IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 3, 2020, accepted March 11, 2020, date of publication April 1, 2020, date of current version April 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2984824

Extracting Explicable Rules for the Identification of Compound–Protein Interactions

LIUCUN ZHU[1], PENGFEI HUANG[1], RUI ZHU[1], FANGXIA GUAN[2], AND WENNA GUO[2]
[1]School of Life Sciences, Shanghai University, Shanghai 200444, China
[2]School of Life Sciences, Zhengzhou University, Zhengzhou 450001, China

Corresponding authors: Fangxia Guan (guanfangxia@126.com) and Wenna Guo (guowenna@zzu.edu.cn)

ABSTRACT Compound–protein interaction (CPI) is one of the essential interaction patterns in living organisms. However, its underlying mechanism has not been fully revealed because of its complicated processes. Determining CPIs with traditional experiments can reveal solid results. However, their defects, such as low efficiency and high cost, are also evident. Designing effective computational methods is an alternative way to determine CPIs. Several methods have been proposed, but such methods can provide limited information to reveal the mechanism of CPIs because most of them are black boxes. In this study, we tried to develop rule-based classifiers for the identification of CPIs. The obtained rules involved gene ontology, KEGG pathway, and molecular ACCess System fingerprint descriptor, which could describe the functional enrichment of CPIs, to constitute the criterion. Although the performance of rule-based classifiers was lower than that of previous black box classifiers, these classifiers could clarify the identification procedures and provide more information on the mechanism of CPIs. The reliability of the obtained rules was also analyzed.

INDEX TERMS Compound–protein interaction, gene ontology, KEGG pathway, molecular fragment, explicable rule, RIPPER.

I. INTRODUCTION

In all living creatures, proteins are one of the major components that support and maintain the fundamental biological processes [1], [2]. However, proteins cannot act alone or independently. They have to interact with one another and other effective compounds to accomplish objective living processes and determine related capabilities [3]. Therefore, compound–protein interaction (CPI) is one of the essential interaction patterns in living creatures [4].

The basic forms of CPI can be artificially divided into various subtypes based on the chemical essentials of interactive chemicals [4], [5]. For instance, protein–protein interaction (PPI) is an essential form of CPIs because proteins are also a specific subtype of compound [5]. Apart from PPIs, another compound subgroup, specifically small molecule compounds, has been identified to participate in CPIs and contribute to the precise regulation of biological processes in humans. This phenomenon reflects the complicated biological contribution of CPIs. Originally, CPIs were first and widely used in pharmacology to help describe *in vivo* pharmacokinetics and pharmacodynamics processes in detail [5]–[8]. Early in 1984, the concept of CPI was introduced to describe specific platelet recognition and aggregation processes induced by a famous drug, namely, aspirin (compounds), and two *in vivo* components, namely, thrombin (protein) and fibrinogen (protein) [9]. However, this complicated and functional group of molecular biological interactions has not been systematically studied for a long time because of the limitation of biological techniques that can be used to identify novel CPIs.

With the development of high-throughput sequencing techniques and mass spectrometry, CPIs have been widely explored, and chemical genomics, a novel biological field, has been presented [10]. Chemical genomics (chemogenomics) aims to systematically screen and identify effective chemicals or small molecules that may directly interact with or be functionally related to human *in vivo* proteins, especially traditional drug targets, such as receptors, kinases, proteases, and transmembrane proteins [10]–[12]. The potential core research objective of chemogenomics is the identification of definite CPIs. However, too many chemicals and *in vivo* proteins are available for one-by-one verification, which is not only expensive but also time consuming.

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou.

Two optimal strategies have been presented to solve this problem: (I) high-throughput automated small molecule/chemical screening technology [13] and (II) computational stimulation prediction [14]–[21]. Experimental approaches based on high-throughput automated screening technologies have been widely used in industrial drug screening and recognition, whereas computational stimulation prediction has been applied to fundamental research on the detailed biological contributions and characteristics of *in vivo* CPIs [13], [14]. The concept and basic biological importance of CPIs have been widely used in fundamental research and industry. However, most studies and applications have introduced the concept of CPIs by taking single interaction as a unit, and summarizing the functional enrichment and distributed characteristics of such CPIs as clusters is difficult. Therefore, CPIs' biological contributions and importance are difficult to be macroscopically described. Most computational methods have focused on the performance of methods. The predicted accuracy is increasing. However, most methods are black boxes, which provide limited biological insights.

The present study is a continuation of one previous study [14], which investigated CPIs with gene ontology (GO) [22], KEGG pathways [23], and molecular ACCess System (MACCS) [24] fingerprint descriptors. A random forest (RF)-based model was built, and several important GO terms were extracted and analyzed. However, our previous study discussed the single relationship between CPIs and one GO term. One GO term or pathway cannot determine CPIs because the underlying mechanism of CPIs is complicated. In the present study, we tried to analyze CPIs with the combination of GO, KEGG pathway, and MACCS fingerprint descriptor by building rule-based classifiers. We used the dataset reported in a previous study [14], which were retrieved from The Binding Database (BindingDB) [25]. After a rigorous feature analysis procedure, the remaining features were fed into a rule learning algorithm, namely, repeated incremental pruning to produce error reduction (RIPPER) algorithm [26]. As a result, a series of classification rules that may contribute to the distinction of events with differential binding affinity and possibility is established during interactions between compounds and proteins. The establishment of rules for CPI recognition may help identify functional and essential CPIs and enhance our understanding on the potential biological and functional characteristics of CPIs.

II. MATERIALS AND METHODS

A. MATERIALS

A total of 22,473 actual CPIs used in the present study were retrieved from the previous study [14], which were extracted from 211,888 CPIs obtained from BindingDB (http://www.bindingdb.org/, accessed in April 2014) [25], a public and web-accessible database focusing primarily onto the interactions of proteins and small/drug-like molecules. These actual CPIs were termed positive samples that included 756 Ensembl IDs (proteins) and 15,914 PubChem IDs (small/drug-like molecules).

Negative samples were necessary to reveal the underlying mechanism of CPIs, that is, which proteins and small/drug-like molecules could interact with one another. Here, we used five different negative sample sets from a previous study [14]. Each set contained 112,365 pairs of proteins and compounds, five times as many as positive samples. Negative samples were also called non-CPIs in the present study. After each negative sample set was combined with the positive sample set, five datasets were accessed and denoted as DS_1, DS_2, DS_3, DS_4, and DS_5.

B. FEATURE REPRESENTATION

Each sample was represented by the same features used in the previous study [14]. Proteins were encoded into 18195-D vectors. Each component indicated the linkage of proteins and one GO term or one KEGG pathway, which was measured in accordance with enrichment theory [27]. Among 18195 protein features, 17916 features were for GO terms and 279 features were about KEGG pathways. The compounds were represented by 166 MACCS fingerprint descriptors. After the features of proteins and compounds were combined, each pair of protein and compound was represented by 18,361 features.

C. FEATURE ANALYSIS

Each sample was represented by numerous features, and all features did not equally contribute to describing the mechanism of CPIs. Feature analysis was performed on each dataset. The protein features were approximately 109 times more than the compound features. The following feature analysis was only executed on protein features.

Feature analysis included two stages. At the first stage, all protein features were evaluated by using their mutual information (MI) to targets. For variables x and y, their MI value can be computed by

$$I(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy \tag{1}$$

where $p(x)$ represents the marginal probabilities of x, and $p(x, y)$ indicates the joint probabilistic distribution of x and y. Features with MI values larger than or equal to 0.01 were selected. At the second stage, the remaining features were analyzed by conducting a powerful feature selection method, namely, minimum redundancy maximum relevance (mRMR) [28], which has wide applications in tackling different biological problems [29]–[33]. This method was used to evaluate the importance of each feature by ranking all features in two lists, namely, MaxRel feature list and mRMR feature list. For the former list, features were sorted on the basis of their relevance to targets. For the latter list, the redundancies between features were further considered. As such, the MaxRel feature list was suitable for analyzing the contribution of a single feature, whereas the mRMR feature list was appropriate for measuring the contribution of a combination of features. The mRMR feature list at this

stage was used because the underlying mechanism of CPIs might involve several features. For each obtained list on each dataset, 500 top-ranked protein features were selected for further analysis because of our limited computational power.

The mRMR program used in this study was accessed at http://home.penglab.com/proj/mRMR/index.htm. mRMR program was executed with default parameters.

D. RULE LEARNING

In contrast to a previous study [14], which aimed to build an effective model for predicting CPIs and extract important GO terms and biological pathways, the present study tried to learn explicable rules to clearly identify the difference between the actual CPIs and general pairs of proteins and compounds. Thus, the underlying mechanism of CPIs was determined.

For each dataset with samples represented by 500 important protein features and 166 compound features, the RIPPER algorithm [26] was applied to the dataset. RIPPER is an improved version of Incremental Reduced Error Pruning [34]. In this algorithm, a greedy strategy was used to produce rules one by one. In each step, RIPPER generated a rule that could cover the remaining training samples as much as possible. The covered training samples were removed, and the following rule was generated on the remaining samples in the same manner. A rule generated by RIPPER was displayed with an IF–THEN clause. For example, a rule can be IF (feature1$>$0.002, featrue2$>$1.2, and feature3$<$1.6), THEN CPI. This rule learning algorithm has been applied to investigate several biological problems [35]–[37].

In the present study, the tool "JRip" in Weka [38] was used, which implemented the abovementioned RIPPER algorithm. Default parameters were used.

E. ACCURACY MEASUREMENT

The performance of the rules was evaluated by tenfold cross-validation to indicate the utility of rules learned by RIPPER algorithm [39]–[41]. With this method, the samples were divided into 10 parts. Each part was singled out individually as the testing dataset, and the remaining parts constituted the training dataset. Lastly, each sample was tested exactly once.

The predicted results yielded by tenfold cross-validation were counted as true positives, true negatives, false positives, and false negatives to calculate four measurements: sensitivity (SN), specificity (SP), accuracy (ACC), and Matthews's correlation coefficient (MCC) [29], [42]–[47]. The formulas of such measurements were presented as follows:

$$\begin{cases} \mathrm{SN} = \dfrac{TP}{TP+FN} \\ \mathrm{SP} = \dfrac{TN}{TN+FP} \\ \mathrm{ACC} = \dfrac{TP+TN}{TP+TN+FP+FN} \\ \mathrm{MCC} = \dfrac{TP \cdot TN - FP \cdot FN}{\sqrt{(TN+FN)\cdot(TN+FP)\cdot(TP+FN)\cdot(TP+FP)}} \end{cases} \tag{2}$$

MCC was selected as a key measurement because it is a balanced measurement, even if class sizes greatly differ. In addition, we also reported other three measurements: recall, precision and F1-measure, to provide a more complete evaluation of classifiers. They can be computed by

$$\begin{cases} \mathrm{Recall} = \dfrac{TP}{TP+FN} \\ \mathrm{Precision} = \dfrac{TP}{TP+FP} \\ \mathrm{F1-measure} = \dfrac{2 \cdot \mathrm{Recall} \cdot \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}} \end{cases} \tag{3}$$

Clearly, recall is same as SN.

III. RESULTS

In this study, we tried to extract explicable rules and to reveal the underlying mechanism of CPIs with GO terms, KEGG pathways, and MACCS fingerprint descriptors. The entire procedures are illustrated in **Fig. 1**.

A. RESULTS OF FEATURE ANALYSIS

In Section II.A, five datasets were constructed. In these datasets, each sample was represented by 18,195 protein features and 166 compound features. Protein features were evaluated through feature analysis. In Section II.C, each protein feature was first assigned a MI value, and features with MI values larger than or equal to 0.01 were selected. As a result, 5,720 features remained on DS_1; 5,684 features remained on DS_2; 5,624 features remained on DS_3; 5,672 features remained on DS_4; and 5,685 features remained on DS_5. A Venn diagram is plotted in **Fig. 2(A)** to show the relationship of these five feature subsets. A total of 5,523 features were the common features of the five feature subsets, indicating that the similarity of any two feature subsets was high. Important protein features (i.e., key GO terms and KEGG pathways) were included.

The remaining protein features on each dataset were analyzed with the mRMR method, resulting in a mRMR feature list. We selected 500 top-ranked features in each list for further analysis. The relationship between these feature sets is shown in **Fig. 2(B)**. In contrast to the Venn diagram in **Fig. 2(A)**, these five feature subsets were quite different; each set contained more than 20% of exclusive features, implying that each dataset included the exclusive information of CPIs. All information extracted from five datasets should be combined to provide a complete overview of the mechanism of CPIs.

B. PERFORMANCE OF THE RULES YIELDED BY RIPPER

In Section III.A, 500 key protein features were extracted on each dataset. They were combined with 166 compound features in representing each pair of proteins and compounds. Explicable rules on each dataset could be learned to determine the mechanism of CPIs. Before the rules were extracted, the utility of rules yielded by RIPPER should be evaluated. Thus, a tenfold cross-validation was executed on each

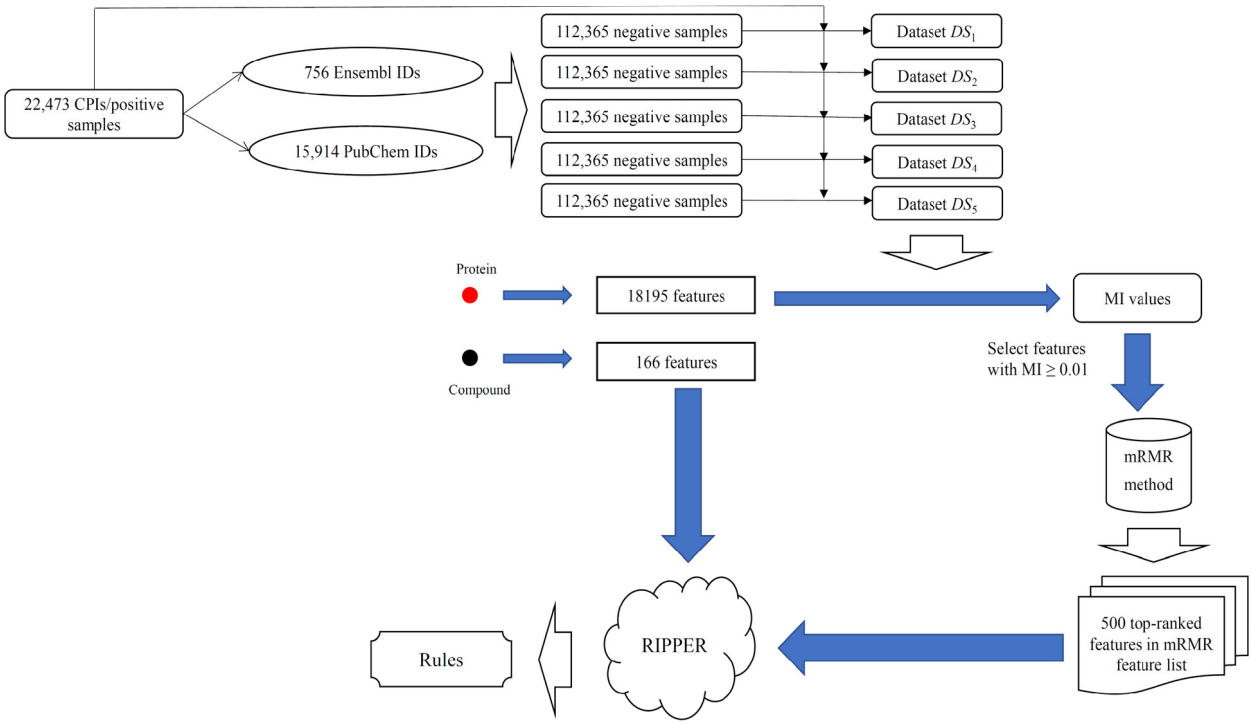


FIGURE 1. Entire procedures of feature analysis and rule learning.

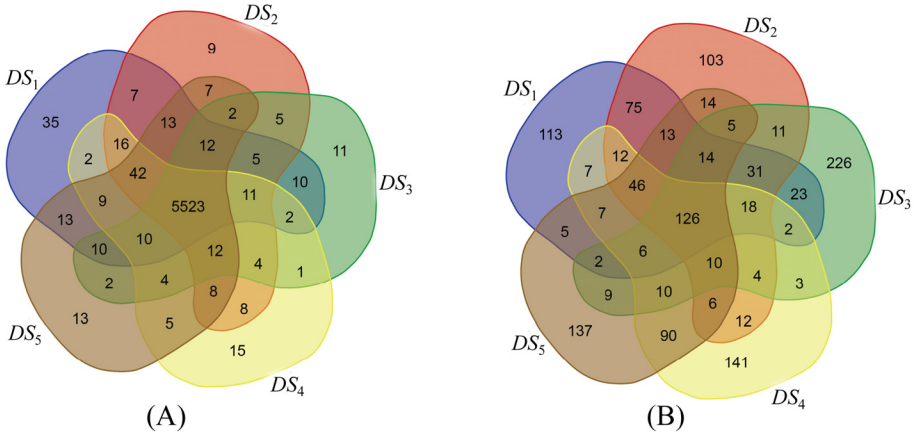


FIGURE 2. Venn diagrams of the protein features extracted from five datasets. (A) Venn diagram of the protein features with MI of ≥ 0.01 on five datasets; (B) Venn diagram of 500 top-ranked protein features in the mRMR feature list on five datasets.

dataset. The predicted results were calculated as SN (recall), SP, ACC, MCC, precision, and F1-measure as mentioned in Section II.E (**Table 1**). SN (recall), SP, ACC, MCC, precision, and recall were approximately 0.579, 0.963, 0.899, 0.606, 0.758, and 0.656, respectively. The performance of RIPPER classifiers was acceptable. The performance of RIPPER classifiers was quite stable with regard to standard deviation (**Table 1**).

In a previous study [14], RF, a powerful classification algorithm [29], [44], [48], [49], was used to identify CPIs. The performance of the optimal RF classifiers and RF classifiers with 500 top-ranked protein features on five datasets is shown in **Fig. 3**. It can be observed that SN of RIPPER classifiers was much lower than those of two RF classifiers, whereas SPs of three classifiers were almost at the same level. The entire performance of RIPPER classifiers was lower than that

TABLE 1. Performance of ripper classifier on five datasets.

Dataset	SN	SP	ACC	MCC	Precision	F1-measure
DS_1	0.583	0.962	0.899	0.606	0.754	0.658
DS_2	0.584	0.962	0.899	0.608	0.756	0.659
DS_3	0.577	0.962	0.898	0.602	0.753	0.653
DS_4	0.584	0.963	0.900	0.610	0.759	0.660
DS_5	0.567	0.965	0.899	0.603	0.766	0.651
Mean	0.579	0.963	0.899	0.606	0.758	0.656
Standard deviation	7.52E−03	1.41E−03	7.24E−04	3.42E−03	5.29E-03	3.88E-03

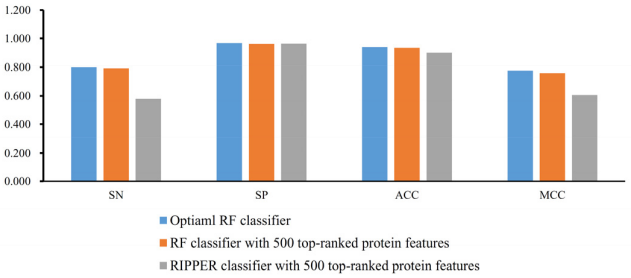


FIGURE 3. Average performance of three different classifiers on five datasets.

TABLE 2. Information of rules on five datasets.

Dataset	Number of rules	Number of used features	Most used features
DS_1	141	284	GO:0033029
DS_2	124	268	GO:0070542
DS_3	117	262	GO:0034694
DS_4	109	251	GO:0032891
DS_5	130	279	GO:0034694

of the two other classifiers. However, RF was a black box, which provided limited information to reveal the mechanism of CPIs. Although the RIPPER classifiers gave low performance, they could yield rules that could clearly display the feature combination that was a key biomarker in determining a CPI.

C. EXPICABLE RULES OBTAINED BY RIPPER

With regard to the acceptable performance of the rules yielded by RIPPER, we applied RIPPER to all the samples in each dataset, resulting in a group of rules. We obtained 141, 124, 117, 109, and 130 rules from the five datasets (Supplementary Material S1). Several features were used for constructing rules in each group. The number of these features is listed in column 3 of Table 2. For rules on DS_1, GO:0033029 was used most. Such features on other four datasets are listed in the last column of Table 2. By investigating all these rules, the mechanism of CPIs could be partly determined. The combination of some features could be a latent biomarker in determining CPIs. In Section IV, some rules would be extensively analyzed to indicate their reasonability.

TABLE 3. First rule yielded by ripper on each of the five datasets.

Dataset	Condition [a]	Result
DS_1	(GO: 0071379 ≥ 1.409) and (GO: 0071393 ≥ 1.214) and (GO: 0042421 ≥ 1.144) and (A131 ≤ 0) and (A133 ≥ 1) and (A92 ≥ 1) and (GO: 0071578 ≥ 0.356) and (A88 ≥ 1) and (A59 ≥ 1) and (A104 ≤ 0)	CPI
DS_2	(GO: 0034694 ≥ 1.939) and (GO: 0033240 ≥ 1.591) and (GO: 0071393 ≥ 1.297) and (A131 ≤ 0) and (A133 ≥ 1) and (GO: 0008747 ≥ 0.754) and (A136 ≥ 1) and (A127 ≤ 0) and (A110 ≥ 1)	CPI
DS_3	(GO: 0034694 ≥ 1.939) and (GO: 0033240 ≥ 1.591) and (GO: 0070530 ≥ 0.558) and (A131 ≤ 0) and (A135 ≥ 1) and (A117 ≥ 1) and (A59 ≥ 1)	CPI
DS_4	(GO: 0034694 ≥ 1.900) and (GO: 0042415 ≥ 1.433) and (GO: 0071393 ≥ 1.297) and (A135 ≥ 1) and (A131 ≤ 0) and (GO:1902186 ≥ 0.496) and (A92 ≥ 1) and (A130 ≥ 1)	CPI
DS_5	(GO: 0034694 ≥ 1.939) and (GO: 0071393 ≥ 1.214) and (GO: 0042421 ≥ 1.144) and (A135 ≥ 1) and (A131 ≤ 0) and (GO: 0030799 ≤ 13.088) and (A92 ≥ 1)	CPI

[a]Features with initials "A" represent MACCS fingerprint descriptors.

IV. DISCUSSION

In this study, five groups of explicable rules that might distinguish unbinding events (non-CPIs) and binding ones (CPIs) were extracted. For the first time, functional annotations (GO [22] and KEGG pathways [23]) and MACCS [24] fingerprint descriptors were applied to quantitatively describe each rule. Other studies have shown that some rules could confirm their reasonability, thereby validating the reliability of all new finding rules.

Five groups of rules that might contribute to the distinction of CPIs and non-CPIs were identified. A total of 621 rules were obtained. Analyzing these rules one by one is difficult because of our limited human resource. According to the principle of RIPPER, the first rule is quite important because such rule is constructed by viewing all the samples in the dataset. Thus, the first rule of each group was screened for a detailed discussion (Table 3).

The first rule on DS_1 contributed to the identification of CPIs. Four GO terms and six MACCS descriptors were involved. The simultaneous satisfaction of four biological processes (GO: 0071379 [cellular response to prostaglandin stimulus], GO: 0071393 [cellular response to progesterone stimulus], GO: 0042421 [norepinephrine biosynthetic process], and GO: 0071578 [zinc II ion transmembrane import])

was included. The interactions involved in such biological processes must participate in the interactions either between hormones (prostaglandin and progesterone) [50], [51] and receptors or between biosynthetic catalyzing enzymes and substrates [52], which are the subtypes of CPIs. Therefore, the interactions enriched in such GO terms may be CPIs, validating the distinction effects of this rule. For the MACCS descriptors, A131 (QH $>$ 1) described that the molecules participating in such interactions must have heteroatoms with at least one hydrogen attached [53]. Considering the complicated interactions, molecules with or without heteroatoms might not contribute to the interaction-associated binding process ($\leq$0), which was also inferred by recent publications. For other MACCS descriptors, A88 presented whether this sulfur-containing atom may contribute to the identification of binding processes, and A104 indicated that compounds with acylamino-like structures may also contribute to the specific subtype of CPIs [54]. Considering that various proteins and effective substrate have a sulfur atom [55] and acylamino (e.g., thioethers) [54], establishing this rule as potential distinguishers for CPIs and non-CPIs is quite appropriate.

For the top rule on dataset DS_2, four GO terms and five MACCS descriptors are presented to contribute to this distinction. GO: 0034694 (response to prostaglandin), GO: 0033240 (positive regulation of cellular amine metabolic process), GO: 0071393 (cellular response to progesterone stimulus), and GO: 0008747 (N-acetylneuraminate lyase activity) are optimal parameters that may contribute to distinguishing CPIs and non-CPIs. All these GO terms may participate in either hormone-associated ligand-receptor interactions or catalysis, which interacts with binding processes and validate the efficacy and accuracy of this rule. For structural-associated MACCS descriptors, molecules must have at least one heteroatom with at least one hydrogen attached. Considering that most interactive compounds can satisfy this rule, this parameter can be considered a complementary filter. For other MACCS descriptors, a specific structure (A110) describing acylamino may also be quite significant for the distinction of CPIs and non-CPIs [55], corresponding to the similar parameter (A104 describing acylamino-like structures) mentioned above.

According to the first rule on DS_3, three GO terms and four MACCS descriptors were used for such distinction. Two GO terms (GO: 0034694 [response to prostaglandin] and GO: 0033240 [positive regulation of cellular amine metabolic process]) also enriched binding-associated interactions on the basis of the quantitative prediction parameters. Considering the detailed biological functions of these GO terms, effective CPIs like interactions between alanine and aspartate aminotransferase [56] and interactions between exogenous prostaglandin E2 and related receptors [57], [58] have been widely reported and validated in such two biological processes, validating the efficacy and accuracy of this rule. For the MACCS descriptors, a specific item named A117 was selected to describe the modified version of a nitrogen–oxygen double bond for detailed discussion [59] and analysis. Considering that this double bond has been identified in multiple drugs, such as isostrychnine [60], the enrichment of this descriptor may be quite reliable.

According to the first rules on DS_4 and DS_5, each dataset contained four GO terms. In such rules, two GO terms (GO: 0034694 [response to prostaglandin] and GO: 0071393 [cellular response to progesterone stimulus]) were shared and discussed, and they reflected the efficacy, accuracy, and correspondence of these two rules. Apart from the two GO terms, GO: 0042415 (norepinephrine metabolic process) and GO: 0042421 (norepinephrine biosynthetic process) were predicted to enrich functional binding interactions with statistical significance ($p < 0.05$). Considering the metabolic and biosynthetic processes, which the two GO terms have described, we can confirm that genes enriching in these biological processes are definitely involved in CPIs. Similar to GO terms, the identified MACCS descriptors have quite high confidence. A131, A135, and A92 that have been discussed and have been identified as quantitative markers for CPI recognition with similar tendency and threshold. These results validated the efficacy and accuracy of these two rules.

Overall, similar to our analysis, all first rules can be validated by recent publications. The remaining rules may also be reasonable for distinguishing CPIs and non-CPIs. Considering all the rules, we can partly reveal the underlying mechanism of CPIs.

V. CONCLUSION

This study investigated CPIs by extracting explicable classification rules with GO terms, KEGG pathways, and MACCS fingerprint descriptors. In contrast to several previous studies, which tried to construct more effective models for the identification of CPIs, the obtained rules could clarify the prediction procedures and indicate the important combination of key features (GO terms, KEGG pathways, and MACCS fingerprints) to determine a CPI. The discussion section suggested that the obtained rules were important for the identification of CPIs.

ACKNOWLEDGMENT

(Liucun Zhu and Pengfei Huang contributed equally to this work.)

REFERENCES

[1] V. Marx, "Proteomics: An atlas of expression," *Nature*, vol. 509, pp. 645–649, May 2014.
[2] H. Zhao, Y. Yang, S. C. Janga, C. C. Kao, and Y. Zhou, "Prediction and validation of the unexplored RNA-binding protein atlas of the human proteome," *Proteins, Struct., Function, Bioinf.*, vol. 82, no. 4, pp. 640–647, Apr. 2014.
[3] A. Persson, S. Hober, and M. Uhlen, "A human protein atlas based on antibody proteomics," *Current Opinion Mol. Therapeutics*, vol. 8, pp. 90–185, Jun. 2006.
[4] C. Ottmann, "New compound classes: Protein-protein interactions," in *Handbook of Experimental Pharmacology*, vol. 232. Berlin, Germany: Springer, 2016, pp. 38–125.
[5] M. R. Arkin, Y. Tang, and J. A. Wells, "Small-molecule inhibitors of protein-protein interactions: Progressing toward the reality," *Chem. Biol.*, vol. 21, no. 9, pp. 1102–1114, Sep. 2014.

[6] M. Tutone, A. Lauria, and A. Almerico, "Leptin and the ob-receptor as anti-obesity target: Recent in silico advances in the comprehension of the protein-protein interaction and rational drug design of Anti-obesity lead compounds," *Current Pharmaceutical Des.*, vol. 20, no. 1, pp. 136–145, Jan. 2014.

[7] O. Cala, F. Guilliere, and I. Krimm, "NMR-based analysis of protein-ligand interactions," *Anal. Bioanal. Chem.*, vol. 406, pp. 56–943, Feb. 2014.

[8] Y. Song and P. Buchwald, "TNF superfamily protein-protein interactions: Feasibility of Small-molecule modulation," *Current Drug Targets*, vol. 16, no. 4, pp. 393–408, Apr. 2015.

[9] G. Agam and A. Livne, "Platelet-platelet recognition during aggregation: Distinct mechanisms determined by the release reaction," *Thrombosis Haemostasis*, vol. 51, no. 2, pp. 145–149, Apr. 1984.

[10] J. Li, J. Yuan, K. C.-C. Cheng, J. Inglese, and X.-Z. Su, "Chemical genomics for studying parasite gene function and interaction," *Trends Parasitol.*, vol. 29, no. 12, pp. 603–611, Dec. 2013.

[11] L. C. B. Olsen and N. J. Faergeman, "Chemical genomics and emerging DNA technologies in the identification of drug mechanisms and drug targets," *Current Topics Med. Chem.*, vol. 12, no. 12, pp. 1331–1345, Jun. 2012.

[12] E. D. Zanders, "Overview of chemical genomics and proteomics," in *Chemical Genomics and Proteomics* (Methods in Molecular Biology), vol. 800. Berlin, Germany: Springer, 2012, pp. 3–10.

[13] C.-W. Chi, A. R. Ahmed, Z. Dereli-Korkut, and S. Wang, "Microfluidic cell chips for high-throughput drug screening," *Bioanalysis*, vol. 8, no. 9, pp. 921–937, May 2016.

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

[15] M. Tsubaki, K. Tomii, and J. Sese, "Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences," *Bioinformatics*, vol. 35, no. 2, pp. 309–318, Jan. 2019.

[16] L. Jacob and J.-P. Vert, "Protein-ligand interaction prediction: An improved chemogenomics approach," *Bioinformatics*, vol. 24, no. 19, pp. 2149–2156, Oct. 2008.

[17] F. Cheng, Y. Zhou, J. Li, W. Li, G. Liu, and Y. Tang, "Prediction of chemical–protein interactions: Multitarget-QSAR versus computational chemogenomic methods," *Mol. BioSyst.*, vol. 8, no. 9, pp. 2373–2384, 2012.

[18] Y. Tabei, "Scalable prediction of Compound–protein interaction on compressed molecular fingerprints," *Mol. Informat.*, vol. 39, nos. 1–2, Jan. 2020, Art. no. 1900130.

[19] Y. Tabei and Y. Yamanishi, "Scalable prediction of compound-protein interactions using minwise hashing," *BMC Syst. Biol.*, vol. 7, no. 6, p. S3, 2013.

[20] K. Tian, M. Shao, Y. Wang, J. Guan, and S. Zhou, "Boosting compound-protein interaction prediction by deep learning," *Methods*, vol. 110, pp. 64–72, Nov. 2016.

[21] H. Liu, J. Sun, J. Guan, J. Zheng, and S. Zhou, "Improving compound–protein interaction prediction by building up highly credible negative samples," *Bioinformatics*, vol. 31, no. 12, pp. i221–i229, Jun. 2015.

[22] The Gene Ontology Consortium, "Gene ontology consortium: Going forward," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D1049–D1056, Jan. 2015.

[23] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: Back to metabolism in KEGG," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D199–D205, Jan. 2014.

[24] M. Sud, "MayaChemTools: An open source package for computational drug discovery," *J. Chem. Inf. Model.*, vol. 56, no. 12, pp. 2292–2297, Dec. 2016.

[25] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "BindingDB: A Web-accessible database of experimentally determined protein-ligand binding affinities," *Nucleic Acids Res.*, vol. 35, pp. D198–D201, Jan. 2007.

[26] W. W. Cohen, "Fast effective rule induction," in *Proc. 12th Int. Conf. Mach. Learn.*, 1995, pp. 115–123.

[27] P. Carmona-Saez, M. Chagoyen, F. Tirado, J. M. Carazo, and A. Pascual-Montano, "GENECODIS: A Web-based tool for finding significant concurrent annotations in gene lists," *Genome Biol.*, vol. 8, p. R3, Jan. 2007.

[28] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

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

[30] L. Chen, X. Pan, X. Hu, Y.-H. Zhang, S. Wang, T. Huang, and Y.-D. Cai, "Gene expression differences among different MSI statuses in colorectal cancer," *Int. J. Cancer*, vol. 143, no. 7, pp. 1731–1740, Oct. 2018.

[31] L. Chen, S. Wang, Y.-H. Zhang, J. Li, Z.-H. Xing, J. Yang, T. Huang, and Y.-D. Cai, "Identify key sequence features to improve CRISPR sgRNA efficacy," *IEEE Access*, vol. 5, pp. 26582–26590, 2017.

[32] J. Li, L. Lu, Y.-H. Zhang, M. Liu, L. Chen, T. Huang, and Y.-D. Cai, "Identification of synthetic lethality based on a functional network by using machine learning algorithms," *J. Cellular Biochem.*, vol. 120, no. 1, pp. 405–416, Jan. 2019.

[33] T. Wang, L. Chen, and X. Zhao, "Prediction of drug combinations with a network embedding method," *Combinatation Chem. High Throughput Screening*, vol. 21, no. 10, pp. 789–797, Feb. 2019.

[34] F. Johannes and G. Widmer, "Incremental reduced error pruning," in *Proc. 11th Annu. Conf. Mach. Learn.*, 1994, pp. 70–77.

[35] X. Pan, T. Zeng, F. Yuan, Y.-H. Zhang, L. Chen, L. Zhu, S. Wan, T. Huang, and Y. Cai, "Screening of methylation signature and gene functions associated with the subtypes of isocitrate dehydrogenase-mutation gliomas," *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 339, Nov. 2019.

[36] S. Zhang, X. Pan, T. Zeng, W. Guo, Z. Gan, Y.-H. Zhang, L. Chen, Y. Zhang, T. Huang, and Y.-D. Cai, "Copy number variation pattern for discriminating MACROD2 states of colorectal cancer subtypes," *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 407, Dec. 2019.

[37] D. Wang, J.-R. Li, Y.-H. Zhang, L. Chen, T. Huang, and Y.-D. Cai, "Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms," *Genes*, vol. 9, no. 3, p. 155, 2018.

[38] I. H. Witten and E. Frank, Eds., *Data Mining:Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann, 2005.

[39] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artif. Intell.*, 1995, pp. 1137–1145.

[40] J. Che, L. Chen, Z.-H. Guo, S. Wang, and Aorigele, "Drug target group prediction with multiple drug networks," *Combinat. Chem. High Throughput Screening*, vol. 23, no. 4, pp. 274–284, 2020.

[41] J.-P. Zhou, L. Chen, and Z.-H. Guo, "iATC-NRAKEL: An efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs," *Bioinformatics*, vol. 36, pp. 1391–1396, 2020.

[42] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica Biophys. Acta–Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.

[43] H. Cui and L. Chen, "A binary classifier for the prediction of EC numbers of enzymes," *Current Proteomics*, vol. 16, no. 5, pp. 383–391, Jul. 2019.

[44] X. Zhao, L. Chen, Z.-H. Guo, and T. Liu, "Predicting drug side effects with compact integration of heterogeneous networks," *Current Bioinf.*, vol. 14, no. 8, pp. 709–720, Dec. 2019.

[45] X. Fan and L. Kurgan, "Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus," *J. Biomol. Struct. Dyn.*, vol. 32, no. 3, pp. 448–464, 2014.

[46] L. Chen, C. Chu, Y.-H. Zhang, M. Zheng, L. Zhu, X. Kong, and T. Huang, "Identification of drug-drug interactions using chemical interactions," *Current Bioinf.*, vol. 12, no. 6, pp. 526–534, Dec. 2017.

[47] R. Zhao, L. Chen, B. Zhou, Z.-H. Guo, S. Wang, and Aorigele, "Recognizing novel tumor suppressor genes using a network machine learning strategy," *IEEE Access*, vol. 7, pp. 155002–155013, 2019.

[48] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[49] X. Zhang, L. Chen, Z.-H. Guo, and H. Liang, "Identification of human membrane protein types by incorporating network embedding methods," *IEEE Access*, vol. 7, pp. 140794–140805, 2019.

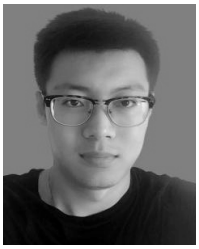
[50] J. C. McGiff, "Interactions of prostaglandins with the Kallikrein–Kinin and Renin-angiotensin systems," *Clin. Sci.*, vol. 59, no. s6, pp. 105s–116s, Jan. 1980.

[51] J. J. Kim, T. Kurita, and S. E. Bulun, "Progesterone action in endometrial cancer, endometriosis, uterine fibroids, and breast cancer," *Endocrine Rev.*, vol. 34, no. 1, pp. 130–162, Feb. 2013.

[52] M. M. Shapiro, V. Chakravartty, and J. E. Cronan, "Remarkable diversity in the enzymes catalyzing the last step in synthesis of the pimelate moiety of biotin," *PLoS ONE*, vol. 7, no. 11, 2012, Art. no. e49440.
[53] M. Prakash, K. Mathivon, D. M. Benoit, G. Chambaud, and M. Hochlaf, "Carbon dioxide interaction with isolated imidazole or attached on gold clusters and surface: Competition between σ H-bond and π stacking interaction," *Phys. Chem. Chem. Phys.*, vol. 16, no. 24, pp. 12503–12509, Jun. 2014.
[54] Y. Song, J. Zhou, Q. Li, A. Lue, and L. Zhang, "Solution properties of the acrylamide-modified cellulose polyelectrolytes in aqueous solutions," *Carbohydrate Res.*, vol. 344, no. 11, pp. 1332–1339, Jul. 2009.
[55] M. Iwaoka and N. Isozumi, "Hypervalent nonbonded interactions of a divalent sulfur Atom. Implications in protein architecture and the functions," *Molecules*, vol. 17, no. 6, pp. 7266–7283, Jun. 2012.
[56] S. Sookoian, "Alanine and aspartate aminotransferase and glutamine-cycling pathway: Their roles in pathogenesis of metabolic syndrome," *World J. Gastroenterol.*, vol. 18, no. 29, p. 3775, 2012.
[57] D. B. Antcliffe, A. M. Wolfer, K. P. O'Dea, M. Takata, E. Holmes, and A. C. Gordon, "Profiling inflammatory markers in patients with pneumonia on intensive care," *Sci. Rep.*, vol. 8, no. 1, pp. 1–9, Dec. 2018.
[58] Y. Yan, G. K. Singh, F. Zhang, P. Wang, W. Liu, L. Zhong, and L. Yang, "Comparative study of normal and rheumatoid arthritis fibroblast-like synoviocytes proliferation under cyclic mechanical stretch: Role of prostaglandin E_2," *Connective Tissue Res.*, vol. 53, no. 3, pp. 246–254, Jun. 2012.
[59] S. V. Shishkina, A. I. Slabko, S. Berski, Z. Latajka, and O. V. Shishkin, "Tuning of character of the N–O bond in HONO from covalent to protocovalent by different types of intramolecular interactions," *J. Chem. Phys.*, vol. 139, no. 12, Sep. 2013, Art. no. 124308.
[60] G. Jacquemot, G. Maertens, and S. Canesi, "Isostrychnine synthesis mediated by hypervalent iodine reagent," *Chem.-A Eur. J.*, vol. 21, no. 21, pp. 7713–7715, May 2015.

LIUCUN ZHU was born in Yangzhou, Jiangsu, China. He received the B.S. degree in biophysics and the Ph.D. degree in biology from Nanjing University, Nanjing, China, in 2005 and 2010, respectively.

From 2010 to 2014, he was a Lecturer with the Institute of System Biology, Shanghai University. In 2015, he turned to work with the School of Life Science. In 2018, he had been promoted to Associate Professor. His research interests include bioinformatics, computational biology, population genetics, and big data research.



PENGFEI HUANG was born in Weifang, Shandong, China, in 1993. He received the B.S. degree in biotechnology from Weifang Medical University, Weifang, Shandong, China, in 2017. He is currently pursuing the M.S. degree in bioinformatics with the School of Life Sciences, Shanghai University, Shanghai, China.

His research interest includes the detection and application of lncRNA, and the application of machine learning in biomedicine. He was a recipient of the Excellent Graduation Thesis of Shandong Province with the paper Establishment of triple Q-PCR detection system for animal food/feed identification, in 2018.

RUI ZHU was born in Lu'an, Anhui, China, in 1994. She received the B.S. degree in information and computing science from Shanghai University, Shanghai, China, in 2016, where she is currently pursuing the Ph.D. degree in mathematics with the Department of Mathematics.

Her research interests include bioinformatics, network biology, computational biology, machine learning, and graph theory.

FANGXIA GUAN was born in Luoyang, Henan, China, in 1969. She received the M.S. degree in clinical medicine from Henan Medical University, Zhengzhou, China, in 1996, and the M.D. degree in clinical medicine from Shandong Medical University, Jinan, Shandong, China, in 1999.

From 1999 to 2006, she has worked with the Albert Einstein Medical College, USA. Since 2006, she has been a Professor with the School of Life Science, Zhengzhou University. She has successively served as the Vice-President with the School of Life Sciences, Zhengzhou University, and the Henan Academy of Medical and Pharmaceutical Sciences. She is the author of two books, more than 100 articles, and 15 inventions. Her research interests included the development and transformation of new technologies for diagnosis and treatment in the field of biomedicine, and the pathogenesis and transformation of prevention with clinical diseases.

Prof. Guan is a member of the American Society for Science and the American Society for Neuroscience. She was a recipient of special government allowances of the State Council, the new century excellent talent in university of Ministry of Education of China, and the leading talent of science and technology innovation in Henan province. She has won three second prizes of science and technology progress of Henan Province. She is also the Chief Editor of *Journal Henan Medical Research*.

WENNA GUO was born in Zhoukou, Henan, China. She received the B.S. degree in mathematics and applied mathematics from Shangqiu Normal University, Shangqiu, Henan, China, in 2013, the M.S. degree in system analysis and integration from Shanghai University, Shanghai, China, in 2016, and the Ph.D. degree in bioinformatics from Nanjing University, Nanjing, China, in 2019.

She is currently working with Zhengzhou University, Zhengzhou, Henan. She is the author of 13 articles. She has found a four-DNA methylation prognostic biomarker of cutaneous melanoma through computer prediction combined with bioinformatics analysis, and the results published in the journals of *ELIFE*. Her research interests include bioinformatics, big data analysis, mathematical modeling, and the pathogenesis and prognosis of cancer.

Dr. Guo was a recipient of the National Postgraduate Scholarship and Outstanding Graduate of Nanjing University.