1 paralogs. *Plant J* 2006; **48**:647–6.
75. Kaku H, Nishizawa Y, Ishii-Minami N, *et al*. Plant cells recognize chitin fragments for defense signaling through a plasma membrane receptor. *Proc Natl Acad Sci USA* 2006;**103**: 11086–91.
76. Rhodes DR, Tomlins S, Varambally S, *et al*. Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol* 2005;**23**:951–9.
77. Schulze S, Henkel SG, Driesch D, *et al*. Computational prediction of molecular pathogen-host interactions based on dual transcriptome data. *Front Microbiol* 2015;**6**:65.