Hindawi
BioMed Research International
Volume 2019, Article ID 4963289, 10 pages
https://doi.org/10.1155/2019/4963289

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 Cai**[6]

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

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 database on collected associations between proteins, as the positive samples, and then we randomly selected proteins to constitute the negative samples. Considering that the random selection of negative samples may influence the results, 10 sets of negative samples were constructed, thereby constituting 10 datasets, each of which contained the same positive samples. Each protein–protein functional association was encoded into a vector by using the GO terms and KEGG pathways. Then, mutual information was adopted to evaluate the importance of all features in each dataset. From the feature lists, in which features were ranked in the decreasing order of their importance, some important features were identified, and their corresponding GO terms or KEGG pathways were obtainable. Finally, we analyzed some most important GO terms and one KEGG pathway to partly uncover the difference between proteins with and without functional associations.

2. Materials and Methods

2.1. Materials. All human protein–protein functional associations used in this study were retrieved from STRING (http://www.string-db.org/, version 9.1) [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 S1.

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.

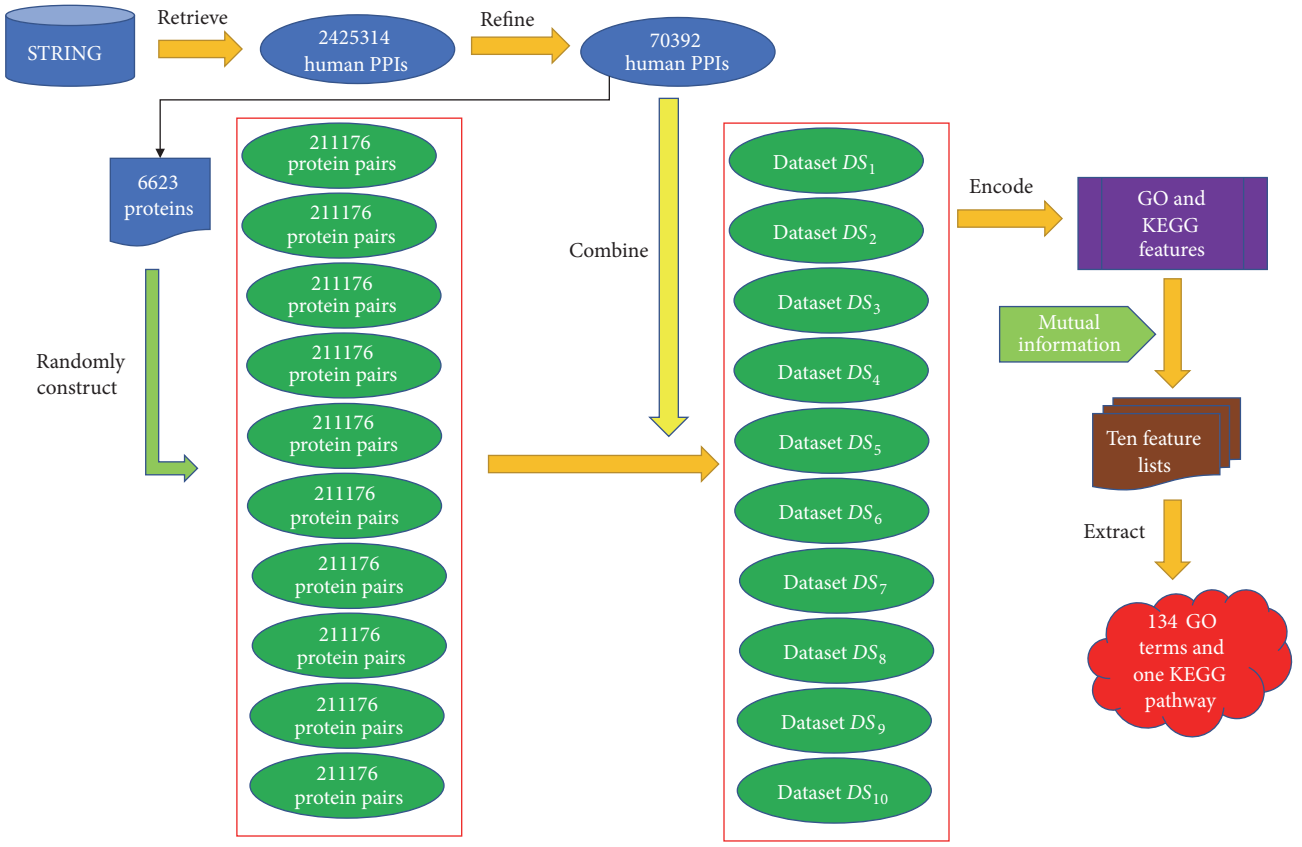

FIGURE 1: The whole procedures for analyzing protein–protein functional associations based on gene ontology (GO) and KEGG pathways. The raw 2,425,314 human PPIs were retrieved from STRING and refined by excluding similar proteins and selecting those validated by experiments, resulting in 70,392 PPIs. 6,623 proteins were involved in investigated PPIs and used to construct ten sets of protein pairs, each of which combined with 70,392 PPIs to constitute ten datasets. Each sample was represented by GO and KEGG features, which were evaluated by mutual information, producing ten feature lists, from which we extracted most important features, corresponding to 134 GO terms and one KEGG pathway.

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

$$v_{\mathrm{GO}}(p) = \left[g_1^p, g_2^p, \ldots, g_{17916}^p\right]^{\mathrm{T}}, \tag{1}$$

where

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

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 $v_{\mathrm{GO}}(p_1)$ and $v_{\mathrm{GO}}(p_2)$ as follows:

$$\begin{aligned} V_{\mathrm{GO}}(P) = v_{\mathrm{GO}}(p_1) \otimes v_{\mathrm{GO}}(p_2) = \left[g_1^{p_1}\right. \\ + g_1^{p_2}, \left|g_1^{p_1} - g_1^{p_2}\right|, \ldots, g_{17916}^{p_1} \\ \left. + g_{17916}^{p_2}, \left|g_{17916}^{p_1} - g_{17916}^{p_2}\right|\right]^{\mathrm{T}}. \end{aligned} \tag{3}$$

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

$$v_{\mathrm{pathway}}(p) = \left[k_1^p, k_2^p, \ldots, k_{279}^p\right]^{\mathrm{T}}, \tag{4}$$

where

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

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

$$\begin{aligned} V_{\mathrm{pathway}}(P) = v_{\mathrm{pathway}}(p_1) \otimes \mathrm{v}_{\mathrm{pathway}}(p_2) = \left[k_1^{p_1}\right. \\ \left. + k_1^{p_2}, \left|k_1^{p_1} - k_1^{p_2}\right|, \ldots, k_{279}^{p_1} + k_{279}^{p_2}, \left|k_{279}^{p_1} - k_{279}^{p_2}\right|\right]^{\mathrm{T}}. \end{aligned} \tag{6}$$

By integrating the GO term and KEGG pathway information of proteins into $P = (p_1, p_2)$, each association can be finally encoded as

$$V(P) = V_{\mathrm{GO}}(P) \oplus V_{\mathrm{pathway}}(P) = \begin{bmatrix} V_{\mathrm{GO}}(P) \\ V_{\mathrm{pathway}}(P) \end{bmatrix}. \quad (7)$$

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:

$$I(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)} dxdy, \quad (8)$$

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

$$L = [f_1, f_2, \ldots, f_N], \quad (9)$$

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.

3.2. Extracting Important GO Terms and KEGG Pathways. Features with high ranks (large MI values) in the MaxRel feature list are more important than those with low ranks (small MI values). For the MI value, we set 0.01 as the threshold to select important features in each MaxRel feature 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

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

Dataset	Number of selected features
DS_1	154
DS_2	154
DS_3	153
DS_4	155
DS_5	149
DS_6	150
DS_7	155
DS_8	152
DS_9	153
DS_{10}	153

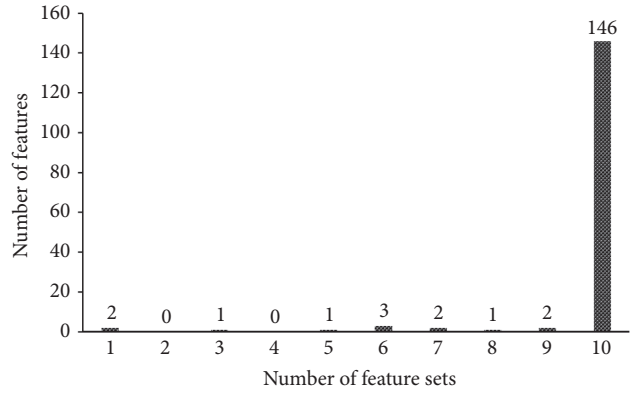

Figure 2: Distribution of 158 selected features: 146, 2, 1, 2, 3, and 4 feature/s in 10, 9, 8, 7, 6, and less than 6 feature sets derived from 10 datasets, respectively.

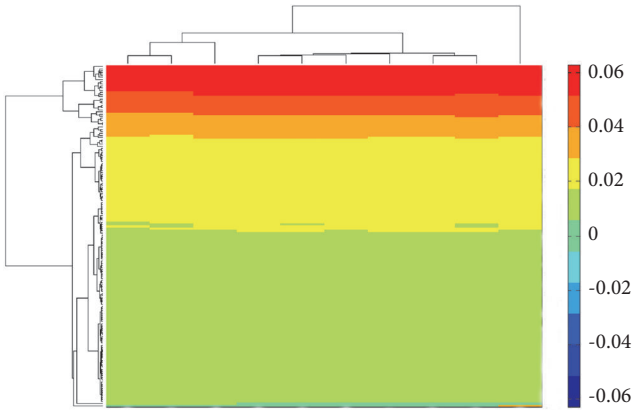

Figure 3: Heat map of MI values of 158 features in the 10 datasets. X-axis represents ten datasets; Y-axis represents 158 features.

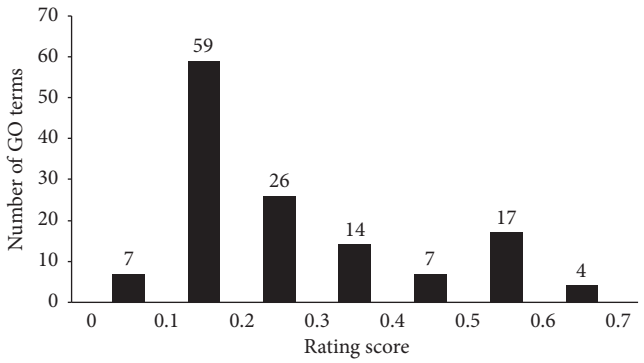

Figure 4: The distribution of the rating scores of 134 selected GO terms.

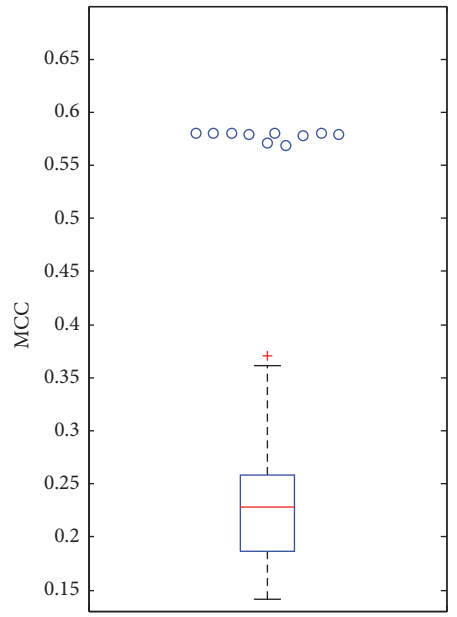

Figure 5: The performance of the random forest (RF) on ten datasets, in which samples were represented by selected 158 features or randomly selected 158 features, evaluated by tenfold cross-validation. The box plot indicates the distribution of MCCs yielded by RF with randomly selected 158 features and the circles represent the MCCs yielded by RF with selected 158 features on ten datasets. It is clear that based on selected 158 selected features, RF produced much better performance, implying the strong associations between these features and PPIs.

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 DS_1 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.

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

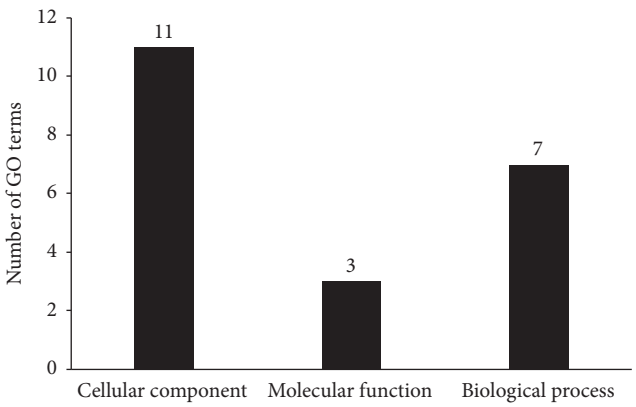

Figure 6: Distribution of 21 GO terms on three groups: cellular component, molecular function, and biological process.

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 the interactions in mTOR signaling pathways [65]. Apart from that, GO: 0044238, describing the primary metabolic process, has also been confirmed to contribute to PPIs. Considering the normal anabolic and catabolic processes, all involving functional PPIs [66–68], it is quite reasonable for genes participating in such biological processes to also participate in effective PPIs. The following three GO terms, GO: 0090304 (nucleic acid metabolic process), GO: 0006139 (nucleobase-containing compound metabolic process), and GO: 0071704 (organic substance metabolic process), may also contribute to PPIs, considering that the nucleus region has been discussed to be quite significant for PPIs [56–58] and proteins turn out to be one of the major subgroups of organic substance in eukaryotic cells; these three GO terms may also actually contribute to the identification of PPIs. As for the remaining GO term, GO: 0044237, it describes a general concept of all the cellular metabolic processes. Considering the analyses listed above, metabolic processes in the cells enrich various actual PPIs and are reasonable to be predicted and screened out as a potential identifier for positive PPIs.

On the basis of the analyses, all 21 GO terms are involved in different aspects of PPI, and they can be used to mark proteins with functional associations. For the remaining GO terms shown in Supplementary Material S4, it is anticipated that they also have associations with 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 (17ZR1412500), National Key R&D Program of China (2018YFC0910403), Shanghai Sailing Program (16YF1413800), 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 ($\surd$: feature is in the feature set; $\times$: feature is not included in the feature set).

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

References

[1] H. B. A. Lodish, P. Matsudaira, C. A. Kaiser et al., *Molecular Cell Biology*, W. H. Freeman, Ed., New York, NY, USA, 2004.

[2] N. Safari-Alighiarloo, M. Taghizadeh, M. Rezaei-Tavirani, B. Goliaei, and A. A. Peyvandi, "Protein-protein interaction networks (PPI) and complex diseases," *Gastroenterology and Hepatology from Bed to Bench*, vol. 7, no. 1, pp. 17–31, 2014.

[3] C. Chakraborty, G. Priya Doss. C, L. Chen, and H. Zhu, "Evaluating protein-protein interaction (PPI) networks for diseases pathway, target discovery, and drug-design using 'In silico pharmacology'," *Current Protein & Peptide Science*, vol. 15, no. 6, pp. 561–571, 2014.

[4] K. Tepper, J. Biernat, S. Kumar et al., "Oligomer formation of tau protein hyperphosphorylated in cells," *The Journal of Biological Chemistry*, vol. 289, no. 49, pp. 34389–34407, 2014.

[5] S. Shukla, U. S. Allam, A. Ahsan et al., "KRAS protein stability is regulated through SMURF2: UBCH5 complex-mediated β-TrCP1 degradation," *Neoplasia (United States)*, vol. 16, no. 2, pp. 115–128, 2014.

[6] B. Zakeri, J. O. Fierer, E. Celik et al., "Peptide tag forming a rapid covalent bond to a protein, through engineering a bacterial adhesin," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 109, no. 12, pp. E690–E697, 2012.

[7] I. M. Rosenberg, Ed., *Protein Analysis and Purification: Benchtop Techniques*, 2nd edition, 2005.

[8] J. Snider, S. Kittanakom, J. Curak, and I. Stagljar, "Split-ubiquitin based membrane yeast two-hybrid (MYTH) system: a powerful tool for identifying protein-protein interactions," *Journal of Visualized Experiments*, no. 36, 2010.

[9] S. Bruckner, F. Hüffner, R. M. Karp, R. Shamir, and R. Sharan, "Topology-free querying of protein interaction networks," *Journal of Computational Biology*, vol. 17, no. 3, pp. 237–252, 2010.

[10] B. Neyshabur, A. Khadem, S. Hashemifar, and S. S. Arab, "NETAL: a new graph-based method for global alignment of protein-protein interaction networks," *Bioinformatics*, vol. 29, no. 13, pp. 1654–1662, 2013.

[11] R. Jansen, H. Yu, D. Greenbaum et al., "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, no. 5644, pp. 449–453, 2003.

[12] J. Shen, J. Zhang, X. Luo et al., "Predicting protein-protein interactions based only on sequences information," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 104, no. 11, pp. 4337–4341, 2007.

[13] X.-Y. Pan, Y.-N. Zhang, and H.-B. Shen, "Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features," *Journal of Proteome Research*, vol. 9, no. 10, pp. 4992–5001, 2010.

[14] S. Hashemifar, B. Neyshabur, A. A. Khan, and J. Xu, "Predicting protein–protein interactions through sequence-based deep learning," *Bioinformatics*, vol. 34, no. 17, pp. i802–i810, 2018.

[15] R. P. Huntley, T. Sawford, P. Mutowo-Meullenet et al., "The GOA database: gene ontology annotation updates for 2015," *Nucleic Acids Research*, vol. 43, no. D1, pp. D1057–D1063, 2015.

[16] The Gene Ontology Consortium, "Gene ontology consortium: going forward," *Nucleic Acids Research*, vol. 43, no. D1, pp. D1049–D1056, 2015.

[17] J. Gillis and P. Pavlidis, "Assessing identity, redundancy and confounds in gene ontology annotations over time," *Bioinformatics*, vol. 29, no. 4, pp. 476–482, 2013.

[18] S. G. Jantzen, B. J. Sutherland, D. R. Minkley, and B. F. Koop, "GO trimming: systematically reducing redundancy in large gene ontology datasets," *BMC Research Notes*, vol. 4, p. 267, 2011.

[19] A. Patil and H. Nakamura, "Filtering high-throughput protein-protein interaction data using a combination of genomic features," *BMC Bioinformatics*, vol. 6, p. 100, 2005.

[20] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein-protein interactions," *Bioinformatics*, vol. 21, supplement 1, pp. i38–i46, 2005.

[21] S. R. Maetschke, M. Simonsen, M. J. Davis, and M. A. Ragan, "Gene ontology-driven inference of protein-protein interactions using inducers," *Bioinformatics*, vol. 28, no. 1, pp. 69–75, 2012.

[22] C. Von Mering, L. J. Jensen, B. Snel et al., "STRING: known and predicted protein-protein associations, integrated and transferred across organisms," *Nucleic Acids Research*, vol. 33, supplement 1, pp. D433–D437, 2005.

[23] A. Franceschini, D. Szklarczyk, S. Frankild et al., "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Research*, vol. 41, no. 1, pp. D808–D815, 2013.

[24] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[25] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[26] L. Liu, L. Chen, Y.-H. Zhang et al., "Analysis and prediction of drug-drug interaction by minimum redundancy maximum relevance and incremental feature selection," *Journal of Biomolecular Structure and Dynamics*, vol. 35, no. 2, pp. 312–329, 2017.

[27] L. Chen, B.-Q. Li, M.-Y. Zheng, J. Zhang, K.-Y. Feng, and Y.-D. Cai, "Prediction of effective drug combinations by chemical interaction, protein interaction and target enrichment of KEGG pathways," *BioMed Research International*, vol. 2013, Article ID 723780, 10 pages, 2013.

[28] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[29] L. Chen, Y.-H. Zhang, G. Lu, T. Huang, and Y.-D. Cai, "Analysis of cancer-related lncRNAs using gene ontology and KEGG pathways," *Artificial Intelligence in Medicine*, vol. 76, pp. 27–36, 2017.

[30] B. Q. Li, L.-L. Zheng, K.-Y. Feng, L.-L. Hu, G.-H. Huang, and L. Chen, "Prediction of linear B-ceel epitopes with mRMR feature selection and analysis," *Current Bioinformatics*, vol. 11, no. 1, pp. 22–31, 2016.

[31] L. Chen, X. Pan, X. Hu et al., "Gene expression differences among different MSI statuses in colorectal cancer," *International Journal of Cancer*, vol. 143, no. 7, pp. 1731–1740, 2018.

[32] Y. Zhang, C. Ding, and T. Li, "Gene selection algorithm by combining reliefF and mRMR," *BMC Genomics*, vol. 9, supplement 2, p. S27, 2008.

[33] Q. Ni and L. Chen, "A feature and algorithm selection method for improving the prediction of protein structural class," *Combinatorial chemistry & high throughput screening*, vol. 20, no. 7, pp. 612–621, 2017.

[34] J. Li, L. Lu, Y.-H. Zhang et al., "Identification of synthetic lethality based on a functional network by using machine learning algorithms," *Journal of Cellular Biochemistry*, vol. 120, no. 1, pp. 405–416, 2019.

[35] L. Chen, Y. Zhang, M. Zheng, T. Huang, and Y. Cai, "Identification of compound–protein interactions through the analysis of gene ontology, KEGG enrichment for proteins and molecular fragments of compounds," *Molecular Genetics and Genomics*, vol. 291, no. 6, pp. 2065–2079, 2016.

[36] Y. Zhou, N. Zhang, B.-Q. Li, T. Huang, Y.-D. Cai, and X.-Y. Kong, "A method to distinguish between lysine acetylation and lysine ubiquitination with feature selection and analysis," *Journal of Biomolecular Structure and Dynamics*, vol. 33, no. 11, pp. 2479–2490, 2015.

[37] L. Chen, Y. Zhang, Q. Zou, C. Chu, and Z. Ji, "Analysis of the chemical toxicity effects using the enrichment of Gene Ontology terms and KEGG pathways," *Biochimica et Biophysica Acta (BBA) - General Subjects*, vol. 1860, no. 11, Part B, pp. 2619–2626, 2016.

[38] X. Ma, J. Guo, and X. Sun, "Sequence-based prediction of RNA-binding proteins using random forest with minimum redundancy maximum relevance feature selection," *BioMed Research International*, vol. 2015, Article ID 425810, 10 pages, 2015.

[39] F. Li, C. Li, M. Wang et al., "GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome," *Bioinformatics*, vol. 31, no. 9, pp. 1411–1419, 2015.

[40] L. Chen, S. Wang, Y. Zhang et al., "Identify key sequence features to improve CRISPR sgRNA efficacy," *IEEE Access*, vol. 5, pp. 26582–26590, 2017.

[41] M. Radovic, M. Ghalwash, N. Filipovic, and Z. Obradovic, "Minimum redundancy maximum relevance feature selection approach for temporal gene expression data," *BMC Bioinformatics*, vol. 18, no. 1, p. 9, 2017.

[42] S. Niu, L.-L. Hu, L.-L. Zheng et al., "Predicting protein oxidation sites with feature selection and analysis approach," *Journal of Biomolecular Structure and Dynamics*, vol. 29, no. 6, pp. 650–658, 2012.

[43] L. Chen, S. Wang, Y.-H. Zhang et al., "Prediction of nitrated tyrosine residues in protein sequences by extreme learning machine and feature selection methods," *Combinatorial Chemistry & High Throughput Screening*, vol. 21, no. 6, pp. 393–402, 2018.

[44] X. Zhao, L. Chen, and J. Lu, "A similarity-based method for prediction of drug side effects with heterogeneous information," *Mathematical Biosciences*, vol. 306, pp. 136–144, 2018.

[45] L. Chen, Y.-H. Zhang, X. Pan et al., "Tissue expression difference between mRNAs and lncRNAs," *International Journal of Molecular Sciences*, vol. 19, no. 11, p. 3416, 2018.

[46] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[47] X. Zhao, L. Chen, Z.-H. Guo, and T. Liu, "Predicting drug side effects with compact integration of heterogeneous networks," *Current Bioinformatics*, 2019.

[48] S. Wang, D. Wang, J. Li, T. Huang, and Y. Cai, "Identification and analysis of the cleavage site in a signal peptide using SMOTE, dagging, and feature selection methods," *Molecular Omics*, vol. 14, no. 1, pp. 64–73, 2018.

[49] S. Wang, Y.-H. Zhang, N. Zhang, L. Chen, T. Huang, and Y.-D. Cai, "Recognizing and predicting thioether bridges formed by lanthionine and β-methyllanthionine in lantibiotics using a random forest approach with feature selection," *Combinatorial chemistry & high throughput screening*, vol. 20, no. 7, pp. 582–593, 2017.

[50] Q. Zhang, X. Sun, K. Feng et al., "Predicting citrullination sites in protein sequences using mRMR method and random forest algorithm," *Comb Chem High Throughput Screen*, vol. 20, no. 2, pp. 164–173, 2017.

[51] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) - Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[52] H. Cui and L. Chen, "A binary classifier for the prediction of EC numbers of enzymes," *Current Proteomics*, vol. 16, no. 5, pp. 381–389, 2019.

[53] L. Chen, C. Chu, Y.-H. Zhang et al., "Identification of drug-drug interactions using chemical interactions," *Current Bioinformatics*, vol. 12, no. 6, pp. 526–534, 2017.

[54] M. Urulangodi, M. Sebesta, D. Menolfi et al., "Local regulation of the Srs2 helicase by the SUMO-like domain protein Esc2 promotes recombination at sites of stalled replication," *Genes & Development*, vol. 29, no. 19, pp. 2067–2080, 2015.

[55] S. Aranda, D. Rutishauser, and P. Ernfors, "Identification of a large protein network involved in epigenetic transmission in replicating DNA of embryonic stem cells," *Nucleic Acids Research*, vol. 42, no. 11, pp. 6972–6986, 2014.

[56] M. Santucci, T. Vignudelli, S. Ferrari et al., "The hippo pathway and YAP/TAZ–TEAD protein–protein interaction as targets for regenerative medicine and cancer treatment," *Journal of Medicinal Chemistry*, vol. 58, no. 12, pp. 4857–4873, 2015.

[57] Y. Li, M. Collins, J. An et al., "Immunoprecipitation and mass spectrometry defines an extensive RBM45 protein–protein interaction network," *Brain Research*, vol. 1647, pp. 79–93, 2016.

[58] P. D. McCrea and C. J. Gottardi, "Beyond β-catenin: prospects for a larger catenin network in the nucleus," *Nature Reviews Molecular Cell Biology*, vol. 17, no. 1, pp. 55–64, 2016.

[59] N. Kory, R. V. Farese, and T. C. Walther, "Targeting fat: mechanisms of protein localization to lipid droplets," *Trends in Cell Biology*, vol. 26, no. 7, pp. 535–546, 2016.

[60] T. Simmen and M. Tagaya, "Organelle communication at membrane contact sites (MCS): from curiosity to center stage in cell biology and biomedical research," *Advances in Experimental Medicine and Biology*, vol. 997, pp. 1–12, 2017.

[61] M. Tagaya and K. Arasaki, "Regulation of mitochondrial dynamics and autophagy by the mitochondria-associated membrane," *Advances in Experimental Medicine and Biology*, vol. 997, pp. 33–47, 2017.

[62] P. Abeyrathna and Y. Su, "The critical role of Akt in cardiovascular function," *Vascular Pharmacology*, vol. 74, pp. 38–48, 2015.

[63] T. J. Rettenmaier, J. D. Sadowsky, N. D. Thomsen et al., "A small-molecule mimic of a peptide docking motif inhibits the protein kinase PDK1," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 111, no. 52, pp. 18590–18595, 2014.

[64] C. Ctortecka, V. Palve, B. M. Kuenzi et al., "Functional proteomics and deep network interrogation reveal a complex mechanism of action of midostaurin in lung cancer cells," *Molecular & Cellular Proteomics*, vol. 17, no. 12, pp. 2434–2447, 2018.

[65] I. Ben-Sahra and B. D. Manning, "mTORC1 signaling and the metabolic control of cell growth," *Current Opinion in Cell Biology*, vol. 45, pp. 72–82, 2017.

[66] G. L. Dornelles, A. Bueno, J. S. de Oliveira et al., "Biochemical and oxidative stress markers in the liver and kidneys of rats submitted to different protocols of anabolic steroids," *Molecular and Cellular Biochemistry*, vol. 425, no. 1-2, pp. 181–189, 2017.

[67] M. G. V. Heiden, L. C. Cantley, and C. B. Thompson, "Understanding the warburg effect: the metabolic requirements of cell proliferation," *Science*, vol. 324, no. 5930, pp. 1029–1033, 2009.

[68] J. Kopitz, "Lipid glycosylation: a primer for histochemists and cell biologists," *Histochemistry and Cell Biology*, vol. 147, no. 2, pp. 175–198, 2017.

[69] J. Xu, M. Zha, Y. Li et al., "The interaction between nitrogen availability and auxin, cytokinin, and strigolactone in the control of shoot branching in rice (Oryza sativa L.)," *Plant Cell Reports*, vol. 34, no. 9, pp. 1647–1662, 2015.

[70] X. E. Feng, Q. J. Wang, J. Gao, S. R. Ban, and Q. S. Li, "Synthesis of novel nitrogen-containing heterocycle bromophenols and their interaction with Keap1 protein by molecular docking," *Molecules*, vol. 22, no. 12, p. 2142, 2017.

[71] P. Chen, D. Xuan, and J. Zhang, "Periodontitis aggravates kidney damage in obese mice by MMP2 regulation," *Bratislava Medical Journal*, vol. 118, no. 12, pp. 740–745, 2018.