Plant Mol Biol (2016) 92:105–116
DOI 10.1007/s11103-016-0498-z

AraPPISite: a database of fine-grained protein–protein interaction site annotations for *Arabidopsis thaliana*

Hong Li[1] · Shiping Yang[1] · Chuan Wang[2] · Yuan Zhou[1] · Ziding Zhang[1]

Received: 6 January 2016 / Accepted: 26 May 2016 / Published online: 23 June 2016

Abstract Knowledge about protein interaction sites provides detailed information of protein–protein interactions (PPIs). To date, nearly 20,000 of PPIs from *Arabidopsis thaliana* have been identified. Nevertheless, the interaction site information has been largely missed by previously published PPI databases. Here, AraPPISite, a database that presents fine-grained interaction details for *A. thaliana* PPIs is established. First, the experimentally determined 3D structures of 27 *A. thaliana* PPIs are collected from the Protein Data Bank database and the predicted 3D structures of 3023 *A. thaliana* PPIs are modeled by using two well-established template-based docking methods. For each experimental/predicted complex structure, AraPPISite not only provides an interactive user interface for browsing interaction sites, but also lists detailed evolutionary and physicochemical properties of these sites. Second, AraPPISite assigns domain–domain interactions or domain–motif interactions to 4286 PPIs whose 3D structures cannot be modeled. In this case, users can easily query protein interaction regions at the sequence level. AraPPISite is a free and user-friendly database, which does not require user registration or any configuration on local machines. We anticipate AraPPISite can serve as a helpful database resource for the users with less experience in structural biology or protein bioinformatics to probe the details of PPIs, and thus accelerate the studies of plant genetics and functional genomics. AraPPISite is available at http://systbio.cau.edu.cn/arappisite/index.html.

Keywords 3D complex structure · *Arabidopsis thaliana* · Database · Domain–domain interaction · Domain–motif interaction · Protein interaction site

Introduction

Protein–protein interactions (PPIs) are heavily involved in a variety of biological processes (Braun et al. 2013). The identification of protein interaction sites is a crucial step to strengthen our understandings about the molecular mechanism and biological significance of PPIs. For instance, in the model plant *Arabidopsis thaliana*, deciphering the characteristics of protein interaction sites could advance research of plant growth, signal transduction and stress response (Arabidopsis Interactome Mapping Consortium 2011; Hashimoto et al. 2012; Lin et al. 2011). Protein interaction sites can be directly annotated from experimentally solved 3D structures of protein complexes. However, about 30 non-redundant heteromers for *A. thaliana* are available in April 2015 release of Protein Data Bank (PDB) (Rose et al. 2015). In contrast, the number of experimentally identified *A. thaliana* PPIs is close to 20,000 according to the statistics from public databases (Chatr-Aryamontri et al. 2013; Lamesch et al. 2012; Orchard et al. 2014). To fill in

Hong Li and Shiping Yang have contributed equally to the work.

Electronic supplementary material The online version of this article (doi:10.1007/s11103-016-0498-z) contains supplementary material, which is available to authorized users.

✉ Yuan Zhou
soontide6825@163.com

✉ Ziding Zhang
zidingzhang@cau.edu.cn

[1] State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China

[2] Department of Plant Biology, Carnegie Institution for Science, Stanford, CA 94305, USA

the big gap, computational methods are playing an important role to add structural information into experimentally determined PPIs.

Template-based docking such as Homology Modeling of Protein Complex (HMPC) and Protein Interactions by Structural Matching (PRISM) is a reliable, efficient and widely-used computational technique to infer the fine-grained interaction site information (Kundrotas and Vakser 2010; Kundrotas et al. 2012; Mosca et al. 2013; Sinha et al. 2010). Based on the idea that homologous protein complexes adopt similar binding modes (Kundrotas et al. 2012), HMPC requires PPIs have homologous template complexes whose 3D structures have been solved. Comparatively, the rationale of PRISM is that if particular surface regions of two interacting proteins resemble a known interface from a template complex, the two interacting proteins possibly mimic the interaction interface (i.e., these proteins interact through analogous regions) (Baspinar et al. 2014). HMPC and PRISM have dramatically expanded the coverage of protein complex structures and provide crucial clues for the identification of protein interaction sites (Fukuhara and Kawabata 2008; Tuncbag et al. 2011).

After the 3D structures of protein complexes are obtained, inferring the physicochemical properties of interaction sites (e.g., the bond types and hot spots) also becomes an easy task. Identifying bond types between interacting residue pairs (i.e., intermolecular bonds), including ionic/electrostatic interaction, hydrogen bond, van der Waals' interaction and cation–π interaction, is conducive to the estimation of binding energy (Krissinel 2010; Krissinel and Henrick 2007; Westermarck et al. 2013). It has also been established that only a small fraction of protein interaction sites called hot spots contribute more significantly to binding energy (Kortemme et al. 2004). Although experimental alanine scanning is a powerful method to determine the contribution of a residue to the affinity of a protein complex, it is generally time-consuming and laborious. To complement the experimental alanine scanning method, various computational methods have been developed to detect hot spots (Morrow and Zhang 2012). One representative method is computational alanine scanning (Kortemme et al. 2004), which measures the change in binding free energy (ΔΔG) when an interacting residue is mutated to alanine. This method can be applied on a large scale to detect hot spots. As reported in the literature (Kortemme et al. 2004), it can correctly predict 79 % of experimental hot spots in a test of 233 mutations.

Even with the assistance from the template-based docking technique, the number of PPIs whose 3D complex structures can be modeled is restricted. In this situation, assigning protein interaction regions (i.e., domains and motifs) instead of elaborate interaction sites to PPIs turns out to be a necessary compromise between the resolution and the coverage. PPIs are often achieved by the reuse of domains or motifs (Mosca et al. 2014) and some databases such as iPfam (Finn et al. 2014b) and Eukaryotic Linear Motif (ELM) (Dinkel et al. 2014) have summarized the domain–domain interactions (DDIs) and the domain–motif interactions (DMIs) in the PDB database. These make it possible to infer the locations of interacting domains/motifs in interacting protein pairs.

To date, several structure-related PPI databases have been established (Higurashi et al. 2009; Mosca et al. 2013; Yao et al. 2014). However, there is still room for improvement. On the one hand, most of them do not provide the information of protein interaction sites. On the other hand, for the databases describing the residue-level interaction sites, they only record the PPIs having experimental structures or adopt a single approach to model the 3D complex structures, reducing the coverage for PPIs with interaction details. Moreover, there is no structure-related PPI database specially designed for *A. thaliana*. To facilitate the research community of plant science, here we constructed a comprehensive database called AraPPISite to provide protein interaction site annotations for *A. thaliana*.

Results

Contents and features

In total, the interaction sites of 7336 *A. thaliana* PPIs are annotated in AraPPISite, which can be classified into two types. In the first type of interaction site annotations, AraPPISite stores 27 PPIs (0.4 %) with experimental complex structures and 3023 PPIs (41.2 %) with predicted complex structures, in which 1677 (22.9 %) and 1346 (18.3 %) PPIs are modeled using HMPC and PRISM respectively (Fig. 1). Although it is possible that multiple complex structures per PPI could be predicted, only the most favored one is kept in AraPPISite (details in "Methods" section). AraPPISite not only provides the fine-grained protein interaction sites of these PPIs, but also lists the bond types, the ΔΔG and the residue conservation of interaction sites to assist users

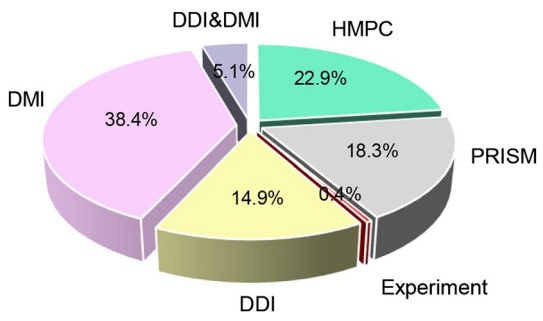


Fig. 1 The proportions of PPIs with different annotation information

to identify critical interaction sites. In the second type of interaction site annotations, AraPPISite assigns the DDI/DMI information for 4286 PPIs (58.4%) whose 3D structures cannot be modeled. Among these, 1097 PPIs (14.9%) have only the DDI information, 2817 PPIs (38.4%) have only the DMI information, and 372 PPIs (5.1%) have both information assigned (Fig. 1). AraPPISite is an open and user-friendly database. Users do not need configure their local machines to visualize the results. The functionality and usage of AraPPISite are detailed in the help page. Moreover, AraPPISite allows users to obtain the full dataset in the download page (http://systbio.cau.edu.cn/arappisite/download.php). Meanwhile, we have also made the data available at https://www.dropbox.com/s/67gqho63wnwvvte/supplementary_material.tar.gz?dl=0.

Accessing specific PPIs

Users can query an interested PPI through either the search page or the network page. In the search page, users could enter a pair of protein IDs to start a search. The Arabidopsis Information Resource (TAIR) identifier, UniProt accession number, Gene name and keyword are supported. In the network page, users can access the PPIs with 3D structures and with DDI/DMI annotations from two different networks, respectively. Taking the network search for the PPIs with 3D structures as an example, users need to submit a TAIR identifier to the search box, and AraPPISite will present the subnetwork of query protein on the left side of the page. Nodes represent proteins and edges represent interactions among the proteins. Users can intuitively observe the interacting partners of query protein from the subnetwork and interactively translate, zoom in or out the view of subnetwork to a suitable scale. Furthermore, the interacting partners of query protein are also listed on the right side of the page. By clicking the "AraPPISite" link of interacting partner, the webpage browser will jump to the result page of PPI between the query protein and the interacting partner.

Visualizing protein interaction sites

The visualization of protein interaction sites with 3D structures can be exemplified in Fig. 2, which illustrates the interaction details between AT5G24270 and AT5G01820 modeled by using HMPC. This result page can be divided into three parts: the basic information table, the 3D structure of protein complex and the list of interaction sites. The basic information table exhibits different resource identifiers, modelling pipeline logs and brief descriptions of protein functions based on the UniProt annotation (UniProt Consortium 2015). More information about the query protein or its template could be obtained by clicking the link of corresponding identifier.

AraPPISite provides the sequence alignments that were used to model the 3D structures. By clicking the "+" button, the alignment result will be shown as in Fig. 2. The protein interaction sites are colored red or blue in sequences. Users can click each residue in the protein sequences to highlight it on the 3D structure, one residue at a time. The selected residue will be colored yellow, rendered using surface representation and labeled with the residue type and its position.

The 3D structure visualization of PPI at atomic resolution is the most fine-grained view of interaction sites provided in AraPPISite (Fig. 2). Once interested interacting residue pairs in the list of interaction sites are selected, users can click the "show" button to highlight the corresponding interacting residue pairs as a stick representation in the 3D structure cartoon. To view the atom-level interaction more clearly, the 3D structure cartoon can be zoomed in and rotated/translated to a suitable scale through mouse operations. It is worth mentioning that AraPPISite allows users to select and highlight more than one pair of interacting residues simultaneously. The highlighted interacting residue pairs can also be swept away from the 3D structure by clicking the "Clear All" button. The PDB file of 3D structure in this page is downloadable by clicking the "download" button above the visualization.

Users can also browse the interacting residue pairs in the list of interaction sites (Fig. 2). To estimate the importance of each interacting residue, AraPPISite provides the information of bond types, $\Delta\Delta G$ and conservation score. The $\Delta\Delta G$ is the change in binding free energy when an interacting residue is mutated to alanine. A $\Delta\Delta G$ value greater than or equal to 1 kcal/mol means the mutated interacting residue may be a hot spot (Kortemme et al. 2004). The residue conservation has also been integrated to predict the hot spot (Caffrey et al. 2004). The lower the score is, the higher conservation the interaction site has (Pupko et al. 2002). The interacting residue pairs can be sorted according to the $\Delta\Delta G$ value or the conservation score to prioritize the putative hot spots.

Visualizing protein interacting domains and motifs

AraPPISite also provides the sequence-level interaction site annotations (i.e., the DDI/DMI annotations) for PPIs whose 3D structures cannot be modeled. Taking the corresponding annotation page between AT1G16010 and AT5G10450 as an example (Fig. 3), the page includes two parts, i.e., the basic information table and the list of DDIs/DMIs. In the basic information table as mentioned in Fig. 2, AraPPISite provides the identifiers, the related links and the brief description of protein function. In the list of DDI/DMI information, AraPPISite presents all possible DDIs and DMIs as well as the corresponding interacting regions of a PPI. The domain and motif definitions are provided by the Pfam (Finn et al.

AraPPISite

Home | Search | Network | Download | Help

PPI: AT5G24270-AT5G01820 **(Homology Modeling of Protein Complex)**

TAIR	Uniprot	Gene name	Template|chain	Range	SeqId	Coverage	Description
AT5G24270	O81223	CBL4	2zfd | A	21-203	0.62	0.82	Calcineurin B-like protein 4
AT5G01820	Q9LZW4	CIPK14	2zfd | B	318-426	0.88	0.25	CBL-interacting serine/threonine-protein kinase 14

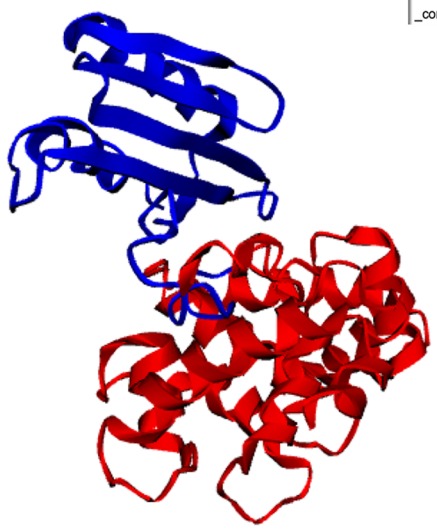
AT5G24270 Target-Template Alignment

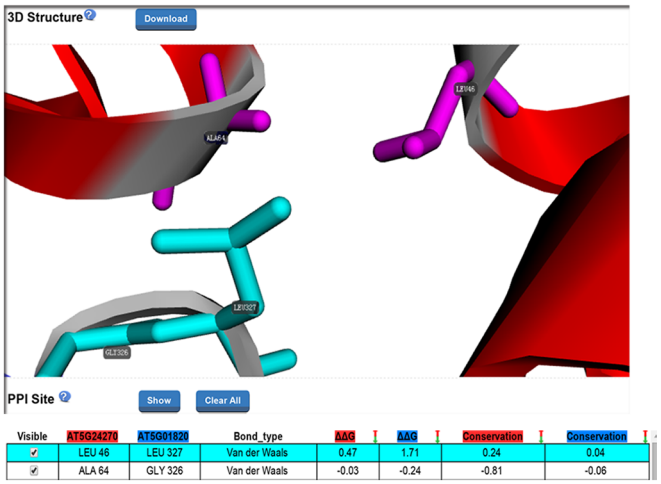
AT5G01820 Target-Template Alignment

3D Structure Download

AT5G24270 Target-Template Alignment

```
2zfd_A      DPELLARDTVFSVSEIEALYELFKKISSAVIDDGLINKEEFQLALFKTNKKESLFADRVFDLFDTKHNGI
AT5G24270   DPELLASVTPFTVEEVEALYELFKKLSSSIIDDGLIHKEEFQLALFRNNRNRNLFADRIFDVFDVKRNGV
_conserve   *******--*-*-*-*-*********-**--*******-*********-------*****-**-**-*-**-
2zfd_A      LGFEEFARALSVFHPNAPIDDKIHFSFQLYDLKQQGFIERQEVKQMVVATLAESGMNLKDTVIEDIIDKT
AT5G24270   IEFGEFVRSLGVFHPSAPVHEKVKFAFKLYDLRQTGFIEREELKEMVVALLHESELVLSEDMIEVMVDKA
_conserve   --*-**-*-*-*****-**----*--*-*-****-*-*****-*-*-****-*-**----*----**---**-
2zfd_A      FEEADTKHDGKIDKEEWRSLVLRHPSLLKNMTLQYLKDITTTF
AT5G24270   FVQADRKNDGKIDIDEWKDFVSLNPSLIKNMTLPYLKDINRTF
_conserve   *--**-*-******--**----*----***-*****-******--**
```





Visible	AT5G24270	AT5G01820	Bond_type	ΔΔG	ΔΔG	Conservation	Conservation
☑	LEU 46	LEU 327	Van der Waals	0.47	1.71	0.24	0.04
☑	ALA 64	GLY 326	Van der Waals	-0.03	-0.24	-0.81	-0.06

PPI Site Show Clear All

Visible	AT5G24270	AT5G01820	Bond_type	ΔΔG	ΔΔG	Conservation	Conservation
☑	LEU 46	LEU 327	Van der Waals	0.47	1.71	0.24	0.04
☑	ALA 64	GLY 326	Van der Waals	-0.03	-0.24	-0.81	-0.06
☐	ALA 64	LEU 327	Van der Waals|H-Bond	-0.03	1.71	-0.81	0.04
☐	LEU 65	LEU 324	Van der Waals	0.89	2.30	-0.16	0.23
☐	LEU 65	GLY 326	Van der Waals	0.89	-0.24	-0.16	-0.06
☐	LEU 65	LEU 327	Van der Waals	0.89	1.71	-0.16	0.04
☐	LEU 65	ARG 336	Van der Waals	0.89	0.32	-0.16	2.37
☐	PHE 66	GLY 321	Van der Waals	0.46	0.13	-0.69	-0.74
☐	PHE 75	PHE 322	Van der Waals	1.86	2.74	-0.69	0.14

◀ **Fig. 2** An example of an AraPPISite entry exhibiting the predicted 3D structure. The webpage consists of a basic information table, a 3D structure of protein complex and a list of interaction sites. The basic information table describes the protein information and the modelling logs. By clicking the "+" button, the sequence alignment between the target and the template will be displayed. The protein interaction sites on the sequences are colored *red* or *blue* in accordance with their respective structures. Moreover, the interacting residue pairs can be highlighted on the 3D structure and ranked according to the values of ΔΔG or residue conservation

2014a) and ELM (Dinkel et al. 2014) databases respectively. Users can easily jump to these resources to search the description of corresponding domain or motif. Because most of the motifs are not firstly discovered in *A. thaliana*, the taxonomy information of the corresponding motif has been made accessible. Users can obtain this information by moving the mouse cursor on the motif identifier (Supplementary Note 1; Table S1).

The detailed view of interaction domains/motifs is displayed in the pop-up box triggered by clicking the "view" button (Fig. 3). Users can clearly observe the domain/motif positions in proteins and jump to related databases by clicking the red/blue color bars in the graphical representation. Special attention should be paid if a protein has repetitive motifs. These motifs are numbered by their positions in the protein sequence and represented as bars filled with identical color (Fig. 3, red bars). The protein sequences are also highlighted so that users can easily locate the interaction domains/motifs in the sequences.

Evaluating the accuracy of predicted complex structures and interaction sites

Although HMPC and PRISM have been widely used to predict protein complex structures and their reliability has been well recognized to the community, the applications of these two methods in our work rely on predicted monomer structures. Undoubtedly, the predicted 3D structures can introduce additional errors for the prediction of complex structures. Therefore, it is very important to benchmark the accuracy of predicted complex structures in AraPPISite. To this end, we used 27 experimentally determined complex structures of *A. thaliana* PPIs as a test set (i.e., golden standard dataset) to assess the performance of HMPC and PRISM. We first predicted the monomer structures of these PPIs through Modeller. For rigorous assessments, the experimental complex structures for these PPIs were excluded from our template library. To cover more data, all the available homologous templates were used to build the monomer structures (see "Methods" section for the sequence identity cutoff of homologous template searching). Then, HMPC and PRISM were employed to build the complex structures from the predicted monomer structures. Note that we may build multiple complex structures for one PPI due to the existence of multiple structures predicted for one monomer.

As a result, 91 protein complex structures involved in 15 PPIs and 156 protein complex structures involved in 20 PPIs are modeled by using HMPC and PRISM, respectively. According to the evaluation criterion of Critical Assessment of Predicted Interactions (CAPRI), which is a community-wide experiment to assess the accuracy of predicted 3D structures of protein complexes, the predicted protein complex structures can be grouped into four categories on the basis of the backbone root-mean-square deviation of interface residues (I_RMS): high accuracy (I_RMS$\leq$1.0 Å), medium accuracy (1.0 Å$<$I_RMS$\leq$2.0 Å), acceptable (2.0 Å$<$I_RMS$\leq$4.0 Å) and incorrect (I_RMS$>$4 Å) (Mendez et al. 2003).

With respect to the models predicted by HMPC, 26 (28.6%) of 91 complex structures achieve high accuracy, 87 (95.6%) complex structures are predicted correctly (i.e., I_RMS$\leq$4.0 Å), and only 4 (4.4%) complex structures are predicted incorrectly (Table 1), which indicates that HMPC has the capacity of predicting the 3D structures of PPIs reliably. Regarding the 156 complex structures predicted by PRISM, 131 (84.0%) and 25 (16.0%) are predicted correctly and incorrectly, respectively (Table 1).

Note that only one complex structure per PPI was adopted in AraPPISite. To ensure that the evaluation is very consistent with the real situation in AraPPISite, we reassessed the performance based on one preferred complex structure per PPI. The selection criteria of the most favored complex structure are the same as those used in AraPPISite (details in "Methods" section). With respect to the 15 unique complex structures predicted by HMPC, 4 (26.7%) complex structures achieve high accuracy, 13 (86.7%) complex structures are predicted correctly, and 2 (13.3%) complex structures are predicted incorrectly (Table 1). Regarding the 20 complex structures predicted by PRISM, 17 (85.0%) and 3 (15.0%) are predicted correctly and incorrectly, respectively (Table 1). Generally, the performance based on one predicted complex structure per PPI is in good agreement with that based on multiple complex structures per PPI. We also calculated the proportions of correctly predicted interacting residues in experimental complex structures to further assess the performance of HMPC and PRISM (Supplementary Note 2; Figs. S1, S2). In line with the performance assessment based on the I_RMS values, the accuracy of PRISM is inferior to that of HMPC, meaning that more cautions should be taken when dealing with the complex structures as well as the corresponding interaction sites inferred from PRISM.

Due to the prediction principle of HMPC, the quality of predicted protein complex structures should be relevant to the sequence identity between interacting proteins and their templates. As expected, among the HMPC-predicted

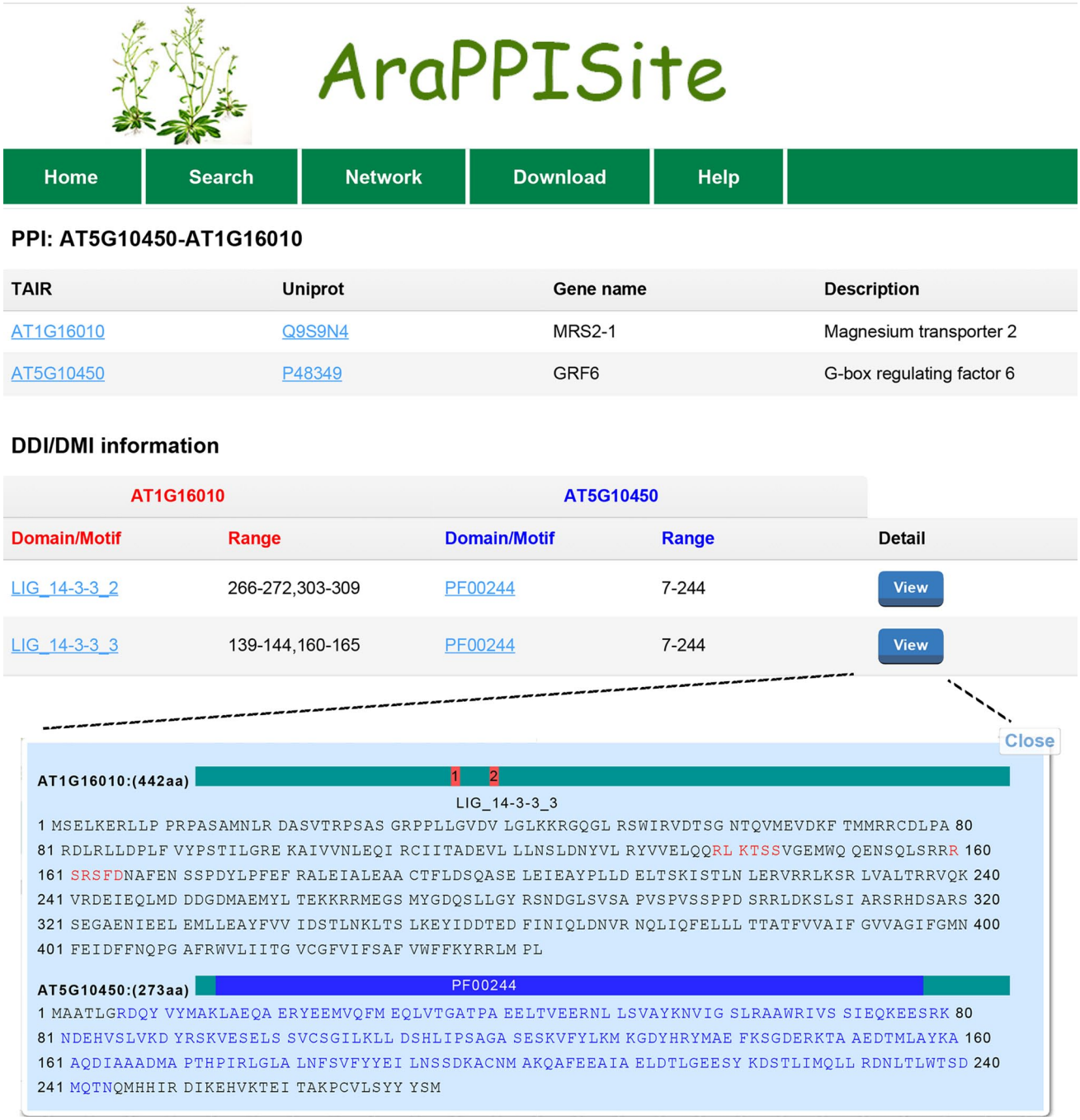


Fig. 3 An example of an AraPPISite entry with the DDI/DMI annotation. The webpage consists of a basic information table and a list of DDIs/DMIs. The basic information table describes the protein information. In the list of DDIs/DMIs, each line presents a possible DDI/DMI and the sequence ranges of domain/motif on the sequences (i.e., the interaction regions). By clicking the "view" button, more details will be graphically displayed. The residues involved in the interactions are colored *red* or *blue* on the sequences

complex structures, the I_RMS values and the sequence identities show a strong negative correlation (Fig. 4a, c). When considering one complex structure per PPI, for instance, the Pearson correlation coefficient (PCC) between the I_RMS values and the sequence identities is –0.7 (Fig. 4c). Therefore, we can further exploit this correlation to assess the confidence of the HMPC complex structures in AraPPISite (Supplementary Note 3; Fig. S1). By contrast, the correlation between the I_RMS values and the sequence identities in the PRISM complex structures is very weak (Fig. 4b, d), which precludes the reliability estimation of the PRISM complex structures in AraPPISite based on the sequence identities.

Table 1 The performance assessment of HMPC and PRISM

	High accuracy (I_RMS≤1.0 Å)	Medium accuracy (1.0 Å<I_RMS≤2.0 Å)	Acceptable (2.0 Å<I_RMS≤4.0 Å)	Incorrect (I_RMS>4 Å)
Multiple predicted complex structures per PPI				
HMPC[a]	26 (28.6%)	30 (33.0%)	31 (34.0%)	4 (4.4%)
PRISM[a]	25 (16.0%)	31 (19.9%)	75 (48.1%)	25 (16.0%)
One predicted complex structure per PPI				
HMPC[b]	4 (26.7%)	5 (33.3%)	4 (26.7%)	2 (13.3%)
PRISM[b]	3 (15.0%)	3 (15.0%)	11 (55.0%)	3 (15.0%)

The number and percentage of predicted complex structures achieving the high accuracy, medium accuracy, acceptable or incorrect prediction is listed in this table

[a]When multiple predicted complex structures per PPI are taken into account, 91 protein complex structures involved in 15 PPIs and 156 protein complex structures involved in 20 PPIs can be predicted using HMPC and PRISM, respectively

[b]When dealing with only one predicted complex structure per PPI, 15 and 20 complex structures are predicted using HMPC and PRISM, respectively

We also resorted to known mutagenesis information to evaluate the accuracy of predicted interaction sites. We downloaded all the *A. thaliana* proteins with mutagenesis annotations from the UniProt database (UniProt Consortium 2015). The mutagenesis information, which was derived from the ad hoc site-directed mutagenesis experiments, records the effects of experimentally mutated residues on the biological properties of proteins, including the reduced or increased interactions with their corresponding partners. Among the 3023 predicted 3D complex structures, only 16 proteins contain the information about how mutated residues can reduce or increase the interactions with partners. Interestingly, 14 out of 33 such interaction-influencing mutated residues are also predicted interaction sites in AraPPISite (Table 2). And 11 out of 16 proteins have at least one mutated residue overlapping with the predicted interaction sites. This adds new experimental evidence to prove the quality of predicted interaction sites in AraPPISite. Taken together, the above two assessment analyses validate that the predicted 3D complex structures and interaction sites deposited in AraPPISite are generally reliable, and thus they are valuable to the community.

Discussion

PPIs provide a systems understanding of protein functionality at the cellular level (Gu et al. 2011; He et al. 2010; Li et al. 2015). To thoroughly understand the molecular mechanism of PPIs, the identification of protein interaction sites is a crucial step. We tried to retrieve *A. thaliana* protein interaction sites from literatures. However, only the interaction domains of 36 PPIs have been reported, among which the interaction domains of 33 PPIs are known for only one of two interacting proteins rather than both of them. In other words, the information about protein interaction sites is largely unknown for *A. thaliana*. Solving the 3D structures of protein complexes is a straightforward method for the identification of interaction sites. Due to the technical limitation, however, no more than 40 non-redundant heterodimers are available in the PDB database, which is negligible in comparison with the number of experimentally identified *A. thaliana* PPIs.

The emergence of computational methods presents a new perspective to capture interaction details. In this study, we designed a database of protein interaction site annotations, AraPPISite, for *A. thaliana*. AraPPISite has advantages over existing structure-related PPI databases. First, AraPPISite models 3023 3D structures of protein complexes and provides the information of protein interaction sites on the basis of the 3D structures. Second, AraPPISite lists the evolutionary and physicochemical properties to assist users to determine critical interaction sites. Third, AraPPISite assigns the DDIs/DMIs to 4286 PPIs without predicted 3D structures, which increases the coverage of PPIs having interaction details. Last but not least, AraPPISite provides the interactive graphical representation of protein interaction sites while does not require any user configuration. Taken together, AraPPISite is an easy-to-use and enriched database resource for plant biologists with less experience in structural biology.

Nevertheless, AraPPISite relies on experimentally derived template protein complex structures. With the advance of structural proteomics, more 3D structures of *A. thaliana* PPIs can be modeled in the future. Regarding the future perspective, we will extend our efforts to model the 3D structures of protein complexes for food crops such as *Oryza sativa* and *Zea mays*, which will accelerate the analyses of protein functions to improve the grain yield. Moreover, we plan to integrate a predictor in AraPPISite that automatically models the protein complex structure by submitting two query protein sequences. We hope AraPPISite will become an important data resource to strengthen our understanding of plant structural interactome at a higher resolution.

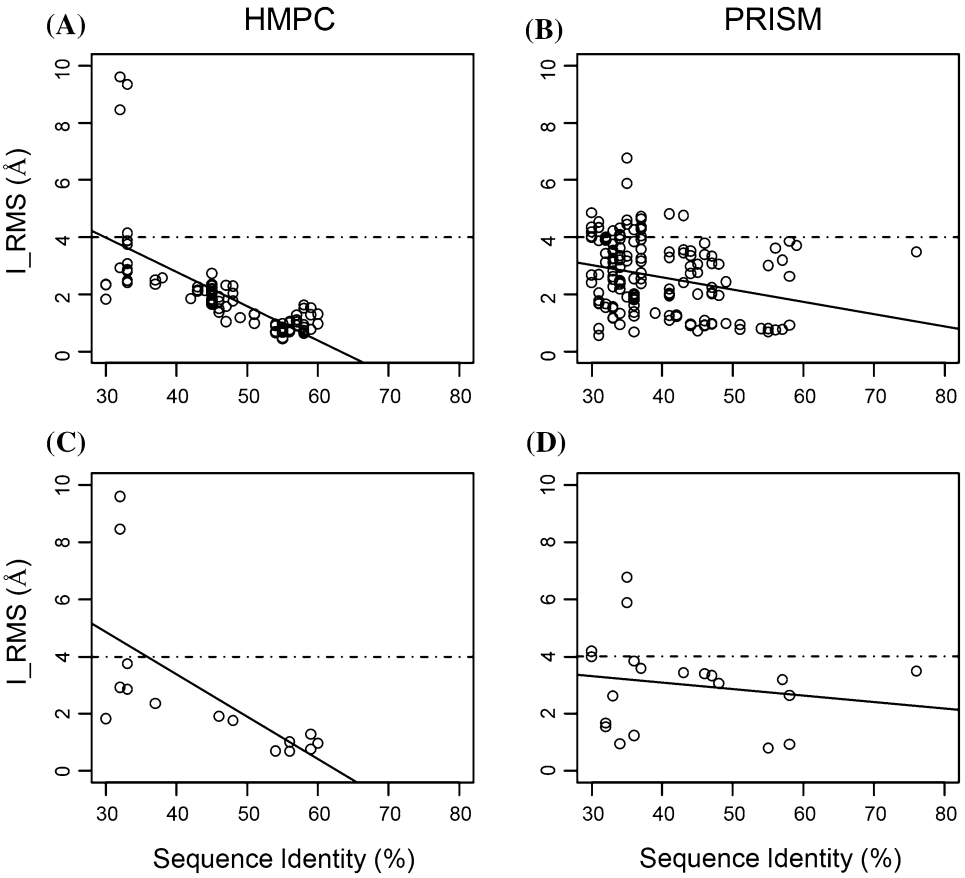


Fig. 4 The relationship between I_RMS values and sequence identities. **a** The relationship based on the predicted 91 HMPC complex structures modeled (PCC = –0.7); **b** the relationship based on the predicted 156 PRISM complex structures (PCC = –0.3); **c** the relationship based on the 15 unique protein complex structures predicted by HMPC (PCC = –0.7); and **d** the relationship based on the 20 unique protein complex structures predicted by PRISM (PCC = –0.2). The *solid line* in each plot represents the linear regression line that models the relationship between the I_RMS values and the sequence identities. The *dashed line* in each plot stands for the I_RMS boundary of correctly and incorrectly predicted complex structures. Note that the sequence identity in X-axis represents the lower one of the two target-template identity values between an interacting protein pair

Methods

Processing experimental protein complex structures

Fifty-four experimental complex structures involved in 34 non-redundant *A. thaliana* binary PPIs were downloaded from the PDB database. The complex structure with the best resolution was recorded for each PPI. When the two interacting proteins mapped multiple chains in the complex structure, the two chains having the largest interaction interface were selected. Then, 27 PPIs whose two chains both covered at least 50 amino acids were stored in our database. The residues from two interacting proteins were defined as the interacting residues if the shortest distance between their atoms was less than 4 Å. The whole interaction site (interface) between two interacting proteins was comprised of all individual interacting residues.

Predicting complex structures of PPIs

Arabidopsis thaliana binary PPIs were collected from three public PPI databases in October 2013, including BioGRID (Chatr-Aryamontri et al. 2013), IntAct (Orchard et al. 2014) and TAIR (Lamesch et al. 2012). To merge the PPIs from different resources, we remapped all protein references to TAIR identifiers using the ID mapping provided by the UniProt database (UniProt Consortium 2015). At the same time, corresponding representative protein sequences were also downloaded from the TAIR database (TAIR10 release). The unmapped PPIs and self-interactions (homo-oligomers) were discarded. Further, the PPIs having experimental complex structures were removed. As a result, 18,647 PPIs were obtained.

To obtain protein interaction sites, modeling the 3D structures of protein complexes was one prerequisite. The first step was to search homologous templates in the PDB database (Rose et al. 2015) using Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990) (Fig. 5). The 3D structure of a PPI could be modeled if the two interacting proteins and their templates shared at least 30% sequence identity (Kundrotas et al. 2012; Mosca et al. 2013). These PPIs were divided into two categories: (1) the templates of two interacting proteins came from different chains of the same protein complex; (2) the templates of two interacting proteins came from different PDB files. For these two categories of PPIs, HMPC and PRISM were separately exploited to model the 3D structures as described in the following paragraphs.

For the first category of PPIs, the next step was to select the best template complexes. The templates having the higher

Table 2 The known mutagenesis information related to predicted interaction sites

Proteins[a]	Mutagenesis[b]	Interaction sites[c]	Partners[a]
AT3G48750 (P24100)	234, 235, 236 (loss of interaction with the partner)	236	AT2G27960 (O23249)
AT2G27960 (O23249)	61 (impaired interaction with the partner)	61	AT3G48750 (P24100)
AT2G45770 (O80842)	71, 109, 326 (reduced interaction with the partner)	71, 109	AT5G03940 (P37107)
AT2G27960 (O23249)	61 (impaired interaction with the partner)	61	AT3G54180 (P25859)
AT4G26080 (P49597)	180 (impaired ABA-mediated binding to the partner)	180	AT4G17870 (O49686)
AT5G57050 (O04719)	168 (impaired ABA-mediated binding to the partner)	168	AT4G17870 (O49686)
AT1G79650 (Q84L33)	47 (abolishes the interaction with the partner)	47	AT4G38630 (P55034)
AT3G02540 (Q84L31)	8, 47 (abolishes the interaction with the partner)	8	AT4G38630 (P55034)
AT5G38470 (Q84L30)	8, 46 (abolishes the interaction with the partner)	46	AT4G38630 (P55034)
AT3G26090 (Q8H1F2)	320 (loss of interaction with the partner)	320	AT2G26300 (P18064)
AT1G02280 (O23680)	45, 46, 47, 48, 49, 50, 130 (impaired interaction with the partner)	45, 46, 47	AT4G02510 (O81283)
AT4G17615 (O81445)	201 (increased interaction with the partner)	NA	AT1G30270 (Q93VD3)
AT5G47100 (Q9LTB8)	201 (increased interaction with the partner)	NA	AT1G30270 (Q93VD3)
AT3G61140 (P45432)	222 (abolishes the interaction with the partner)	NA	AT5G42970 (Q8L5U0)
AT3G59060 (Q84LH8)	31, 32, 37, 38 (loss of binding to the partner)	NA	AT2G18790 (P14713)
AT4G29810 (Q9S7U9)	99, 220, 226 (loss of binding to the partner)	NA	AT4G08500 (Q39008)

[a]The interactions between proteins (column 1) and their partners (column 4) have predicted complex structures in AraPPISite. The identifiers in parentheses are the corresponding UniProt accession numbers

[b]This column lists the known residue mutations that affect the interactions between proteins and the corresponding partners. Keywords describing the mutational effects on the corresponding interactions are listed in parentheses. The numbers in this column denote the residue numbers in sequences. All the mutagenesis information is extracted from the UnitProt database

[c]This column lists the mutated residues belonging to the predicted interaction sites in AraPPISite

sequence identity with the interacting proteins were added to the candidate set. Preference was given to the candidates of X-ray structures over NMR structures. The candidate with the best resolution was considered as the best template to model the 3D structures of interacting proteins separately.

Then, Modeller (version 9.14) (Sali and Blundell 1993) was employed to generate five models for each interacting protein. The model with the lowest DOPE score was selected for each interacting protein and unaligned residues at N-terminal or C-terminal were truncated (Mosca et al. 2013). We computed the exact sequence identity and the coverage between the models and templates according to the alignment results of the salign tool in Modeller. Further, protein–peptide (peptide means no more than 50 amino acids in the model) interactions and peptide–peptide interactions were removed. It should also be noted that Modeller inevitably produced the knotted structure when a long insertion (i.e., longer than 15 residues) occurred in the model compared with the template (http://salilab.org/modeller/FAQ.html#14). Therefore, to guarantee the quality of protein complex structures, we discarded the PPIs containing a long insertion in the interaction interface.

Finally, HMPC was used to model the 3D structures of PPIs. We applied symmetry operations on template atom coordinates to generate the complete complex structure according to REMARK 350 of the PDB file. The template

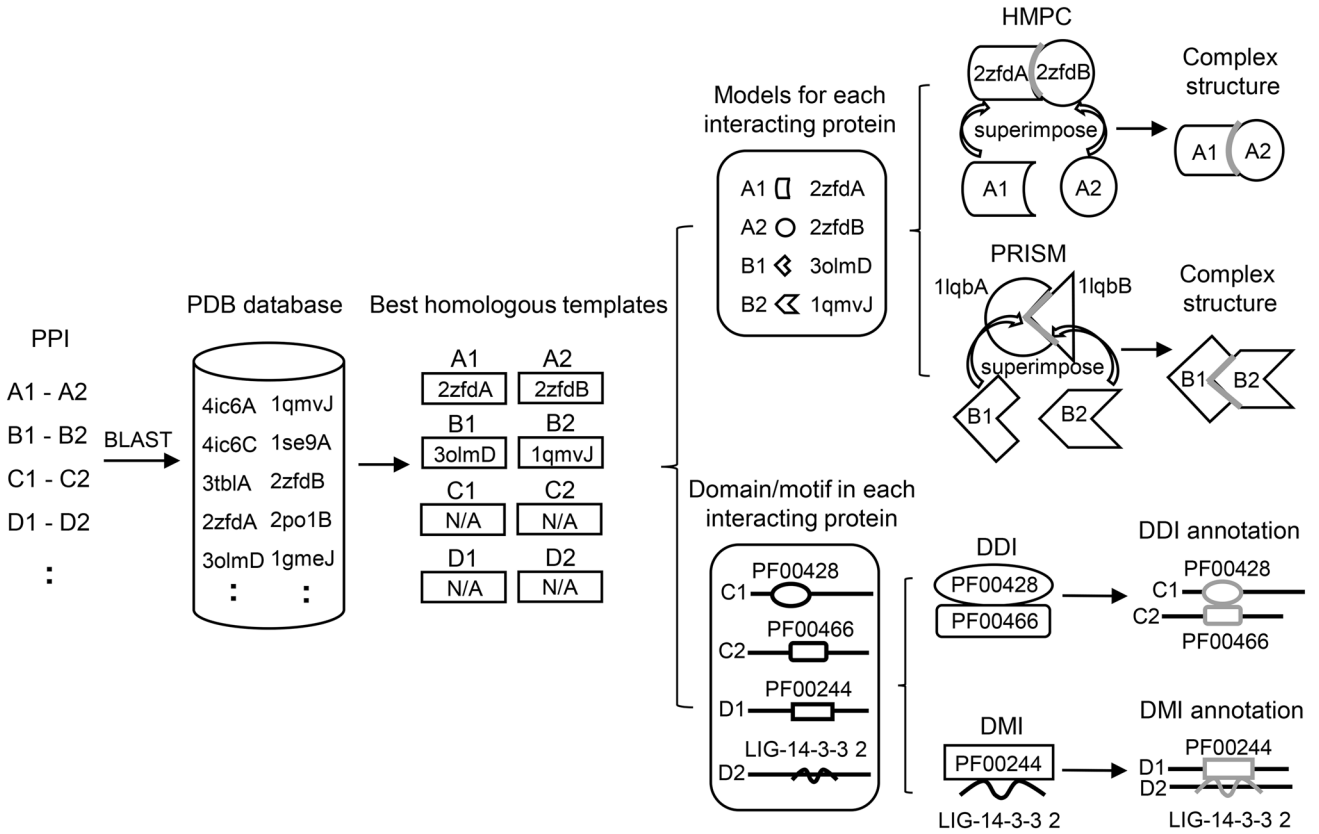


Fig. 5 Overview of the protein interaction site annotation pipeline. There are four illustrative PPI instances (A1–A2, B1–B2, C1–C2, D1–D2). Firstly, homologous templates are searched using BLAST. Then, Modeller is used to generate the 3D models of monomers if the templates are available (the case of A1, A2, B1 and B2). For A1 and A2, their templates are from the same complex structure (PDB entry: 2zfd). The complex structure is thus built by HMPC, namely the two monomer structures (A1 and A2) are superimposed onto the corresponding chains of template complex structure (2zfdA and 2zfdB) to infer the predicted complex structure of A1–A2. *Gray* represents PPI regions. For B1 and B2, their templates are from different PDB structures. Therefore, PRISM was used to search the complex structures sharing high interface residue similarity with B1 and B2. Here, one complex (PDB entry: 1lqb) is selected and the complex structure of B1–B2 can be modeled by PRISM, adopting the binding mode of 11qb (represented by *gray*). For C1–C2 and D1–D2 whose homologous templates are not available, the known DDI information (PF00428 and PF0046) of C1–C2 is annotated from the iPfam database, while the known DMI information (PF00244 and LIG-14-3-3 2) of D1–D2 is annotated from the ELM database

complex with the largest interaction interface was considered. We employed TM-align (Zhang and Skolnick 2005) to align the separated structures of two monomers to the template complex structure. Thus, the 3D structures of protein complexes for the first category of PPIs were obtained. Similar to the processing of experimental complex structures, a distance cutoff of 4 Å between any atom pair was also used to identify interacting residues based on the complex structures.

For the second category of PPIs, we built the structural models for individual interacting proteins as described above, but constructed the complex structures by using the PRISM software. It was worth mentioning that the templates used in PRISM might be different from those exploited to build the structures of individual interacting proteins. By default, PRISM automatically selected the template for the protein complex and adopts the template's binding mode for two interacting proteins in order to build the protein complex structure. We considered the complex structure with minimum energy as the 3D structure of the protein complex for the second category of PPIs, and the corresponding interacting residues were further calculated.

Quantifying the physicochemical and evolutionary properties of protein interaction sites

We quantified the physicochemical and evolutionary properties of protein interaction sites, including the bond types, the ΔΔG upon alanine mutation and the residue conservation. Ionic bond, hydrogen bond and van der Waals' interaction were defined in accordance with iPfam (Finn et al. 2014b). Cation–π interactions were identified if the distance between the cationic group of Lys or Arg and the aromatic ring center of Phe, Tyr or Trp was less than 6 Å (Gallivan

and Dougherty 1999). Computational alanine scanning (Kortemme et al. 2004) used a free energy function consisting of a linear combination of multiple models to predict the ΔΔG. The positive ΔΔG value means that the interaction between two proteins is destabilized when an interacting residue is mutated to alanine. In contrast, the negative ΔΔG value means the mutation of interacting residue has a stabilizing effect on the PPI. To capture the conservation of interaction sites, we extracted orthologs from the OMA orthology database (Altenhoff et al. 2015). Further, Rate4Site (Pupko et al. 2002) was employed to estimate the residue conservation based on the protein multiple sequence alignment generated by the Clustal Omega software (Sievers et al. 2011). The lower conservation score means the higher residue conservation.

Annotating DDIs and DMIs

For those PPIs that did not satisfy the criteria of 3D structure modeling, the presence of DDIs and DMIs was assigned. First, the protein domain information was taken from the TAIR database (Lamesch et al. 2012). Then, the motifs in the ELM database (Dinkel et al. 2014) were searched against the protein sequences using PatMatch (Yan et al. 2005). Inter-chain DDIs inferred from the PDB database were downloaded from the iPfam database and experimentally validated DMIs were downloaded from the ELM database. Using these data, the information of DDIs and DMIs in the PPIs could be annotated.

Computing I_RMS values

To evaluate the accuracy of predicted complex structures, we calculated I_RMS values. We first superimposed the predicted complex structure onto the corresponding experimental complex structure using MM-align, which is a structural alignment algorithm for comparing protein complex structures (Mukherjee and Zhang 2009). Then, we extracted interface residues from the experimental complex structure and selected their equivalent residues in the predicted complex structure. Note that, in accordance with the criteria of CAPRI, here the interface residues refer to those having at least one pairwise atom distance ≤10 Å (Mendez et al. 2003). The value of I_RMS between the predicted and experimental complex structure was computed as:

$$I\_RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

where N is the number of interface residues in the experimental complex structure and δ_i is the C_α atom distance between the ith interface residue and its equivalent in the predicted complex structure.

Implementing the web server

The web server was implemented on a Linux operating system with CentOS-6.2, Apache 2.2.15 and MySQL 5.1.52. For the 3D structure visualization, we employed 3Dmol.js (Rego and Koes 2015), which provides an interactive and hardware-accelerated 3D representation without any configuration on local machines. In addition, Sigmajs Exporter plugin of Gephi (https://marketplace.gephi.org/plugin/sigmajs-exporter/) was used to display the networks.

Acknowledgments We are also grateful to Xin Yi from Prof. Zhen Su's Lab and Dr. Xiaobao Dong for valuable comments on the construction of the database.

Funding This work was supported by grants from the National Natural Science Foundation of China (31471249 and 31271414).

Author contributions H.L. performed the analyses and drafted the manuscript. S.Y. constructed the database. Y.Z. and Z.Z. supervised the study. C.W., Y.Z. and Z.Z. revised the manuscript. H.L., C.W., Y.Z. and Z.Z. provided suggestions for the database construction.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Altenhoff AM et al (2015) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. Nucleic Acids Res 43:D240–D249. doi:10.1093/nar/gku1158

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410. doi:10.1016/S0022-2836(05)80360-2

Arabidopsis Interactome Mapping Consortium (2011) Evidence for network evolution in an Arabidopsis interactome map. Science 333:601–607. doi:10.1126/science.1203877

Baspinar A, Cukuroglu E, Nussinov R, Keskin O, Gursoy A (2014) PRISM: a web server and repository for prediction of protein–protein interactions and modeling their 3D complexes. Nucleic Acids Res 42:W285–W289. doi:10.1093/nar/gku397

Braun P, Aubourg S, Van Leene J, De Jaeger G, Lurin C (2013) Plant protein interactomes. Annu Rev Plant Biol 64:161–187. doi:10.1146/annurev-arplant-050312-120140

Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES (2004) Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci 13:190–202. doi:10.1110/ps.03323604

Chatr-Aryamontri A et al (2013) The BioGRID interaction database: 2013 update. Nucleic Acids Res 41:D816–D823. doi:10.1093/nar/gks1158

Dinkel H et al (2014) The eukaryotic linear motif resource ELM: 10 years and counting. Nucleic Acids Res 42:D259–D266. doi:10.1093/nar/gkt1047

Finn RD et al (2014a) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230. doi:10.1093/nar/gkt1223

Finn RD, Miller BL, Clements J, Bateman A (2014b) iPfam: a database of protein family and domain interactions found in the Protein

Data Bank. Nucleic Acids Res 42:D364–D373. doi:10.1093/nar/gkt1210
Fukuhara N, Kawabata T (2008) HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures. Nucleic Acids Res 36:W185–W189. doi:10.1093/nar/gkn218
Gallivan JP, Dougherty DA (1999) Cation–pi interactions in structural biology. Proc Natl Acad Sci USA 96:9459–9464
Gu H, Zhu P, Jiao Y, Meng Y, Chen M (2011) PRIN: a predicted rice interactome network. BMC Bioinform 12:161. doi:10.1186/1471-2105-12-161
Hashimoto K et al (2012) Phosphorylation of calcineurin B-like (CBL) calcium sensor proteins by their CBL-interacting protein kinases (CIPKs) is required for full activity of CBL-CIPK complexes toward their target proteins. J Biol Chem 287:7956–7968. doi:10.1074/jbc.M111.279331
He F, Zhou Y, Zhang Z (2010) Deciphering the Arabidopsis floral transition process by integrating a protein–protein interaction network and gene expression data. Plant Physiol 153:1492–1505. doi:10.1104/pp.110.153650
Higurashi M, Ishida T, Kinoshita K (2009) PiSite: a database of protein interaction sites using multiple binding states in the PDB. Nucleic Acids Res 37:D360–D364. doi:10.1093/nar/gkn659
Kortemme T, Kim DE, Baker D (2004) Computational alanine scanning of protein–protein interfaces. Sci STKE 2004:pl2. doi:10.1126/stke.2192004pl2
Krissinel E (2010) Crystal contacts as nature's docking solutions. J Comput Chem 31:133–143. doi:10.1002/jcc.21303
Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372:774–797. doi:10.1016/j.jmb.2007.05.022
Kundrotas PJ, Vakser IA (2010) Accuracy of protein–protein binding sites in high-throughput template-based modeling. PLoS Comput Biol 6:e1000727. doi:10.1371/journal.pcbi.1000727
Kundrotas PJ, Zhu Z, Janin J, Vakser IA (2012) Templates are available to model nearly all complexes of structurally characterized proteins. Proc Natl Acad Sci USA 109:9438–9441. doi:10.1073/pnas.1200678109
Lamesch P et al (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40:D1202–D1210. doi:10.1093/nar/gkr1090
Li H, Zhou Y, Zhang Z (2015) Competition–cooperation relationship networks characterize the competition and cooperation between proteins. Sci Rep 5:11619. doi:10.1038/srep11619
Lin M, Zhou X, Shen X, Mao C, Chen X (2011) The predicted Arabidopsis interactome resource and network topology-based systems biology analyses. Plant Cell 23:911–922. doi:10.1105/tpc.110.082529
Mendez R, Leplae R, De Maria L, Wodak SJ (2003) Assessment of blind predictions of protein–protein interactions: current status of docking methods. Proteins 52:51–67. doi:10.1002/prot.10393
Morrow JK, Zhang S (2012) Computational prediction of protein hot spot residues. Curr Pharm Des 18:1255–1265
Mosca R, Ceol A, Aloy P (2013) Interactome3D: adding structural details to protein networks. Nat Methods 10:47–53. doi:10.1038/nmeth.2289
Mosca R, Ceol A, Stein A, Olivella R, Aloy P (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res 42:D374–D379. doi:10.1093/nar/gkt887
Mukherjee S, Zhang Y (2009) MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. Nucleic Acids Res 37:e83. doi:10.1093/nar/gkp318
Orchard S et al (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 42:D358–D363. doi:10.1093/nar/gkt1115
Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18(Suppl 1):S71–S77
Rego N, Koes D (2015) 3Dmol.js: molecular visualization with WebGL. Bioinformatics 31:1322–1324. doi:10.1093/bioinformatics/btu829
Rose PW et al (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res 43:D345–D356. doi:10.1093/nar/gku1214
Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815. doi:10.1006/jmbi.1993.1626
Sievers F et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. doi:10.1038/msb.2011.75
Sinha R, Kundrotas PJ, Vakser IA (2010) Docking by structural similarity at protein–protein interfaces. Proteins 78:3235–3241. doi:10.1002/prot.22812
Tuncbag N, Gursoy A, Nussinov R, Keskin O (2011) Predicting protein–protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. Nat Protoc 6:1341–1354. doi:10.1038/nprot.2011.367
UniProt Consortium (2015) UniProt: a hub for protein information. Nucleic Acids Res 43:D204–D212. doi:10.1093/nar/gku989
Westermarck J, Ivaska J, Corthals GL (2013) Identification of protein interactions involved in cellular signaling. Mol Cell Proteomics 12:1752–1763. doi:10.1074/mcp.R113.027771
Yan T et al (2005) PatMatch: a program for finding patterns in peptide and nucleotide sequences. Nucleic Acids Res 33:W262–W266. doi:10.1093/nar/gki368
Yao Q et al (2014) P(3)DB 3.0: From plant phosphorylation sites to protein networks. Nucleic Acids Res 42:D1206–D1213. doi:10.1093/nar/gkt1135
Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33:2302–2309. doi:10.1093/nar/gki524