Maize (Zea mays) is one of the most important crops worldwide. To understand the biological processes underlying various traits of the crop (e.g., yield and response to stress), a detailed protein-protein interaction (PPI) network is highly demanded. Unfortunately, there are very few such PPIs available in the literature. Therefore, in this work, we present the Protein-Protein Interaction Database for Maize (PPIM), which covers 2,762,560 interactions among 14,000 proteins. The PPIM contains not only accurately predicted PPIs but also those molecular interactions collected from the literature. The database is freely available at http://comp-sysbio.org/ppim with a user-friendly powerful interface. We believe that the PPIM resource can help biologists better understand the maize crop.

Maize (Zea mays) is one of the most important crops in the world. Understanding the molecular mechanisms underlying various traits of maize (e.g., response to drought and salinity) is important to improve the quality and yield of the crop. Although the maize genome sequence has unraveled the gene components of the crop, most traits involve complex interactions among molecules. Some protein-protein interactions (PPIs) have been experimentally determined in maize. For example, the CENTORADIALIS8 protein was found to interact with the floral activator DLF1 protein with yeast two-hybrid assays (Danilevskaya et al., 2008), and barren stalk1 was found to interact with barren inflorescence2 with pulldown assays (Skirpan et al., 2008). Unfortunately, unlike other model organisms, there are very few molecular interactions available for maize. Therefore, a comprehensive maize interactome map is highly demanded.

Recently, with more information about maize available, it has become practical to investigate the interactions between maize molecules. For example, with accumulating gene expression data, a gene coexpression network has been built to identify gene modules that play important roles in conditions of interest. With this idea, Downs et al. (2013) constructed a gene coexpression network based on gene expression data from 50 maize tissues and identified some gene modules that are important for development. By comparing the maize and rice (Oryza sativa) coexpression networks, Ficklin and Feltus (2011) identified some conserved gene modules between the two species, indicating their essential roles in crops. With protein abundance and phosphorylation data in different maize tissues across seven developmental stages, Walley et al. (2013) built a protein coexpression network to present kinase-substrate relationships. The metabolic network MaizeCyc (Monaco et al., 2013), containing enzyme catalysts, proteins, and other metabolites, has also been constructed. Focusing on maize kernel development, the expression quantitative trait loci have been investigated with RNA sequencing data (Fu et al., 2013), and the gene regulations underlying endosperm cell differentiation have been identified (Zhan et al., 2015).

Despite the above efforts to identify possible interactions between molecules, no comprehensive interactome is available for maize. Most current approaches construct gene coexpression networks; however, these only describe the associations between genes and
cannot tell which genes have real interactions. Under these circumstances, we present a comprehensive Protein-Protein Interaction Database for Maize (PPIM), which provides both our predicted physical and functional interactions as well as molecular interactions collected from the literature and public databases. To our knowledge, the PPIM is the most comprehensive database for maize to date. The user-friendly powerful interface accompanying the database can help biologists better explore the database.

Figure 1. Contributions of the nine organisms to predicted physical PPIs for maize. A, Distribution of maize orthologous proteins in the nine organisms. B, Distribution of predicted physical interactions supported by distinct organisms.
RESULTS AND DISCUSSION

A Maize PPI Network

Based on the physical PPIs collected from nine model organisms (Fig. 1), there were 19,134 PPIs among 3,706 proteins that were predicted by the interolog approach. Figure 1 shows the contribution of distinct organisms to predicted maize PPIs. It is not surprising that Arabidopsis (Arabidopsis thaliana), as the model organism of plants, has more orthologous proteins with maize than other organisms (Fig. 1A). As the most well-studied species, Saccharomyces cerevisiae and Homo sapiens have the most complete interactomes and, therefore, contribute most to the predicted PPIs (Fig. 1B).

To predict the functional PPIs, we trained a support vector machine (SVM) model based on the integration of six distinct types of data sets as described in “Materials and Methods.” A 10-fold cross-validation was utilized to evaluate the SVM model, where the 10-fold cross-validation was repeated 100 times and the same number of negative samples as positive samples were randomly selected to train the SVM model each time. Table I shows the performance of the trained model over the benchmark data set, where the average of the 10-fold cross-validation was used as the final output of the model. It can be seen that our trained model has high precision, indicating the reliability of our predicted PPIs. With the trained model, we finally predicted 2,734,000 functional interactions among 10,793 proteins. In addition, the decision score provided by the SVM model was used as the confidence score for each predicted PPI.

We put the predicted physical and functional interactions together and constructed a complete interactome map for maize. The interactome consisting of 2,752,298 PPIs and 12,691 proteins was deposited in the PPIM, which is freely accessible at http://comp-sysbio.org/ppim. According to their confidence scores, the interactions were further grouped into three categories: low-confidence, medium-confidence, and high-confidence interactions. The physical PPIs as well as the functional PPIs with the top 5% highest decision scores were regarded as the high-confidence interactions, while those functional PPIs with decision scores between the top 50% and the top 5% were treated as medium-confidence interactions; the rest of the functional PPIs were considered as low-confidence interactions. As a result, the predicted interactome consists of 155,845 high-confidence PPIs, 1,229,771 medium-confidence PPIs, and 1,366,682 low-confidence PPIs. Recently, a kinase-substrate interaction network was constructed based on the protein expression profiles (Walley et al., 2013), and a gene regulatory network consisting of interactions between transcription factors (TFs) and genes was inferred for maize leaf development (Yu et al., 2015). Both data sets provide possible functional interactions between proteins/genes and, therefore, were deposited in the PPIM. In addition, we collected experimentally determined PPIs from public databases such as UniProt (2015.10; UniProt Consortium, 2015), BioGrid (2015.9; Stark et al., 2006), DIP (2015.7; Xenarios et al., 2001), IntAct (2015.10; Orchard et al., 2014), and MINT (2012.10; Licata et al., 2012) and obtained 28 PPIs. To further extend the experimentally determined collection, we performed text mining to the journal abstracts retrieved from Medline and extracted 74 pairs of proteins that have been clearly stated to have interactions. As a result, 98 experimentally determined PPIs were identified from both public databases and the literature. In summary, the maize interactome deposited in the PPIM comprises 2,762,560 interactions among 14,000 proteins. Table II summarizes the PPIs deposited in the PPIM.

Distinct Proteins Obtain Specific Functions by Interacting with Different Partners

The maize genome has a unique architecture with large amounts of duplicated genes due to the whole-genome duplication 5 million years ago (Haberer et al., 2005). Although duplicated genes tend to have similar functions (Chen et al., 2013), distinct members of the same gene family should interact with specific partners to obtain their specific functions. To see whether the members from the same family tend to interact with the same partners in maize in our prediction, we investigated the interaction partners of those members belonging to the same family. For this, the inhomogeneity IH of interaction partners of those proteins belonging to the same family F was defined as follows:

\[
IH_F = \frac{|P_1 \cap P_2 \cap \ldots \cap P_n|}{|P_1 \cup P_2 \cup \ldots \cup P_n|}
\]

Table II. Number of molecular interactions deposited in the PPIM and the number of proteins involved

<table>
<thead>
<tr>
<th>Category</th>
<th>Predicted Physical PPIs</th>
<th>Predicted Functional PPIs</th>
<th>Experimentally Determined PPIs</th>
<th>Kinase-Substrate Interactions</th>
<th>TF-Gene Interactions</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of proteins</td>
<td>3,706</td>
<td>10,793</td>
<td>112</td>
<td>771</td>
<td>2,293</td>
<td>14,000</td>
</tr>
<tr>
<td>No. of interactions</td>
<td>19,134</td>
<td>2,734,000</td>
<td>98</td>
<td>1,130</td>
<td>10,091</td>
<td>2,762,560</td>
</tr>
</tbody>
</table>
interacting and union of the two sets, and \(|\cdot|\) denotes the number of elements in the set. Note that the interactions between members from the same family \(F\) were excluded from consideration. Figure 2 shows the distribution of the inhomogeneity for maize proteins belonging to families of different size, where the proteins with high sequence similarity (\(E\) value below \(1e^{-10}\) and identity above 80 with BLAST; Altschul et al., 1990) were assigned to the same family. From the results, we can see that the more genes a family has the more inhomogeneous its members’ interaction partners will be, implying that the members from large gene families have specific functions by interacting with distinct interaction partners.

Validation of Maize PPIs

Interacting Proteins Tend to Have Similar Functions

In the literature, it has been observed that interacting proteins tend to have similar functions. Therefore, we investigated the functional annotations of interacting protein pairs against those of noninteracting protein pairs. With functional annotations obtained from the agrIGO database (Du et al., 2010), Figure 3A shows the proportion of protein pairs that have at least one same functional term for both interacting and noninteracting protein pairs, where the three functional categories from Gene Ontology (GO) databases were investigated. It can be seen that the protein pairs predicted to interact tend to have similar functions, indicating the reliability of our predicted PPIs. Considering the fact that the functional annotations about biological process and molecular function have been used to construct the SVM model, it is not surprising to see a high proportion of our predicted PPIs share these two kinds of annotations.

Since the cellular component annotations were not used to train the model, we further used the annotations to validate our predicted PPIs. Given a pair of proteins \((i\text{ and } j)\), their functional similarity was defined as follows:

\[
S_{ij} = \frac{|C_i \cap C_j|}{|C_i \cup C_j|} \tag{2}
\]

where \(C_i\) and \(C_j\) denote the set of cellular component annotations associated with the \(i\)th and \(j\)th genes, respectively, \(\cap\) and \(\cup\) separately indicate the intersection and union of the two sets, and \(|\cdot|\) is the number of elements in the set. Figure 3B shows the distribution of protein pairs according to their functional similarity defined by Equation 2. Obviously, the protein pairs from the PPIIM rather than the noninteracting protein pairs have higher similarity (\(P < 2.2e^{-15}\), one-sided Wilcoxon signed-rank test), which is consistent with the fact that proteins belonging to the same cellular component are more likely to interact with each other.

Validation of Maize PPIs with Pathway Information

Considering that the interacting proteins are more likely to be involved in the same pathway, we used the pathway information from CornCyc of the Plant Metabolic Network to check whether our predicted interacting protein pairs tend to be in the same pathway (Caspi et al., 2012). Since the pathways from CornCyc were inferred with methods different from those for pathways from MaizeCyc that were used to construct our SVM model, the CornCyc pathways are independent and ideal data with which to evaluate our predicted PPIs. There are 6,510 proteins involved in the CornCyc pathways, and these proteins form 103,269 predicted PPIs in the PPIIM, where 22.08% (22,803) of the PPIs belong to at least one common CornCyc pathway. Compared with noninteracting protein pairs, we found that our predicted PPIs tend to be in the same pathway \((P < 2.2e^{-16}\), Fisher’s exact test), which demonstrates that our predicted PPIs are reliable.

In addition to CornCyc, we also explored the pathway information from other species, including AraCyc (version 6.0; Mueller et al., 2003) for Arabidopsis, SorghumCyc (version 1.1) for Sorghum bicolor, and RiceCyc (version 3.3) for rice (Dharmawardhana et al., 2013). The orthologs of maize proteins were first identified in the three organisms. Figure 4 shows the Venn diagram of the predicted PPIs for the PPIIM, AraCyc, SorghumCyc, and RiceCyc based on pathway information. From the results, we can see that a lot of our predicted PPIs can be validated with pathway information in other species, implying the high accuracy of our predicted PPIs.

Validation of Maize PPIs with Interactions from the Literature

By investigating those 98 experimentally determined PPIs, 56 pairs with corresponding proteins were found in our predicted interactome; the others cannot be
described by the six types of information considered here. Among the 56 experimentally determined PPIs, 22 (39.29%) PPIs can be found in our predicted interactome, indicating the predictive power of our approach. In the kinase-substrate network, there are 1,130 interactions involving 771 proteins, out of which 61 (5.4%) interactions can be found in the PPIM. Among the 10,091 interactions between 254 TFs and

Figure 3. A, Proportions of interacting and noninteracting protein pairs that have the same functional terms according to three categories of biological processes, molecular functions, or cellular components. B, Distribution of protein pairs according to their functional similarities for interacting and noninteracting protein pairs.
2,114 genes, 421 (4.2%) interactions can be found in the PPIM.

Furthermore, we also validated our predicted PPIs with those predicted by other approaches and PPIs from other species. The STRING database provides predicted PPIs for various species (Jensen et al., 2009), where there are 12,066,705 interactions predicted for maize. With the CornCyc pathway as the gold standard, we noticed that only 3,003 (0.7%) of the 427,999 interactions from STRING for those proteins from CornCyc were found to be in at least one same pathway. Compared with the significant enrichment of PPIs from the PPIM (22.08%) in the CornCyc pathway, we can say that our predicted PPIs are more reliable.

In addition to direct validation of our interactome with predicted maize PPIs from the literature, the established interactions from other species were also utilized to validate our predicted maize PPIs. Here, we considered two plant model organisms, Arabidopsis and rice. The PPIs for Arabidopsis were obtained from AraNet (http://www.inetbio.org/aranet/), which is currently the most comprehensive Arabidopsis interactome (Lee et al., 2015). The PPIs from rice were obtained from our previous work (http://comp-sysbio.org/dipos/; Sapkota et al., 2011). By investigating their interologs in Arabidopsis and rice, 0.56% and 12.46% of the predicted interactions from the PPIM, respectively, were validated in the two species. Table III summarizes the predicted interactions deposited in the PPIM that can be validated with information from various sources, where 689,204 (25.04%) PPIs from the PPIM can be validated, indicating the reliability of our predicted PPIs.

### Usage of the PPIM

All our predicted PPIs, including both physical and functional ones, along with those from public databases and the literature (e.g. the kinase-substrate interactions; Walley et al., 2013) were deposited in the PPIM. Furthermore, the pathway and GO annotations about maize proteins were also collected from public databases, which can help us better understand the functions of maize proteins. In addition, the cross-links to other databases (e.g. MaizeCyc) were also provided in the PPIM.

Except for the rich content in the PPIM, a powerful and user-friendly interface was provided. For example, a network consisting of query proteins and their interaction partners as well as the interactions among them

---

**Table III. Number of predicted PPIs that can be validated with information from other sources**

<table>
<thead>
<tr>
<th>Data Source</th>
<th>No. of PPIs</th>
<th>Organism</th>
<th>Overlap with the PPIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimentally determined PPIs</td>
<td>98</td>
<td>Maize</td>
<td>22</td>
</tr>
<tr>
<td>TF-target interactions</td>
<td>10,091</td>
<td>Maize</td>
<td>383</td>
</tr>
<tr>
<td>Kinase-substrate interactions</td>
<td>1,130</td>
<td>Maize</td>
<td>61</td>
</tr>
<tr>
<td>STRING</td>
<td>12,066,705</td>
<td>Maize</td>
<td>412,546</td>
</tr>
<tr>
<td>AraNet</td>
<td>244,220</td>
<td>Arabidopsis</td>
<td>15,395</td>
</tr>
<tr>
<td>DIPOS</td>
<td>886,440</td>
<td>Rice</td>
<td>342,807</td>
</tr>
<tr>
<td>Total</td>
<td>127,930,18</td>
<td>align</td>
<td>689,204 (25.04% of the PPIM)</td>
</tr>
</tbody>
</table>
will be shown on the Web page, where the visualization was accomplished with Cytoscape plugins. Given a set of maize genes, their enriched pathways from MaizeCyc and functions from GO will also be shown. The PPIM can be freely accessed at http://comp-sysbio.org/ppim.

### CONCLUSION

Maize is one of the most important crops in the world. To help us better understand the molecular mechanisms that underlie various traits of maize, we predicted the physical and functional maize PPIs to get a complete interactome map for maize. We hereby present the PPIM, which contains 2,762,560 PPIs covering 14,000 maize proteins. To our knowledge, the PPIM is the most comprehensive PPI database to date. The powerful and friendly interface of the PPIM can help biologists to better utilize the database. We believe that the PPIM resource can help biologists better understand and breed the maize crop.

### MATERIALS AND METHODS

#### Gene Coexpression Similarity

The gene expression profiles across 50 tissues of maize (Zea mays), including embryo, anther, cob, ear, endosperm, husk, leaf, ovule, pericarp, root, silk, stalk, and tassel, at multiple stages of development were retrieved from the Gene Expression Omnibus database with the accession number GSE44743 (Ficklin and Feltus, 2011). The data were preprocessed with the robust multichip average method by utilizing the BioConductor package (version 2.6) of R (version 14.0; Gentleman et al., 2004; R Development Core Team, 2010). The Pearson average method by utilizing the BioConductor package (version 2.6) of R (version 14.0; Gentleman et al., 2004; R Development Core Team, 2010). The Pearson correlation coefficient between each pair of genes was calculated based on their expression profiles, and the correlation coefficient was further transformed into a value between 0 and 1 as follows:

\[
\text{SCE} = \frac{(\text{CE} + 1)}{2}
\]

where CE denotes the correlation coefficient and SCE is the scaled correlation coefficient. The SCE was used as the coexpression similarity of a pair of genes.

#### Functional Similarity

The functional annotations for maize proteins were obtained from the agriGO database. Specifically, we only considered the two functional categories of biological process and molecular function annotations due to the scarce information for molecular components. To see the relationships between distinct functional terms, the hierarchical structure of the terms was downloaded from the Gene Ontology Consortium (Ashburner et al., 2000). Given a pair of terms \(i\) and \(j\) of a certain functional category, their similarity \(\text{TS}_{ij}\) was defined as follows:

\[
\text{TS}_{ij} = \frac{d_i + d_j}{d_i + d_j + 2}
\]

where \(d_i\) denotes the distance between the nearest common parent term of the term \(i\) and the root in the functional hierarchical tree, \(d_j\) denotes the distance between term \(j\) and the root, and the same is true for \(d_j\).

Since each protein may be annotated with multiple functional terms, the functional similarity \(FS\) between the protein pair \(p\) and \(p'\) was defined as below:

\[
\text{FS} = \max_{\{pGOp, p'GOp\}} \text{TS}_{ij}
\]

where GO\(_p\) and GO\(_{p'}\) represent the set of GO terms for proteins \(p\) and \(p'\), respectively.

#### Domain Interaction

The amino acid sequences for 38,914 maize proteins were downloaded from the maizeGDB database (http://www.maizegdb.org; Lawrence et al., 2008). The domains of maize proteins were annotated by querying sequences against the Pfam database (Coggill et al., 2008). Since the PPIs were accomplished with domain-domain interactions (DDIs; Zhao et al., 2010), we used DDIs to predict PPIs. The DDIs were downloaded from the DOMINE database (Raghavachari et al., 2008), and only the high-confidence DDIs were considered here. Given a pair of proteins \(i\) and \(j\), the probability of the protein pair interacting was defined as below:

\[
\text{PD} = \frac{\text{SI}_{ij}}{D_i D_j}
\]

where \(D_i\) and \(D_j\) represent the number of domains within protein \(i\) and protein \(j\), respectively, and \(\text{SI}_{ij}\) denotes the number of interacting domain pairs between protein \(i\) and protein \(j\).

#### Sequence Similarity

With the amino acid sequences for maize proteins, the sequence similarity \(SS\) between protein \(i\) and protein \(j\) was defined as follows.

### Table IV. The 18 species used to construct the phylogenetic profile

<table>
<thead>
<tr>
<th>Species</th>
<th>Taxonomy Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sorghum bicolor</td>
<td>4558</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>29760</td>
</tr>
<tr>
<td>Oryza sativa ssp. japonica</td>
<td>39947</td>
</tr>
<tr>
<td>Triticum aestivum</td>
<td>4565</td>
</tr>
<tr>
<td>Brachypodium distachyon</td>
<td>15368</td>
</tr>
<tr>
<td>Hordeum vulgare</td>
<td>4513</td>
</tr>
<tr>
<td>Glycine max</td>
<td>3847</td>
</tr>
<tr>
<td>Populus spp.</td>
<td>3689</td>
</tr>
<tr>
<td>Medicago truncatula</td>
<td>3880</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>3702</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>6239</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>7227</td>
</tr>
<tr>
<td>Escherichia coli (strain K12)</td>
<td>83333</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>9606</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>10090</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>10116</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>559292</td>
</tr>
<tr>
<td>Schizosaccharomyces pombe (strain 972/ATCC 24843)</td>
<td>284812</td>
</tr>
</tbody>
</table>

### Table V. The six kinds of information used to describe a pair of proteins

<table>
<thead>
<tr>
<th>Category</th>
<th>Gene Expression</th>
<th>Domain Information</th>
<th>Biological Process</th>
<th>Molecular Function</th>
<th>Amino Acid Sequence</th>
<th>Evolutionary Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of proteins annotated</td>
<td>30,574</td>
<td>17,050</td>
<td>14,244</td>
<td>18,428</td>
<td>38,914</td>
<td>38,914</td>
</tr>
</tbody>
</table>
Evolutionary Similarity

Since PPIs are known to be conserved across species, the proteins that are conserved in multiple species are more likely to interact with each other. In this work, we investigated the phylogenetic profiles of maize proteins across 18 organisms. Table IV shows the 18 organisms and their corresponding taxonomy identifiers provided by the National Center for Biotechnology Information taxonomy database (Benson et al., 2009; Sayers et al., 2009).

The homologous proteins of those maize proteins were identified in the 18 species using BLAST with E values less than $e^{-15}$ and sequence identity above 40%. In this way, an evolution profile vector was constructed for each maize protein, and each element denotes whether the protein has a homologous protein in the corresponding species, where the element is 1 if the maize protein has a homologous protein in the corresponding species and 0 otherwise. With the evolution profile vectors, the evolutionary similarity $ES$ between protein $i$ and protein $j$ was defined as below:

$$ES = 1 - \frac{H_L}{L}$$

where $H_L$ denotes the Hamming distance between protein $i$ and protein $j$ and $L$ is the length of the protein vectors (i.e. 18).

Gold Standard Functional Interactions

Since no gold standard functional interactions are available for maize, we used the pathway information from the MaizeCyc database (version 2.2) as the benchmark data. In particular, a pair of neighbor proteins within the same pathway was regarded to functionally interact with each other. Finally, 53,978 functional interactions were obtained and used as positive samples. A similar number of protein pairs out of all possible protein pairs, except the positive samples, were used as negative samples.

Prediction of Functional Interactions

With the gold standard functional interactions obtained above, we aimed to predict new functional interactions by integrating the six distinct kinds of data described above (Table V), where only the 11,068 proteins with all six types of information were considered. To predict the functional interactions, we utilized the SVM here due to its good performance. The LIBSVM toolbox was employed to train the SVM model (Chang and Lin, 2011), where the Gaussian kernel was adopted and the parameters were optimized with 10-fold cross-validation.

Prediction of Physical Interactions

To predict the physical interactions among maize proteins, we used the interolog approach as described in our previous work (Zhao et al., 2009). In particular, the physical PPIs for nine model organisms (i.e. Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Escherichia coli, Homo sapiens, Mus musculus, Rattus norvegicus, Saccharomyces cerevisiae, and Schizosaccharomyces pombe) were collected from seven commonly used sources: BioGrid (2013.9), DIP (2013.7), IntAct (2013.8), MINT (2012.10), HPRD (2013.10; Keshava Prasad et al., 2009), The Arabidopsis Information Resource (2009.5; Garcia-Hernandez et al., 2002), and MPIDB (2009.11; Goll et al., 2008). The interactions for Arabidopsis from two systematic experiments were also collected (Braun et al., 2011; Mukhtar et al., 2011).

With the above interactions from model organisms, a pair of proteins were regarded to interact with each other if their corresponding orthologous proteins interact within any of the nine organisms. The orthologs of maize proteins in other organisms were identified with reciprocal best hits through BLAST with $E$ value and identity cutoffs of $e^{-15}$ and 40%. Specifically, given maize protein $i$ and its ortholog in the 4th organism, a normalized confidence score $C^i_k$ about the orthologous relationship was defined as follows:

$$SS = \frac{BS_i}{\sqrt{BS_i BS_j}}$$

where $BS_i$ denotes the bit score describing the sequence similarity between protein $i$ and protein $j$ and the bit score is the output of BLAST.

$$C^i_k = \frac{BS_{ik} BS_{jk}}{BS_{ik} BS_{j}}$$

where $BS_{ik}$ is the bit score obtained when aligning maize protein $i$ against its ortholog in the 4th organism, and the opposite is true for $BS_{ik}$. In addition, a confidence score $CS_{ij}$ was defined for each predicted interaction pair ($i$ and $j$) as described in our previous work (Sapkota et al., 2011):

$$CS_{ij} = \frac{\sum_{k=1}^{n} C^i_k C^j_k}{n}$$

where $n_i$ is the number of databases supporting the predicted interaction and $n$ is the number of reference organisms (i.e. 9).

Collection of Experimentally Determined PPIs from the Literature

To extract experimentally determined PPIs from published articles, text mining was performed on the journal abstracts retrieved from Medline. A pair of proteins were regarded to possibly interact if the proteins of interest cooccur in one abstract together with at least one interaction term described in controlled vocabularies of interaction types from the HUPO Proteomics Standards Initiative (http://www.proteomecommons.org; Mayer et al., 2013) and the organism name maize or Zea mays occurs in the abstract. The resultant candidates were further manually curated to keep only those clearly described functional interactions.

LITERATURE CITED


