Research Article

Analysis of Protein–Protein Functional Associations by Using Gene Ontology and KEGG Pathway

Fei Yuan,1 Xiaoyong Pan,2 Lei Chen,3,4 Yu-Hang Zhang,5 Tao Huang,5 and Yu-Dong Cai6

1Department of Science & Technology, Binzhou Medical University Hospital, Binzhou 256603, Shandong, China
2BASF & IDLab, Ghent University, Ghent, Belgium
3College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
4Shanghai Key Laboratory of PMMP, East China Normal University, Shanghai 200241, China
5Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
6School of Life Sciences, Shanghai University, Shanghai 200444, China

Correspondence should be addressed to Fei Yuan; snowhawkyrf@outlook.com and Yu-Dong Cai; cai.yud@126.com

Received 4 January 2019; Revised 4 June 2019; Accepted 26 June 2019; Published 18 July 2019

Academic Editor: Hesham H. Ali

Copyright © 2019 Fei Yuan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Protein–protein interaction (PPI) plays an extremely remarkable role in the growth, reproduction, and metabolism of all lives. A thorough investigation of PPI can uncover the mechanism of how proteins express their functions. In this study, we used gene ontology (GO) terms and biological pathways to study an extended version of PPI (protein–protein functional associations) and subsequently identify some essential GO terms and pathways that can indicate the difference between two proteins with and without functional associations. The protein–protein functional associations validated by experiments were retrieved from STRING, a well-known database on collected associations between proteins from multiple sources, and they were termed as positive samples. The negative samples were constructed by randomly pairing two proteins. Each sample was represented by several features based on GO and KEGG pathway information of two proteins. Then, the mutual information was adopted to evaluate the importance of all features and some important ones could be accessed, from which a number of essential GO terms or KEGG pathways were identified. The final analysis of some important GO terms and one KEGG pathway can partly uncover the difference between proteins with and without functional associations.

1. Introduction

Protein is the material foundation of all living things [1]. Protein–protein interaction (PPI) plays an extremely significant role in the growth, reproduction, and metabolism of any life, even in a single cell [2, 3]. Proteins can be easily clustered in three methods: (I) homology of protein subunits, (II) stability of interactions, and (III) combination mode of subunits [4–6]. By connecting related proteins, PPI initiates the action of various functional or structural proteins in every single cell [4]. Given that proteins influence different biological processes, even in single cells, conducting a study on PPI to further determine protein functions and life activities is a relevant endeavor.

PPI has been thoroughly studied both in experimental and computing scenarios. To study PPI via experiments, coimmunoprecipitation, Western blot, and yeast two-hybrid systems are generally adopted [7, 8]. As for computational methods, several algorithms have been developed to identify PPI, and the two main ones are the topology-free approaches and the graph-based approaches, which are based on distances between proteins and specialized clustering techniques, respectively [9, 10]. Some other computational methods predict PPIs from protein sequences using machine learning. Jansen et al. developed a Bayesian network to integrate multiple genomic features to predict PPIs [11]. Shen et al. trained a support vector machine classifier using conjoint triad features derived from sequences [12]. Pan et al. first used latent Dirichlet allocation model to extract latent topic features from the conjoint triad features, then the learned topic features were fed into a random forest classifier to predict PPIs [13]. Hashemifar et al. trained a deep learning model
to predict PPIs using evolutionary information with random projection and data augmentation [14]. In addition, with the development and innovation of computational technologies, the use of updated algorithms has allowed researchers to predict and study PPIs conveniently and accurately alongside the utilization of different databases and methods.

Gene ontology (GO) is a bioinformatic concept that was originally proposed to unify the representation of genes and gene products of many species [15, 16]. The ontology covers three main domains, namely, (I) cellular component, (II) molecular function, and (III) biological process, which can easily cluster all genes and gene products with a directed acyclic graph (DAG) [16]. For convenience, the ontological domains are widely used in computational biology to avoid redundancy of different annotations of a single functional or structural gene [17, 18]. GO terms, which have been updated given the development of biological science, can summarize the specific role of genes and their products in living cells, and they are regarded as powerful tools in computational biology science [16]. Different kinds of PPIs are also included in the various terms of GO annotations. The specific locations or functions of PPIs in cells have been investigated to easily describe and distinguish the several kinds of GO terms. The GO annotations contain informative signals for PPIs. For example, Patil and Nakamura trained a machine learning classifier to infer PPIs using features derived from sequence similarity, shared GO terms and domains [19]. Ben-Hur et al. used a kernel method to integrate sequences, GO annotations, local network properties and homologous interactions for predicting PPIs [20]. Stefan et al. generated features for proteins from GO DAG; then, the extracted features were fed into a random forest classifier to predict PPIs [21]. However, these studies only adopted the GO annotations to construct the model for predicting PPIs. They did not analyze which GO annotations were highly related to the determination of PPIs. In addition, genes can be clustered into several biological pathways. Some essential pathways may be highly related to PPIs.

In this study, we investigated an extended version of PPI (protein–protein function associations) by using GO terms and KEGG pathways. Considering the fact that few PPI studies with computational methods investigated which GO terms were highly related to the determination of PPIs, the purpose of this study was to identify key GO terms or KEGG pathways that can indicate the difference between two proteins with and without functional associations. We first extracted protein–protein functional associations with experiment validations reported in Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) [22, 23], a well-known public database on several collected associations between proteins from various organisms. These associations have been derived from the following four sources: (I) genomic context, (II) high-throughput experiments, (III) (conserved) coexpression, and (IV) previous knowledge. To obtain the human protein–protein functional associations in this database, we downloaded a file named “protein.links.detailed.v9.1.txt.gz” and then extracted lines starting with “9606” (i.e., the code of human in STRING). A total of 2,425,314 human protein–protein functional associations involving 20,770 proteins were accessed. The purpose of this study is to identify some important GO terms or KEGG pathways that can indicate the difference between two proteins with and without functional associations. Thus, we refined the 20,770 proteins as follows: (1) utilize CD-HIT [24] to discard similar proteins such that the similarity between any two remaining proteins was less than 0.25 and (2) exclude proteins whose GO term or KEGG pathway information was not available, from which we obtained 8,916 proteins. The derived proteins can comprise 588,154 human protein–protein functional associations. Furthermore, we selected 70,392 human protein–protein functional associations among the above-mentioned associations. The “Experimental” scores of these associations are larger than zero, meaning that they are validated by solid experiments. These associations involved 6,623 human proteins. For convenience, these associations were termed positive associations in this study and are provided in Supplementary Material SI.

To extract the difference between positive associations and any two proteins without functional associations, some negative associations are necessary. Given that negative associations are substantially more than the positive ones, we constructed 211,176 differing pairs of proteins, which were thrice as many as positive associations, and each of them was produced as follows: (1) random selection of two different proteins from 6,623 proteins, and (II) these two proteins cannot comprise an association reported in STRING. The obtained negative and positive associations constituted a dataset. Considering that the produced negative associations may influence the results, we randomly produced 10 sets of negative associations. Each of the sets, together with the positive associations, constituted a dataset, thus producing 10 datasets, which were denoted as $DS_1, DS_2, \ldots, DS_{10}$. By analyzing these datasets, some essential information for protein–protein functional associations can be discovered. The whole procedures are illustrated in Figure 1.
2.2. Representation of Protein–Protein Function Associations.

GO terms [16] and KEGG pathways [25] are always used to elucidate and describe molecular functions, cellular components, and biological and signal processes of genes. From Gene Ontology Consortium [16], 17,916 GO terms were retrieved. Accordingly, a protein \( p \) can be encoded as

\[
v_{\text{GO}}(p) = \left[ g_{p}^{1}, g_{p}^{2}, \ldots, g_{p}^{17916} \right]^T, \tag{1}
\]

where

\[
g_{p}^{i} = \begin{cases} 
1 & \text{If } p \text{ is annotated by the } i\text{-th GO term} \\
0 & \text{Otherwise.} 
\end{cases} \tag{2}
\]

For two proteins \( p_1 \) and \( p_2 \) that comprised either a positive association or a negative association \( P = (p_1, p_2) \), because there was no order information in \( P \), i.e., \((p_1, p_2) \) was identical to \((p_2, p_1) \), it was not appropriate to simply combine the features of \( p_1 \) and \( p_2 \). To exclude the order information of \( P \), we adopted the following scheme that has been used in some studies [26, 27]. For \( P = (p_1, p_2) \), it was encoded into a vector by using \( v_{\text{GO}}(p_1) \) and \( v_{\text{GO}}(p_2) \) as follows:

\[
V_{\text{GO}}(P) = v_{\text{GO}}(p_1) \otimes v_{\text{GO}}(p_2) = \left[ g_{1}^{p_1} + g_{1}^{p_2}, \left| g_{1}^{p_1} - g_{1}^{p_2} \right|, \ldots, g_{17916}^{p_1} + g_{17916}^{p_2}, \left| g_{17916}^{p_1} - g_{17916}^{p_2} \right| \right]^T. \tag{3}
\]

Moreover, according to KEGG [25], there were 279 pathways, based on which the protein \( p \) can be represented by

\[
v_{\text{pathway}}(p) = \left[ k_{1}, k_{2}, \ldots, k_{279} \right]^T, \tag{4}
\]

where

\[
k_{i} = \begin{cases} 
1 & \text{If } p \text{ is annotated by the } i\text{-th KEGG pathway} \\
0 & \text{Otherwise.} 
\end{cases} \tag{5}
\]

Similarly, \( P = (p_1, p_2) \) can be encoded into

\[
V_{\text{pathway}}(P) = v_{\text{pathway}}(p_1) \otimes v_{\text{pathway}}(p_2) = \left[ k_{1}^{p_1} + k_{1}^{p_2}, \left| k_{1}^{p_1} - k_{1}^{p_2} \right|, \ldots, k_{279}^{p_1} + k_{279}^{p_2}, \left| k_{279}^{p_1} - k_{279}^{p_2} \right| \right]^T. \tag{6}
\]
By integrating the GO term and KEGG pathway information of proteins into \( P = (p_1, p_2) \), each association can be finally encoded as

\[
V(P) = V_{GO}(P) \oplus V_{pathway}(P) = \begin{bmatrix} V_{GO}(P) \\ V_{pathway}(P) \end{bmatrix}.
\]

(7)

A total of 36,390 features were used to represent each positive association or negative association. The information of each GO term or KEGG pathway was contained by these two features.

2.3. Feature Evaluation with Mutual Information. As mentioned in Section 2.2, several features were used to represent each protein–protein functional association. However, not all are highly related to sufficiently determine the differences between positive and negative associations, i.e., not all GO terms and KEGG pathways can be used to mark the associations. Here, we adopted the mutual information (MI) of each feature and target (class labels of samples) to evaluate the importance of each feature. The evaluations use the following equation to access the relationship between the two variables of \( x \) and \( y \):

\[
I(x, y) = \int \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy,
\]

(8)

where \( p(x) \) and \( p(y) \) are the marginal probabilistic density of variables \( x \) and \( y \), while \( p(x, y) \) is their joint probabilistic density.

Given a dataset in which each sample is represented by \( N \) features, after the MI values of all features were calculated, features were sorted by their MI values in decreasing order, thereby producing a feature list named MaxRel feature list, which is formulated as

\[
L = [f_1, f_2, \ldots, f_N],
\]

(9)

where \( f_i \) represents a feature in the dataset.

To quickly implement the program of MI, we adopted the program of minimum redundancy maximum relevance (mRMR) method [28], which integrates the MI program. This program has been applied in solving several complicated biological problems [26, 29–45].

3. Results

3.1. Results of the Feature Evaluation. As mentioned in Section 2.2, each association in the 10 datasets was represented by 36,390 features. We calculated the MI value of each feature in each of the datasets \( DS_1, DS_2, \ldots, DS_{10} \). Subsequently, ten MaxRel feature lists could be accessed. A part of these 10 lists is provided in Supplementary Material S2.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of selected features</th>
</tr>
</thead>
<tbody>
<tr>
<td>( DS_1 )</td>
<td>154</td>
</tr>
<tr>
<td>( DS_2 )</td>
<td>154</td>
</tr>
<tr>
<td>( DS_3 )</td>
<td>153</td>
</tr>
<tr>
<td>( DS_4 )</td>
<td>155</td>
</tr>
<tr>
<td>( DS_5 )</td>
<td>149</td>
</tr>
<tr>
<td>( DS_6 )</td>
<td>150</td>
</tr>
<tr>
<td>( DS_7 )</td>
<td>155</td>
</tr>
<tr>
<td>( DS_8 )</td>
<td>152</td>
</tr>
<tr>
<td>( DS_9 )</td>
<td>153</td>
</tr>
<tr>
<td>( DS_{10} )</td>
<td>153</td>
</tr>
</tbody>
</table>

Table 1: Number of selected features in each MaxRel feature list.

![Figure 2: Distribution of 158 selected features: 146, 2, 1, 2, 3, and 4 feature/s in 10, 9, 8, 7, 6, and less than 6 feature sets derived from 10 datasets, respectively.](image)

list, thus producing 10 feature sets denoted as \( F_1, F_2, \ldots, F_{10} \). The numbers of selected features in these sets are listed in Table 1. In the tabulation, the sizes of the 10 feature sets are nearly the same. After the features in these 10 sets were combined, 158 features were obtained (Supplementary Material S3). The obtained number (i.e., 158) did not differ much from the size of each feature set, which indicates that the majority of the 158 features were included in each set. In particular, among the 158 features, 146 features were included in all 10 feature sets, while 2, 1, 2, 3, and 4 feature/s were included in nine, eight, seven, six, and less than six feature sets (Figure 2), respectively. Considering that the negative associations in each of the 10 datasets somewhat differed, we predicted that the random selection of negative PPIs will not have a strong influence on the selection of the 158 features; i.e., the features can effectively determine the difference between positive and negative associations. Figure 3 shows a heat map of MI values of the 158 features in the 10 datasets. In the figure, the MI values of each of the 158 features in the 10 datasets are nearly the same. Similarly, the distributions of the MI values of the 158 features in the 10 datasets are nearly the same, which validates the above-mentioned results. Subsequently, an extensive investigation to further uncover the mechanism of proteins with function associations was conducted.

A careful checking showed that the important 158 features were derived from 134 GO terms and one KEGG pathway (Supplementary Material S4). To further evaluate their...
importance, we adopted a calculation technique called rating score measurement for each GO term. In this paper, the rating score is expressed as the sum of MI values of the related features in the 10 MaxRel feature lists. The scores are also provided in Supplementary Material S4. The rating score for the KEGG pathway (hsa03010) was 0.107, while the distribution of rating scores for 134 GO terms is illustrated in Figure 4.

3.3. Analysis of the Importance of Selected Features. As mentioned in Section 3.2, we finally selected 158 features that were deemed to be highly related to PPIs. To confirm such conclusion, we did the following test. For each of ten datasets mentioned in Section 2.1, each sample in the dataset was represented by these 158 features. And we also randomly constructed 100 feature sets, each of which consisted of 158 features. Samples in $D_S$ were represented by features in each of these feature sets to comprise 100 datasets. The classic classification algorithm, random forest (RF) [44, 46–50], was performed on all above-mentioned datasets, evaluated by tenfold cross-validation. The predicted results were counted as Matthews correlation coefficient (MCC) [40, 44, 47, 51–53], which are shown in Figure 5. It can be observed that the RF with selected 158 features yielded the MCCs between 0.55 and 0.60, while the RF with randomly selected 158 features generated the MCCs around 0.23. Clearly, the selected 158 features can capture the essential properties of PPIs, thereby providing more powerful distinguishing ability. Investigation on these features can help uncover the mechanisms of PPIs.

4. Discussion

As mentioned in Section 3.2, 134 GO terms and one KEGG pathway were regarded important in determining the difference between positive and negative associations. This section gave a detailed analysis on them.

4.1. Analysis of Key GO Terms. Analyzing above-mentioned 134 GO terms one by one is difficult. Here, we selected the most important 21 GO terms with rating scores larger than 0.5 for detailed analysis, which are listed in Table 2. 21 GO terms can be clustered into three groups: cellular component, molecular function, and biological process [15]. The distribution of the aforementioned 21 GO terms on these three groups is shown in Figure 6. Eleven GO terms are clustered into cellular component, three terms into molecular function, and seven terms into biological process. All these GO terms can be proven or inferred as associated with PPIs in published literature, as to be discussed below.

Cellular Component GO Terms. As described above, eleven of the 21 GO terms clustered as cellular components refer to the
Table 2: Information of most important 21 GO terms.

<table>
<thead>
<tr>
<th>GO term ID</th>
<th>GO term</th>
<th>Rating score</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0044260</td>
<td>cellular macromolecule metabolic process</td>
<td>0.688</td>
<td>Biological process</td>
</tr>
<tr>
<td>GO:0043170</td>
<td>macromolecule metabolic process</td>
<td>0.640</td>
<td>Biological process</td>
</tr>
<tr>
<td>GO:0044428</td>
<td>nuclear part</td>
<td>0.618</td>
<td>Cellular component</td>
</tr>
<tr>
<td>GO:1901363</td>
<td>heterocyclic compound binding</td>
<td>0.600</td>
<td>Molecular function</td>
</tr>
<tr>
<td>GO:0032991</td>
<td>protein-containing complex</td>
<td>0.593</td>
<td>Cellular component</td>
</tr>
<tr>
<td>GO:0097159</td>
<td>organic cyclic compound binding</td>
<td>0.591</td>
<td>Molecular function</td>
</tr>
<tr>
<td>GO:0031981</td>
<td>nuclear lumen</td>
<td>0.590</td>
<td>Cellular component</td>
</tr>
<tr>
<td>GO:0044238</td>
<td>primary metabolic process</td>
<td>0.589</td>
<td>Biological process</td>
</tr>
<tr>
<td>GO:0003676</td>
<td>nucleic acid binding</td>
<td>0.583</td>
<td>Molecular function</td>
</tr>
<tr>
<td>GO:0090304</td>
<td>nucleic acid metabolic process</td>
<td>0.569</td>
<td>Biological process</td>
</tr>
<tr>
<td>GO:0071704</td>
<td>organic substance metabolic process</td>
<td>0.556</td>
<td>Biological process</td>
</tr>
<tr>
<td>GO:0044237</td>
<td>cellular metabolic process</td>
<td>0.552</td>
<td>Biological process</td>
</tr>
<tr>
<td>GO:0005634</td>
<td>nucleus</td>
<td>0.549</td>
<td>Cellular component</td>
</tr>
<tr>
<td>GO:0044446</td>
<td>intracellular organelle part</td>
<td>0.547</td>
<td>Cellular component</td>
</tr>
<tr>
<td>GO:0044424</td>
<td>intracellular part</td>
<td>0.537</td>
<td>Cellular component</td>
</tr>
<tr>
<td>GO:0044422</td>
<td>organelle part</td>
<td>0.536</td>
<td>Cellular component</td>
</tr>
<tr>
<td>GO:0070013</td>
<td>intracellular organelle lumen</td>
<td>0.529</td>
<td>Cellular component</td>
</tr>
<tr>
<td>GO:0005622</td>
<td>intracellular</td>
<td>0.523</td>
<td>Cellular component</td>
</tr>
<tr>
<td>GO:0043233</td>
<td>organelle lumen</td>
<td>0.521</td>
<td>Cellular component</td>
</tr>
<tr>
<td>GO:0031974</td>
<td>membrane-enclosed lumen</td>
<td>0.514</td>
<td>Cellular component</td>
</tr>
<tr>
<td>GO:0006139</td>
<td>nucleobase-containing compound metabolic process</td>
<td>0.506</td>
<td>Biological process</td>
</tr>
</tbody>
</table>

Figure 6: Distribution of 21 GO terms on three groups: cellular component, molecular function, and biological process.

Part of a single cell and its specific extracellular environment, taking account for more than 52% of selected GO terms [16]. Comparing with molecular function and biological process as other two GO categories, which mostly reflect the indirect and functional relationships between different proteins, cellular component reflects the direct interactive relationships. Thus, the enrichment of functional clustered GO terms in such GO category indicated that subcellular localization and regional protein distribution may contribute more to the distinction of positive and negative associations. Direct PPIs which take the majority of all PPIs relied on the direct molecular interactions between proteins. The participants of most positive associations must share similar physical subcellular localizations, while those of the negative ones do not have to. Therefore, comparing to molecular function and biological process, it is quite reasonable for the cellular component category of GO terms to take the majority of all the enriched biological processes contributing to the recognition of positive PPIs.

The cellular component GO term with the highest rating score was GO: 0044428, describing the nuclear part of the eukaryotic cells, involving in chromosomes housing and replicating. Such processes involve multiple effective PPIs, like Esc2 and Rad51 [54, 55]. Therefore, the functional enrichment of genes involved in such cellular component may be more probable to participate in an actual PPI, contributing to the recognition of positive PPIs. Similarly, GO: 0031981, describing the nuclear lumen region, and GO: 0005634, describing a more general region of the cell, nucleus, may also involve in multiple PPIs. It has been widely reported that the nucleus region involves multiple subgroup of PPIs, regulating the expression and replication of genes [56–58]. Therefore, having nucleus as one of the busiest regions in cells, genes identified in such region may actually tend to be participating in certain PPIs.

Apart from the nucleus region of the cell, according to our results, we also identified that cellular regions associated with functional organelles may also be related to PPIs. GO: 0044422, describing the organelle part of cells, GO: 0070013, describing intracellular organelle lumen, GO:0044446, describing the intracellular organelle part, and GO:0043233, describing organelle lumen, have all been screened out as the potential cellular components that may be
associated with positive PPIs [59]. Similar with the nucleus regions, comparing to extracellular matrix and other intracellular regions, the organelles and its related biochemical reactions space involve in more actually interacting PPIs [59– 61]. Therefore, PPIs that locate in such region tend out to actually happen, indicating that these GO terms contribute to describing an effective gene cellular component features of genes that actually participate in PPIs.

Apart from such specific GO terms, we also identified some more general ones, like GO: 0032991 (protein-containing complex), GO: 0044424 (intracellular part), GO: 0005622 (intracellular), and GO: 0031974 (membrane-enclosed lumen). They all describe the regions that enrich significant biological processes of the cells. Therefore, actual PPIs tend to enrich in such region, revealing the specific PPI distribution pattern in the eukaryotic cells.

**Molecular Function GO Terms.** Three molecular function associated GO terms were extracted. The top GO term was GO: 1901363, describing heterocyclic compound binding. According to recent publications, various PPIs can actually be functional enriched in the heterocyclic compound binding, like the interactions between PDK1 and AKT in the eukaryotic cells [62, 63]. Therefore, genes that participate in such molecular function may tend to be more probable to actually contribute to PPIs. Similarly, the other molecular function GO term, named GO: 0097159, which describes organic cyclic compound binding, also involves various PPIs, like interactions among TBK1, PDPK1, and AURKA [64]. As for the last term, GO: 0003676, it describes the nucleic acid binding. As analyzed above, the nucleus region, where nucleic acid binding processes mostly occur, can distinguish the positive and negative PPIs due to its relative high interaction frequency [56–58]. Therefore, it is quite reasonable to speculate that such molecular function GO term may also be related to PPIs.

**Biological Process GO Terms.** Apart from above-mentioned cellular component and molecular function associated GO terms, we also identified a group of functional enrichment results that can be clustered into the biological processes cluster. All these GO terms describe effective metabolic processes in the cells. GO: 0044260 and GO: 0043170 describe the macromolecule metabolic processes. According to recent publications, such metabolic processes involve various PPIs, like interactions among TBK1, PDPK1, and AURKA [64]. As for the last term, GO: 0003676, it describes the nucleic acid binding. As analyzed above, the nucleus region, where nucleic acid binding processes mostly occur, can distinguish the positive and negative PPIs due to its relative high interaction frequency [56–58]. Therefore, it is quite reasonable to speculate that such molecular function GO term may also be related to PPIs.

For the remaining GO terms, we also identified a group of functional enrichment results that can be clustered into the biological processes cluster. All these GO terms describe effective metabolic processes in the cells. GO: 0044260 and GO: 0043170 describe the macromolecule metabolic processes. According to recent publications, such metabolic processes involve various PPIs, like interactions among TBK1, PDPK1, and AURKA [64]. As for the last term, GO: 0003676, it describes the nucleic acid binding. As analyzed above, the nucleus region, where nucleic acid binding processes mostly occur, can distinguish the positive and negative PPIs due to its relative high interaction frequency [56–58]. Therefore, it is quite reasonable to speculate that such molecular function GO term may also be related to PPIs.

For the remaining GO terms, we also identified a group of functional enrichment results that can be clustered into the biological processes cluster. All these GO terms describe effective metabolic processes in the cells. GO: 0044260 and GO: 0043170 describe the macromolecule metabolic processes. According to recent publications, such metabolic processes involve various PPIs, like interactions among TBK1, PDPK1, and AURKA [64]. As for the last term, GO: 0003676, it describes the nucleic acid binding. As analyzed above, the nucleus region, where nucleic acid binding processes mostly occur, can distinguish the positive and negative PPIs due to its relative high interaction frequency [56–58]. Therefore, it is quite reasonable to speculate that such molecular function GO term may also be related to PPIs.

For the remaining GO terms, we also identified a group of functional enrichment results that can be clustered into the biological processes cluster. All these GO terms describe effective metabolic processes in the cells. GO: 0044260 and GO: 0043170 describe the macromolecule metabolic processes. According to recent publications, such metabolic processes involve various PPIs, like interactions among TBK1, PDPK1, and AURKA [64]. As for the last term, GO: 0003676, it describes the nucleic acid binding. As analyzed above, the nucleus region, where nucleic acid binding processes mostly occur, can distinguish the positive and negative PPIs due to its relative high interaction frequency [56–58]. Therefore, it is quite reasonable to speculate that such molecular function GO term may also be related to PPIs.

**KEGG Pathways.** For the remaining GO terms, we also identified a group of functional enrichment results that can be clustered into the biological processes cluster. All these GO terms describe effective metabolic processes in the cells. GO: 0044260 and GO: 0043170 describe the macromolecule metabolic processes. According to recent publications, such metabolic processes involve various PPIs, like interactions among TBK1, PDPK1, and AURKA [64]. As for the last term, GO: 0003676, it describes the nucleic acid binding. As analyzed above, the nucleus region, where nucleic acid binding processes mostly occur, can distinguish the positive and negative PPIs due to its relative high interaction frequency [56–58]. Therefore, it is quite reasonable to speculate that such molecular function GO term may also be related to PPIs.

**4.2. Analysis of Other GO Terms and KEGG Pathways.** As for other GO terms extracted in this study, although not so relevant with PPIs as such GO terms described in Section 4.1, some of them have also been reported to be functionally related to certain PPIs. For instance, GO: 0006807, describing nitrogen compound metabolism, has been widely reported to be functionally related to compound-protein interactions but not protein–protein interactions [69, 70]. However, when extensively studying biological processes of such GO term, we found out that various specific PPIs are just like the interactions between the protein products of TIMP1 and MMP2 [71]. Therefore, in this study, some identified GO terms have not been directly reported to contribute to the PPIs. However, by digging deep into the actual biological processes, molecular functions and cellular components of them, we actually found that various novel identified PPIs are associated with these GO terms.

Furthermore, one KEGG pathway hsa03010 was obtained in our study. It describes the ribosome associated pathway. Considering that genes/proteins that participate in such pathway may interact with each other, forming the complex of ribosome, such KEGG pathway, may also contribute to the distinction of positive and negative PPIs.

**5. Conclusions**

This study investigated protein–protein functional associations based on GO terms and KEGG pathways. By using mutual information, we identified important GO terms and KEGG pathways that can describe the difference between actual associations and pairs of proteins without associations and help understand the mechanisms of protein interactions. A possible future research direction is to further use these GO terms and KEGG pathways to build a computational method for inferring novel associations between proteins, enriching the biological functional annotation of proteins.

**Data Availability**

The original data used to support the findings of this study are available at STRING dataset and in supplementary information files.

**Conflicts of Interest**

The authors declare that there are no conflicts of interest regarding the publication of this paper.
Acknowledgments

This study was supported by the National Natural Science Foundation of China (31701151), Natural Science Foundation of Shanghai (17ZR142500), National Key R&D Program of China (2018YFC0910403), Shanghai Sailing Program (16YF143800), the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245), the fund of the Key Laboratory of Stem Cell Biology of Chinese Academy of Sciences (201703), and Science and Technology Commission of Shanghai Municipality (STCSM) (18dz2271000).

Supplementary Materials

Supplementary 1. 70392 protein–protein functional associations.

Supplementary 2. A part of MaxRel feature list on ten datasets obtained by mutual information of each feature.

Supplementary 3. Selected 158 features and their occurrences in the 10 feature sets of the 10 datasets (√: feature is in the feature set; X: feature is not included in the feature set).

Supplementary 4. Extracted important GO terms and their rating scores.

References


References:


