Molecular Therapy
Nucleic Acids
Original Article

Conjoint Feature Representation of GO and Protein Sequence for PPI Prediction Based on an Inception RNN Attention Network

Lingling Zhao,[1] Junjie Wang,[1] Yang Hu,[2] and Liang Cheng[3,4]

[1]Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China; [2]Department of Computer Science, School of Life Science and Technology, Harbin Institute of Technology, Harbin 150001, China; [3]NHC and CAMS Key Laboratory of Molecular Probe and Targeted Theranostics, Harbin Medical University, Harbin 150028, Heilongjiang, China; [4]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, Heilongjiang, China

Protein-protein interactions (PPIs) are pivotal for cellular functions and biological processes. In the past years, computational methods using amino acid sequences and gene ontology (GO) annotations of proteins for prioritizing PPIs have provided important references for biological experiments in the wet lab. Despite the current success, sequence information and ontological annotation in semantic representation have not been integrated into current methods. We propose a deep-learning-based PPI prediction methodology conjointly featuring sequence information and GO annotation. First, we adopt a word-embedding tool, the NCBI-blueBERT model pre-trained on PubMed, to map the GO terms into their semantic vectors. Then, the GO semantic vectors and protein sequence vector serve as the input of the proposed inception recurrent neural network (RNN) attention network (IRAN). The IRAN captures the spatial relationship and the potential sequential feature of the protein sequence and ontological annotation semantics. The extensive experimental results on 12 benchmarks demonstrate that our method achieves superiority over state-of-the-art baselines. In the yeast dataset of a binary PPI prediction, our method improved the performance with the Matthews correlation coefficient increasing from 94.2% to 98.2% and the accuracy from 97.1% to 98.2%. The analogous results were also obtained in other comparison evaluations.

INTRODUCTION

Protein-protein interactions (PPIs) drive numerous molecular processes and cellular activities, such as differentiation and cell-cell communication. The identification and characterization of PPIs are considered essential to understand the mechanisms of biological processes. Toward this end, various high- or low-throughput techniques have been used. In this manner, numerous direct PPIs have been elaborated, yielding the well-known databases, including the Database of Interacting Proteins (DIP),[1] IntAct molecular interaction database (IntAct),[2] and Molecular INTeraction database (MINT).[3] However, due to the vast combinatorial space of protein interactions, wet experiment-based identification has only limited coverage at a relatively high and time-consuming cost. Fortunately, these collected PPIs have formed ever-increasing and constantly updated datasets that cover various species, thereby promoting the emergence of the predictive computational models of PPIs. Guided by an inexpensive identification of PPI candidates through computational screening, the costs and failure rates of experimental methods can be reduced significantly.

Two primary issues have been addressed in the computational methods of PPI prediction: the feature representation of proteins and classification algorithm over the feature space. Protein domains, protein structure information, gene neighborhood, gene fusion, coevolution of proteins, and phylogenetic profiles have been attempted as the PPI descriptors.[4] Nevertheless, these pieces of information are not always available, and such unavailability restricts their application. Sequence-based approaches have become an active research area due to the explosive growth of sequence data. Various predefined features can be extracted from protein sequences, such as autocovariance (AC),[5] auto-cross covariance (ACC),[5] composition-transition-distribution (CTD) descriptors,[6] local protein sequence descriptors (LDs),[7] or other numerical features. Although various descriptors have been retrieved to represent the latent facets of PPIs, the coverage of these descriptors is still limited, and many research efforts have been made to overcome this restriction by using some complex learning strategies, such as the hybrid of descriptors, ensemble of classifiers,[8,9] or even the ensemble of distinct deep neural networks.[10]

Conversely, to avoid the hand-crafted feature engineering, deep learning methods have been successfully applied in the PPI predictions based on raw protein sequences and other prediction tasks. Different neural network architectures have been attempted, leading

Received 11 May 2020; accepted 21 August 2020;
https://doi.org/10.1016/j.omtn.2020.08.025.

Correspondence: Yang Hu, Department of Computer Science, School of Life Science and Technology, Harbin Institute of Technology, Harbin 150001, China.
E-mail: huyang@hit.edu.cn
Correspondence: Liang Cheng, NHC and CAMS Key Laboratory of Molecular Probe and Targeted Theranostics, Harbin Medical University, Harbin 150028, Heilongjiang, China.
E-mail: liangcheng@hrbmu.edu.cn

Table 1. Summary of the Benchmark Datasets

Dataset	Species	No. of Proteins	Size (Positive/ Negative)
SC full	*Saccharomyces cerevisiae*	4,424	17,257/48,594
SC balanced	*Saccharomyces cerevisiae*	4,424	17,257/17,257
Yeast	yeast	2,497	5,594/5,594
MM	*Mus musculus*	1,088	500/500
AT	*Arabidopsis thaliana*	756	541/541
SP	*Schizosaccharomyces pombe*	904	742/742
DM	*Drosophila melanogaster*	658	321/321
EC	*Escherichia coli*	589	1,167/1,167

to a series of raw sequence-based deep learning PPI predictors: DeepPPI[11] and DNNPPI[12] seek to retrieve the latent feature from the deep neural networks (DNNs); Pei and colleagues[13] have investigated a stacked autoencoder (SAE) for PPI prediction; DPPI[14] and Wang et al.'s[15] study both use the convolution neural networks (CNNs), capturing the spatial features of the input sequence; DeepSequencePPI[16] introduced recurrent neural networks (RNNs) for the first time by considering the ordering information of protein sequences; and PIPR[17] proposed in 2019 developed a Siamese residual network incorporating a residual RNN and CNN to leverage both the local spatial features and contextualized information of protein sequences. It is notable that these models achieve promising predicting performances despite relying only on the amino acid sequence, demonstrating the powerful feature representation and learning capacity of deep learning.

Gene ontology (GO) annotation is another characterization method for proteins. The GO is a hierarchical vocabulary for annotating gene or gene product functions and their relationships with respect to molecular function (MF), cellular component (CC), and biological process (BP).[18,19] It is known that proteins interact with each other in protein complexes or functional modules, and a protein complex refers to a group of interacting proteins at the same cellular location, whereas functional modules consist of proteins that commonly participate in a particular cellular process or molecular function at a different time and place. These two protein interactions are both closely related to the annotation of GO. Therefore, GO annotation is considered as one of the most important indicators to characterize PPIs, and several studies have inferred the interaction possibility by evaluating the semantic similarity of GO annotations to proteins. Conventionally, the semantic similarity is quantified by the number of sharing GO terms annotating to the compared proteins.[20] This scheme provides a simple but rough estimation to the semantic similarity of GO annotation. To overcome the limitation, several studies have focused on the semantic retrieval from GO-directed acyclic graphs (DAGs), such as go2ppi[21] and PPI-MetaGo,[16] which represent the protein's GO annotation according to the hierarchical structure of the GO DAG. Meanwhile, development of the language representation techniques has significantly inspired biomedical fields to mine the semantics of entities from numerous available textual resources or self-defined corpora. Word embedding (WE), especially word2vec[22] proposed in 2013, has achieved great success and yielded a series of biomedical semantic similarity measures, including Onto2Vec[23] and OPA2Vec.[24] However, because word2vec is not sensitive to context, it cannot distinguish the different semantics of polysemous words. To overcome the drawback, contextualized representation models, including ELMO[25] and more recently Google's BERT,[26] have been proposed since 2018. They are designed using whole sentences as the context, thereby resulting in better polysemy and nuance handling. In addition, by applying the transfer learning technique, two BERT versions, NCBI-blueBERT[27] and clinic-BERT,[28] have been trained on a textual database, such as PubMed abstracts (https://www.ncbi.nlm.nih.gov/pubmed/), for specific domain tasks.

In this study, we approached the problem of PPI prediction by the combination of protein sequence and GO annotation, and propose a DNN to encode both the sequential and spatial features of protein primary sequences and the embedding vector of GO annotation (GOSeqPPI for short). Unlike the other NLP-based methods, which utilize the word2vec[23] model to embed the GO annotation to a set of vectors, we used a pre-trained BERT to project the ontology annotation to a high-dimensional vector space. Compared with word2vec, BERT is contextual word-embedding (CWE), which means that BERT can encode each word in an ontology term dependent on the context. Moreover, BERT does not need to maintain a large vocabulary, thereby avoiding the out-of-vocabulary words problem. Furthermore, the BERT model was more capable of understanding the true semantic meaning of the text by presenting the state-of-the-art results of various NLP tasks, including question answering and natural language inference. In the present study, we conducted four experiments to evaluate the performance of the proposed method. First, we verified the effectiveness of the joint feature representation in PPI prediction and further evaluated our model on various binary PPI prediction tasks utilizing various datasets. Then, an evaluation on two PPI network prediction tasks was executed to access the generalization ability. Finally, a challenging multiclass classification on PPI type was examined on both GOSeqPPI and the state-of-the-art methods.

RESULTS

Datasets

The following three types of binary PPI datasets were utilized in this study: the *Saccharomyces cerevisiae* (SC) full dataset with large but imbalanced samples, two datasets with balanced large samples including the yeast dataset and the pruned SC full dataset, and five datasets with small samples. These datasets cover various species and follow different sample distributions and thus can adequately validate the performance of GOSeqPPI and the baseline methods. Their sizes and descriptions are summarized in Table 1. To keep an unbiased and consistent test, in the comparison experiment with PPI-MetaGo and go2ppi, we utilized the same GO and GO annotation datasets as theirs. In other evaluations, the GO and GO annotation data were

Table 2. Results on the Balanced SC Dataset

Methods	Prec	Accu	Sen	Spec	F-Score	MCC	AUC
DeepPPI	0.9493	0.9254	0.9005	0.9508	–	0.8520	0.9754
EnsDNN	0.9545	0.9529	0.9512	0.9548	0.9529	0.9059	0.97
SeqPPI	0.938	0.926	0.912	0.939	0.925	0.851	0.978
GOSeqPPI	0.961	0.956	0.949	0.962	0.955	0.912	0.988

from QuickGO at https://www.ebi.ac.uk/QuickGO/. GO release version 2019-11-27 and GO annotation 2019-11-25 were used.

The Evaluation of Joint Feature Representation

To access the conjoint feature representation capacity, especially the impact of semantic feature of GO annotation on the entire protein, we constructed an only-sequence-feature-based PPI (SeqPPI) predictor for comparison with GOSeqPPI as well as two other baseline models, DeepPPI and EnsDNN, on the balanced SC datasets. The SeqPPI predictor shares the same network architecture with the sequence block of GOSeqPPI. This means that SeqPPI can be considered as a sequence-feature classifier separated from the GOSeqPPI model; therefore, the comparison between SeqPPI and GOSeqPPI can indicate the effect of GO annotation on the prediction performance. Similar to SeqPPI, DeepPPI designed a DNN to learn the representations of proteins only from protein descriptors. Another sequence-based DNN predictor, EnsDNN, enhances the classification capacity by integrating 27 DNNs to leverage complementary information about several descriptors of protein sequences. To keep the consistence with the baseline predictors, we constructed a 4-fold cross-validation (CV) utilizing the aforementioned five balanced SC datasets on which DeepPPI and EnsDNN had previously been trained and tested. The average results are listed in Table 2.

It can be observed that the performances of SeqPPI and DeepPPI were similar in terms of almost all measures. This might be due to the fact that both of them use the raw protein sequence as a feature and a DNN for feature representation and classification. In addition, the tested dataset is balanced and of large size, which is sufficient to train the models efficiently, resulting in a close learning ability. In contrast, EnsDNN had a better performance than SeqPPI or DeepPPI. The superiority could be attributable to the ensemble strategy of numerous different DNNs and the inclusion of the handcraft features from sequence. Although GOSeqPPI only develops one DNN, it performs best in terms of the precision (Prec), accuracy (Accu), specificity (Spec), Matthews correlation coefficient (MCC), and area under receiver operating characteristic curve (AUC) relative to the other predictors. According to the comparison with SeqPPI, Accu increased with 3%, Prec 2.4%, Sen 3.7%, Spec 2.3%, F-score 3%, MCC 6.3%, and AUC 0.9%, suggesting that inclusion of the GO annotation semantic feature significantly improves prediction performance.

Binary PPI Prediction

In this experiment, the imbalanced SC full dataset, the balanced yeast dataset, and the multiple series small-scale datasets were utilized to assess the robustness of GOSeqPPI. We conducted a 10-fold CV to obtain the prediction results and selected earlier trained and tested models on these datasets as baseline approaches, including PPI-MetaGo, go2ppi-RF, DPPI, DeepSequencePPI, and PIPR. DeepSequencePPI is also a sequence-based method with a deep learner consisting of gated recurrent units (GRU), thus adapting itself in the condition of large samples. Aside from the sequence-based deep learners, such as DeepPPI, DPPI, or EnsDNN, we also selected PPI-MetaGo and go2ppi-RF for comparison. go2ppi-RF is a classic GO-driven PPI predictor relying on merely GO annotation, whereas PPI-MetaGo combines sequence-based and GO-based features with an ensemble stacked generalization.

Table 3 illustrates the details of the seven metrics on these datasets. As shown in Table 3, the GOSeqPPI provided the best results on the SC full, *Escherichia coli* (EC), *Arabidopsis thaliana* (AT), *Schizosaccharomyces pombe* (SP), and *Mus musculus* (MM) datasets in terms of all of the metrics, and the least Accu improvement was reached with 2.2% on the EC dataset, 9.4% on the AT dataset, 3.1% on the SP dataset, 2% on the *Drosophila melanogaster* (DM) dataset, and 3% on the SC full dataset. On the DM dataset, the Accu, Sen, F-score, MCC, and AUC were significantly higher than those in GOSeqPPI or PPI-MetaGo, whereas the Prec and Spec were slightly lower than in PPI-MetaGo, indicating that GOSeqPPI performs better than PPI-MetaGo in the identification of the interacted protein pairs, but somewhat worse in the non-interacted protein pairs, probably because the DM dataset has only 642 protein pairs, which cannot engage the training of the deep network in GOSeqPPI. Meanwhile, the notable superiority to PPI-MetaGo and DeepSequencePPI in the distinct cases suggests the effectiveness of GO semantic embedding based on BERT and the temporal-spatial feature representation by the inception RNN.

On the yeast dataset, GOSeqPPI also exhibited better prediction performance over PIPR and DPPI, both of which are sequence driven and deep learning based. PIPR incorporates a well-designed deep residual recurrent convolutional neural network in the Siamese architecture, and, to our best knowledge, PIPR is the state-of-the-art method in the binary PPI prediction and provides the best prediction results on the yeast dataset. The GOSeqPPI outperformed PIPR with Accu from 97.1% to 98.2%, Sen from 97.2% to 99.8% (almost all of the positive samples were correctly classified), F-score from 97.1% to 98.3%, and MCC from 94.2% to 98.2%, but it was slightly inferior to PIPR in terms of Prec (from 97% to 96.8%) and Spec (from 97.0% to 96.8%). The comparison results demonstrate that GOSeqPPI has better identification ability on the interacted protein pairs and

Table 3. Comparison Results of Proposed Models and Baselines

Dataset	Methods	Prec	Accu	Sen	Spec	F-Score	MCC	AUC
SC	PPI-MetaGo	0.934	0.924	0.912	0.936	0.923	0.848	0.972
Full	DeepSequencePPI	0.942	0.932	0.920	0.922	0.931	0.864	0.978
	GOSeqPPI	0.902	0.960	0.950	0.963	0.925	0.899	0.989
Yeast	DPPI	0.967	0.946	0.922	–	0.944	–	–
	PIPR	0.970	0.971	0.972	0.970	0.971	0.942	–
	GOSeqPPI	0.968	0.982	0.998	0.968	0.983	0.982	0.996
EC	PPI-MetaGo	0.922	0.902	0.879	0.925	0.900	0.805	0.950
	go2ppi-RF	0.937	0.905	0.869	0.941	0.902	0.813	0.951
	GOSeqPPI	0.953	0.925	0.895	0.954	0.921	0.854	0.975
AT	PPI-MetaGo	0.830	0.808	0.778	0.837	0.801	0.619	0.866
	go2ppi-RF	0.875	0.789	0.684	0.895	0.764	0.596	0.810
	GOSeqPPI	0.901	0.884	0.864	0.903	0.881	0.769	0.934
SP	PPI-MetaGo	0.935	0.929	0.922	0.935	0.928	0.858	0.965
	go2ppi-RF	0.901	0.885	0.865	0.904	0.882	0.771	0.941
	GOSeqPPI	0.955	0.958	0.961	0.955	0.957	0.916	0.981
DM	PPI-MetaGo	0.885	0.869	0.857	0.882	0.867	0.744	0.916
	go2ppi-RF	0.853	0.843	0.832	0.854	0.841	0.688	0.889
	GOSeqPPI	0.881	0.887	0.906	0.870	0.886	0.783	0.950
MM	PPI-MetaGho	0.808	0.786	0.754	0.818	0.779	0.575	0.860
	go2ppi-RF	0.836	0.738	0.604	0.812	0.695	0.500	0.762
	GOSeqPPI	0.905	0.868	0.822	0.912	0.855	0.744	0.933

comparable negative classification ability to PIPR, and this might be because of the joint use of the GO-annotation semantic features in GOSeqPPI.

The PPI Network Prediction

The PPI network enables the understanding of how cell life works by identifying protein complexes and their functions, as well as the protein classification. PPI prediction plays a crucial role in the construction of a PPI network. In this study, we applied GOSeqPPI in two kinds of PPI network prediction, a one-core PPI network of CD9 and a crossover network of the Wnt-related network. A one-core PPI network contains only a core protein and multiple attached proteins.

To evaluate the generalization of our method, we utilized two independent datasets to train and test the GOSeqPPI, respectively; that is, we trained GOSeqPPI based on the balanced SC dataset and then applied the trained GOSeqPPI to infer the interaction pairs in the CD9 and Wnt-related networks of humans. As the core of a typical one-core network, CD9 is an important tetraspanin protein and interacts with 16 associated proteins. The prediction results revealed that all of the PPI pairs in this network were identified by our method (Figure 1). In contrast, LightGBM-PPI hit 14 PPIs, and Shen's work[32] hit 13 PPIs.

The Wnt-related pathway is indispensable in signal transduction, and the predicted components of this pathway have been verified with the yeast II hybrid experiments, comprising 96 interacting pairs. The prediction results marked in the Wnt-related network are shown in Figure 2, where the blue and red lines indicate true and false prediction, respectively. Among the 96 PPI pairs, only the protein pair of CER1-WNT4 was missing in GOSeqPPI, which significantly outperforms Zhou's[9], Shen's,[32] and Ding's[33] work alongside LightGBM with accuracies of 87/96, 73/96, 91/96, and 89/96), respectively.

Multi-class PPI Prediction

Aside from determining whether two proteins interact, another meaningful task about PPI prediction is to examine the type of the interaction between two proteins. Following the evaluation used in the work of PIPR, we used the same experimental configurations and the same datasets SHS27k and SHS148k by a 10-fold CV. To fit the multiclass classification task, the binary output of the network in GOSeqPPI was replaced with a multiple dimensional vector, which indicates the sample probability belonging to each category. There are seven interaction types in the datasets SHS27k and SHS148k: reaction, binding, ptmod, activation, inhibition, catalysis, and expression. The type distribution of the samples in the two datasets is shown in Figures 3 and 4. More detailed information about the datasets can be found in Table 4, where no. of PNA denotes the number of proteins with no GO annotations (PNA), and no. of PPNA denotes the number of protein pairs involved at least one protein with no GO annotations (PPNA). Note that 20 proteins in the SHS148k dataset did not have any GO annotation, and they were involved in 655 protein pairs,

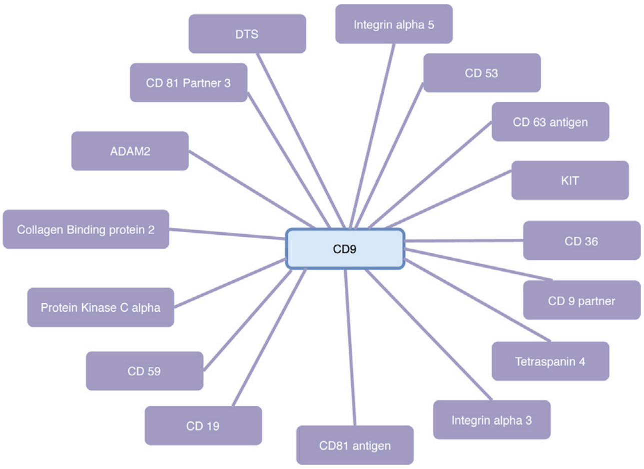

Figure 1. The Predicted CD9 PPI Network

meaning that their GO annotation semantic feature cannot be represented. For these proteins, we filled their GO annotation semantic matrices with randomly generated data following a Gaussian distribution N(0,1).

It can be observed for both SHS27k and SHS148k that the sample distribution on the interaction type was nearly balanced except the type "expression." On this multiclass classification task where few PPI predictors were attempted, PIPR outperformed all of the baselines according to the CTD and AC descriptors and different machine learning models, including support vector machine (SVM), random forest, AdaBoost, k-nearest neighbor (KNN), and logistic regression. Therefore, in this study, we compared GOSeqPPI with the state-of-the-art PIPR in terms of accuracy and report our results in Table 5. Among the baselines listed, rand means random guessing, and zero rule refers to predicting the majority class. The prediction accuracies of the tested models were much impaired in comparison with the ones on the binary PPI prediction tasks. The main reasons for this observation lie in the challenge of multiclass classification. GOSeqPPI achieved the best accuracy on SHS27k dataset, and the accuracy promotion relative to PIPR was 2.6%, whereas on the SHS148k dataset, GOSeqPPI slightly outperformed PIPR. In addition, it can be observed that both performances of PIPR and GOSeqPPI improved when the size of the dataset increased, but the promotion rate of GOSeqPPI (1.3%) was lower than that of PIPR (3.9%). The reason for this observation lies in the inevitable deviation caused by the random semantic feature adopted when dealing with the absence of GO annotation. This influence results in the decline of the prediction performance of our model, but also indicates that the PPI prediction can benefit from the inclusion of the semantic feature of GO annotation. Additionally, although feature noise existed in both the training and testing processes, GOSeqPPI still provided comparable multi-category prediction to PIPR, again exhibiting the robustness of the whole methodology.

Protein Interactions and Disease

In this subsection, we explore the application of protein interactions as a translational approach to the study of human diseases. PPIs are also involved in the mechanisms leading to healthy or diseased states in organisms due to the central role of such interactions in biological processes. Diseases are often caused by mutations affecting the binding interfaces of, or leading to, biochemically dysfunctional allosteric changes in proteins.[29,30] Therefore, candidate proteins that presumably cause a certain disease can be in turn inferred by the known related proteins. Based on this judgment, we constructed a human protein-disease dataset along with 8,127 protein pairs, which were selected from the manually curated dataset of InnateDB (https://www.innatedb.com) by removing the protein pairs involving a protein or two proteins without any related disease. Then, the GOseqPPI model was trained and tested on this PPI dataset, where the training dataset contained 7,552 samples and the testing dataset had 575 samples. Note that there are 4,352 positive and 3,775 negative PPI samples, and this distribution can be considered approximately balanced.

According to the PPI testing results, we marked the protein pairs in which the proteins had at least one common related disease. To analyze the relationship between PPI and their common diseases, we prioritized protein pairs in accordance with their PPI values and counted the number of protein pairs with common diseases (PPCDs) distributed in 14 zones, including the zone of protein pairs with the top 5, 10, 20, 50, 100, 200, and 250 predicted PPI values (marked by top x for short), and the zone of protein pairs with the bottom 5, 10, 20, 50, 100, 200, and 250 predicted PPI values (marked by bottom x for short). As reported in Table 6, in the first seven zones(top 5 to top 250), with decreasing predicted PPI values, the ratio of PPCDs decreased accordingly, and this trend was relatively stable. However, in the last seven zones (bottom 250 to bottom 5), the ratio of PPCDs decreased sharply. This difference is presumably because of the variation among the predicted PPI values of these zones. Although the dividing strategy remained the same in the top and bottom zones, the predicted PPI values in the top zones were closely distributed and were all larger than 0.8955, whereas the ones in the bottom zone had a significant difference (dropped from 0.6332 to 8.494e−5). Our analyses showed that the Pearson coefficient of the PPI values and the ratio of PPCDs was 0.9799 with a p value of 8.95e−10, demonstrating that there was strong relatedness between PPI and the ratio of PPCDs, and the PPI prediction could in turn be utilized to infer the candidate of related proteins of a given disease. For example, the proteins that interact with the related proteins of the disease will be recommended as the candidates, and the recommendation confidence can be prioritized in accordance with their PPI values with these known related proteins.

DISCUSSION

To date, numerous machine learning-based computational methods have been proposed to address PPI prediction tasks. However, there is still room to improve the prediction performance, especially the robustness, generalization, and precision. Meanwhile, a vast number of wet experiments collect and verify increasingly more protein

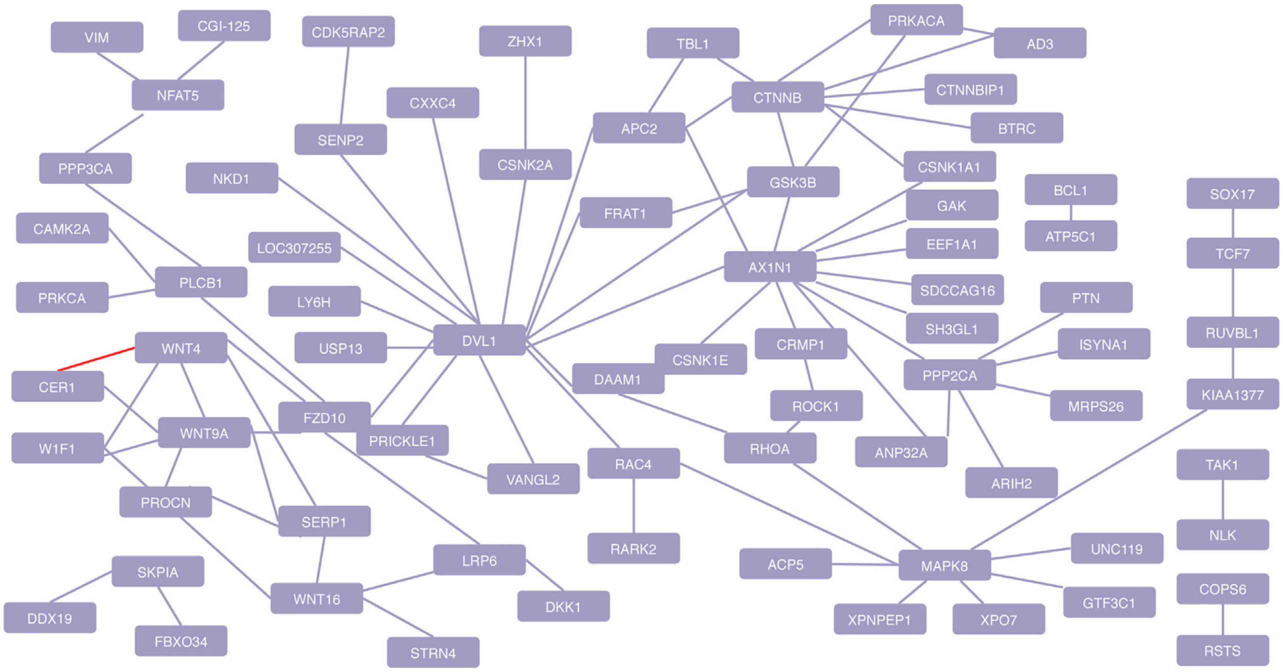

Figure 2. The Predicted Wnt-Related Pathway PPI Network

interactions, providing more high-quality and large-scale datasets for machine learning methods to optimize and adjust their models. In particular, a series of deep learning-based PPI methods have emerged and become gradually prominent due to their powerful representation and learning capacities from non-handcrafted raw information. It is notable that although feature engineering is not required in deep learning, feature selection for protein characterization is still crucial. In this study, we put forward GOSeqPPI, a deep-learning-based PPI prediction methodology with the conjoint feature of protein sequence and GO annotation semantic embedding. GO annotations and protein sequence share rather limited common information and are of low correlation, which enlarges the coverage of feature space and helps identify PPIs more efficiently. Instead of the GO annotation embedding based on GO DAG, we used the semantic encoding of GO annotation from NCBI-blueBERT pretrained by the corpus, including the PubMed database and clinic notes. The contextual sensitivity and focusing on the biomedical corpus enabled NCBI-blueBERT to capture more accurate semantic features for protein GO annotation. The proposed network incorporates an inception block and a binary directional GRU (Bi-GRU) block, capturing the potential sequential and spatial relationship in protein sequences as well as the semantic vectors of GO annotation. Subsequently, the global pooling-attention layer compresses the feature space and provides a compact global feature for each protein. The whole framework integrates the information originating from two different perspectives, protein physical sequence and human knowledge description, and projects them into an abstract feature space and reconstructs the proteins in this space by a nonlinear mapping trained by the proposed neural network.

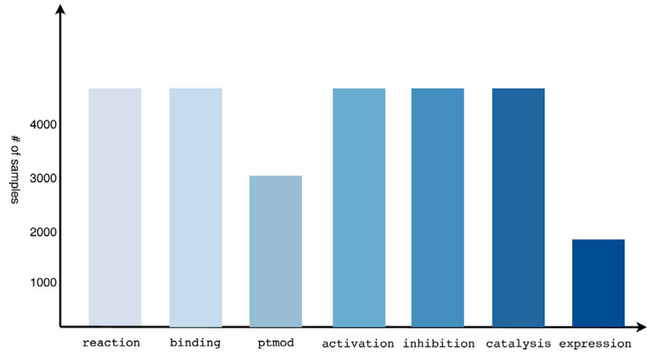

Figure 3. The Sample Distribution in Each Category on the SHS27k Dataset

Extensive evaluations on both the large-scale and small-scale as well as balanced and imbalanced datasets have been made, and the experimental results demonstrate that our method retains remarkable superiority over the baselines on the binary PPI classification, the multi-categories PPI type classification task, and PPI network prediction. Notably, although on even the small-size datasets, such as the DM with only 642 protein pairs, GOSeqPPI outperforms the baselines, and more samples will still be more conducive to the performance amelioration resulting from more training. In addition, the hybrid feature that integrates the information originating from both the physical and human knowledge spaces, in particular, the raw sequence and GO annotation semantic to characterize proteins, can be applied to portray other biomedical entities or concepts, such as compounds and genes. Moreover, the proposed

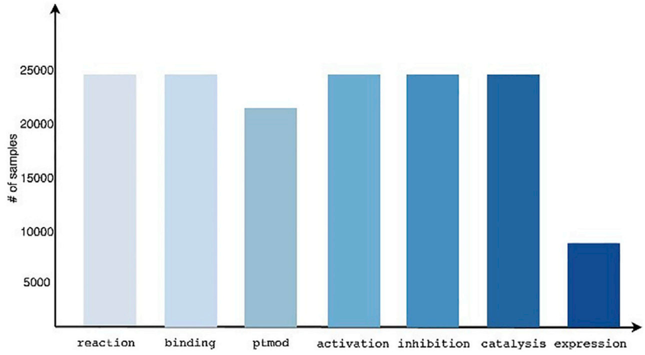

Figure 4. The Sample Distribution on Each Category on the SHS148k Dataset

inception RNN attention network can also be used in other biological downstream classification or regression tasks[31].

MATERIALS AND METHODS

In this section we present a GO-aided PPI prediction method. Figure 5 provides an overview of the entire framework, which consists of the following five layers: input layer, pre-encoding layer, feature representation layer, classification layer, and output layer. Protein sequence and GO annotations for each protein are at the bottom of this framework, and then the protein sequence is encoded by 0-1 one-hot encoding, while GO annotations are embedded by the BERT model. Both types of pre-encoded features are re-encoded by the inception RNN attention network (IRAN) deep network, resulting in a concatenated feature vector with the structural, functional, and cellular location information about each protein. A full-linked block works on the top of feature representation and makes a classification for PPI or PPI type, or performs a regression task, such as affinity prediction.

Input Representation

In our proposed framework, the raw amino acid sequence and GO annotation serve as the input of the entire framework, where the protein sequence is encoded by one-hot coding, and thus the sequence can be converted into a one-hot coding matrix $S = [s_1; s_2; \ldots, s_L]$, where L is the length of the raw amino acid sequence. For the GO annotation composed of GO terms, several types of information can be candidates to represent GO terms, namely, term name, description, and term ID. In this study, term names were used to represent GO annotations due to their dense information compared with description or term ID. Meanwhile, the conciseness of term names facilitates the feasible semantic embedding in the following pre-encoding layer. Let a protein P be annotated by GO terms $g_1; g; \ldots, g_M$, where M is the number of terms annotating this protein, then the GO annotations for any protein can be viewed as a set of GO terms $A = \{g_1; g_2; \ldots, g_M\}$. A GO term g is in the set A if and only if g is used to annotate P. Furthermore, since a term name generally consists of several words, g_i is described by a word sequence and denoted by $[w_1; w_2; \ldots, w_N]$, where N is the number of words involved in a term name.

Table 4. Summary of Datasets SHS27k and SHS148k

Dataset	No. of proteins	No. of Protein Pairs	No. of Proteins without GO Annotations	No. of Protein Pairs Involving One or Two Proteins without GO Annotations
SHS27k	1690	26944	0	0
SHS148k	5189	148050	20	655

The Pre-encoding of GO Annotation Based on BERT

To capture the semantic information from the GO annotation for each protein, we adopted a pre-trained BERT model to map each GO term into a high-dimension feature vector. In contrast to the state-of-the-art GO representation methods based on word2vec, we exploited the NCBI-blueBERT model specifically trained for the biomedical domain as the semantic embedding approach. BERT is a word-based natural language-processing model that was trained on the following two unsupervised prediction tasks: (1) the masked language model (MLM) task (15% of the words were masked), and (2) the next sentence prediction task. Specifically, given a term name g_i including N words, the BERT model first uses word piece tokenization, which means one word may break into several pieces. The aim of tokenization is to achieve a balance between vocabulary size and out-of-vocabulary words. Additionally, two special tokens, start [CLS] and end tokens [SEP], are added for each sentence. Finally, the term name g_i is represented by an ordering set of tokens $g_i = \{c_1, c_2, \ldots, c_N\}$. The input E_i provides the representation of each token c_i by summing the corresponding token, segment, and position embeddings. Further detailed information can be found in Devlin et al.[26] Through the word embedding, each term name is converted into a two-dimension vector $T_i = [v_1, v_2, \ldots, v_{ni}]$, where v_i denotes the vector converted from the i-th token (a word or a word piece) and n_i is the number of tokens after tokenization. Then, all of the terms annotating a protein form a semantic feature matrix $T = [T_1, T_2, \ldots, T_M]$ with a width of 768, which is fixed by the BERT model. Furthermore, to keep the same size for all proteins, the complementary data obtained by filling zeros are exploited to obtain the input of the same size. Here, we selected 256 or 512, according to the specific task, as the height of the semantic feature matrix. Then the final semantic feature matrix is $T' = [T_1, T_2, \ldots, T_M, [00\ldots0]\ldots, [00\ldots0]]$.

The Inception RNN Attention Network

In this study, we developed an IRAN to capture both the temporal and spatial properties of the heterogeneous presentation of proteins, and this network consists of three blocks, an inception-CNN block, a Bi-GRU block, and a global-pooling-attention layer. The whole architecture is described in Figure 6. The inception-CNN block is an ensemble of CNN that combines different sizes of convolution kernel and then concatenates the spatial features together. For the temporal feature extraction, we constructed a GRU-based block to learn the sequential feature of the protein sequence and GO term vector. Finally, an attention layer was linked on top of both the inception and GRU blocks, enabling the neural network to automatically pay attention to a certain part of the given features. On top of the attention layer, the

Table 5. Percentage Accuracy for PPI Interaction Type Prediction

Methods	Rand	Zero Rule	PIPR	GOSeqPPI
SHS27k	14.28	16.70	59.56	61.16
SHS148k	14.28	16.21	61.91	61.99

model is completed with one fully connected (FC) layer to accomplish the classification.

The Inception-CNN Branch

CNN applies multiple convolution kernels to capture local features from a sliding window; nevertheless, deep CNN suffers from vanishing and exploding gradients and is prone to overfitting. To overcome these shortcomings, we used an inception network to leverage the spatial characterizations of input vectors. The inception block consists of the following convolution operations in a "shallow" mode: (1) a 1D convolution kernel with size …; (2) a 1D convolution kernel is stacked after a 1D convolution kernel; (3) a 1D convolution kernel following a 1D convolution kernel; and (4) a 1D convolution kernel. These different convolution kernel sizes ensure that both the "sparse" and "non-sparse" features, thereby increasing the width of the network and the adaptability of the network to scale. By stacking two 1 convolution kernels before the 3 convolution and 5 convolution kernels, the network parameters are reduced significantly. The architecture of the inception network is shown in Figure 6.

If the pre-coded input (either the protein one-hot matrix S' or the semantic embedding matrix T' of the GO annotation) is uniformly denoted by X, then through the inception-CNN unit, four middle-layer feature representations can be calculated by the Equations 1–4 as follows, where the hidden states h_1^k, h_3^k, and h_5^k correspond to the output of the convolution operations 1–4, and w and b represent the weight and bias, respectively:

$$\mathrm{h}_1^{\mathrm{k}} = \mathrm{relu}\left(\mathrm{w}_1 \times \mathrm{x} + \mathrm{b}_1^{\mathrm{k}}\right) \tag{1}$$

$$\mathrm{h}_5^{\mathrm{k}} = \mathrm{relu}\left(\mathrm{w}_3 \times \left(\mathrm{relu}\left(\mathrm{w}_1 \times \mathrm{x} + \mathrm{b}_1^{\mathrm{k}}\right)\right) + \mathrm{b}_3^{\mathrm{k}}\right) \tag{2}$$

$$\mathrm{h}_5^{\mathrm{k}} = \mathrm{relu}\left(\mathrm{w}_5 \times \left(\mathrm{relu}\left(\mathrm{w}_1 \times \mathrm{x} + \mathrm{b}_1^{\mathrm{k}}\right)\right) + \mathrm{b}_5^{\mathrm{k}}\right) \tag{3}$$

$$\mathrm{h}_3^{\mathrm{k}} = \mathrm{relu}\left(\mathrm{w}_3 \times \mathrm{x} + \mathrm{b}_3^{\mathrm{k}}\right). \tag{4}$$

Table 6. Accuracy for PPI Interaction Type Prediction

PPI zone	Predicted PPI Values	No. of Protein Pairs with Common Diseases	Ratio of Protein Pairs with Common Diseases (%)
Top 5	0.9986–0.9995	5	100
Top 10	0.9977–0.9995	9	90
Top 20	0.9969–0.9995	18	90
Top 50	0.9937–0.9995	45	90
Top 100	0.9849–0.9995	92	92
Top 200	0.9432–0.9995	170	85
Top 250	0.8955–0.9995	211	84.4
Bottom 250	8.494e−5 to 0.6332	103	41.2
Bottom 200	8.494e−5 to 0.2279	71	35.5
Bottom 100	8.494e−5 to 0.0142	28	28
Bottom 50	8.494e−5 to 0.0046	13	26
Bottom 20	8.494e−5 to 0.0014	3	15
Bottom 10	8.494e−5 to 0.0005	1	10
Bottom 5	8.494e−5 to 0.0002	0	0

Bidirectional GRU Branch

As a variant of long short-term memory (LSTM), the GRU recurrently encodes the input data according to their ordering relationship and retains important features through updateu_k and reset gates r_k to ensure that they will not be lost during long-term propagation. For the k-th embedding vector v_k, GRU computes the hidden stateh_k along with the previous stateh_k according to Equations 5–8:

$$r_k = \sigma(W_r v_k + U_r h_{k-1} + b_r) \tag{5}$$

$$u_k = \sigma(W_u v_k + U_u h_{k-1} + b_u) \tag{6}$$

$$\tilde{h}_k = tanh(W_c v_k + U(r_k \odot h_{k-1})) \tag{7}$$

$$h_k = (1 - u_k) \odot h_{k-1} + u_k \odot \tilde{h}_k, \tag{8}$$

where W_r, W_u, W_u and U_r, U_u, U denote the weight matrices that would be learned in the training of the GRU; b_r and b_u are biases for the reset gate and the update gate, respectively; σ represents the sigmoid function; and $\odot$denotes element-wise production. In the GRU network, the state transmission is unidirectional from the front to the back. Bi-GRU is composed of two GRUs superimposed together to capture the influence on the current hidden state h_k from both the forward and backward directions. If the forward GRU encodes the sequence $v_1, \ldots, v_l$ and the hidden state at k-th position is denoted as $\overleftarrow{h}_k$ while the backward GRU encodes the sequence $v_l, \ldots, v_1$ and the output is $\overrightarrow{h}_k$, then the resultant output of the Bi-GRU k-th position is the concatenation $[\overleftarrow{h}_k, \overrightarrow{h}_k]$.

Multi-level Feature Representation

For each protein, the inception CNN and Bi-GRU branches execute a series of operations over the GO-embedding and protein sequence-encoding matrices, respectively, and then each branch generates four different local feature representations. After the inception of CNN and Bi-GRU blocks, and considering that different global pooling strategies tend to perceive different features from the inception CNN or Bi-GRU blocks, we utilized a global max pooling, global average pooling, and the attention mechanism to produce global feature representational vectors f_{maxp}, f_{avep}, and f_{att}, respectively, and then concatenated these three features into an entire global feature $F = [f_{maxp}, f_{avep}, f_{att}]$.

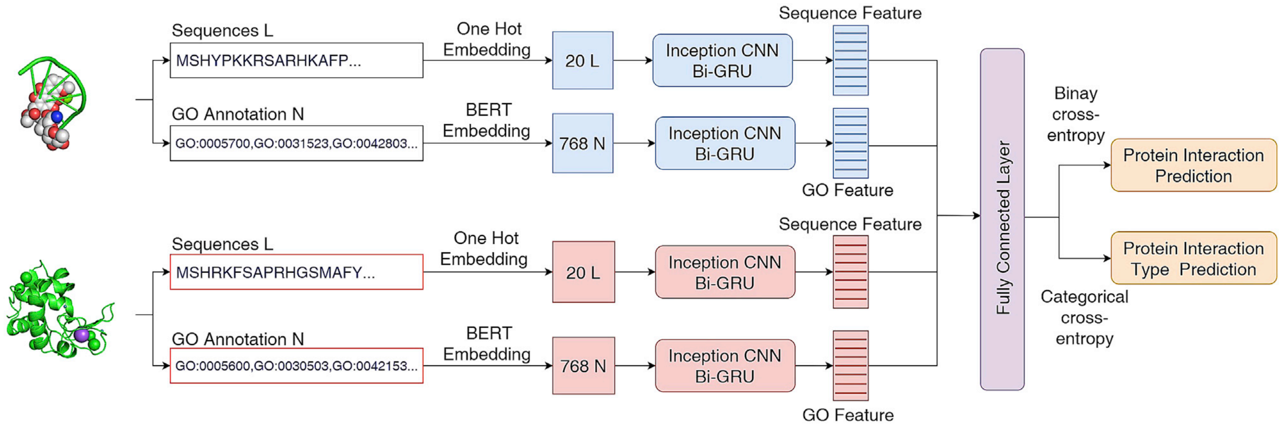

Figure 5. The Framework of the GOSeqPPI

Attention Mechanism

Given an input X, the attention mechanism can guide the network to focus on the important features. Pertinent to the problem at hand, attention mechanism can help us identify relevant regions on protein sequence and GO term for the input. Assume that the output of inception CNN or Bi-GRU layer is X, and two linear transformations of X, Query Q and Key K, which are defined by

$$Q = W_Q^T X, \quad K = W_k^T X$$

where W_Q, W_K are the weight matrices for these two linear transformations, respectively. Then the attention matrix A is computed as follows:

$$A(Q, K) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

where d_k is the dimension of the K and the softmax is a function. For every position in the inception CNN and Bi-GRU output, the

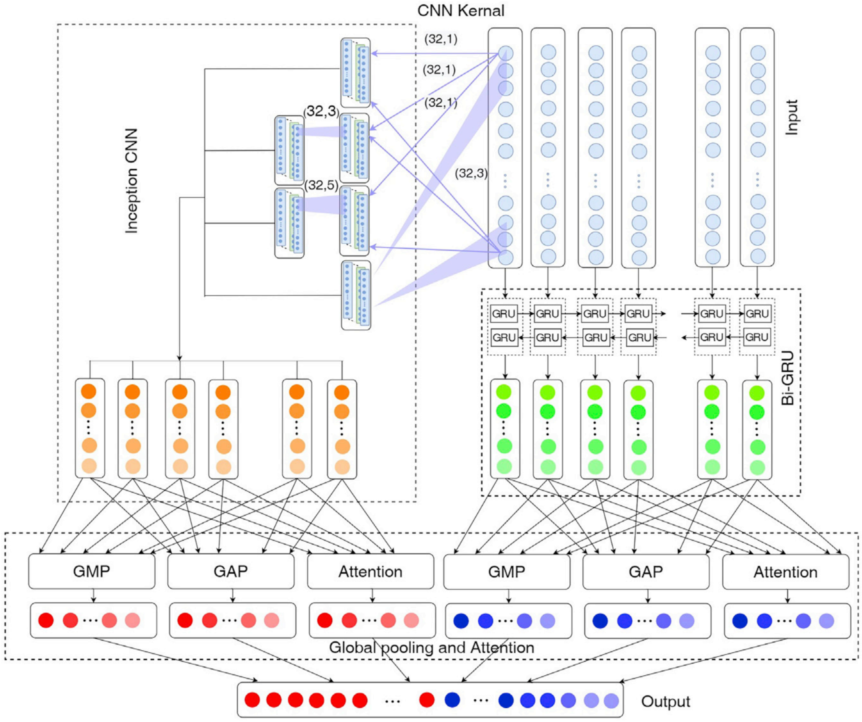

Figure 6. The Architecture of the Proposed Inception RNN Attention Network

Table 7. The Architecture and Parameters in the Proposed IRAN

Input Layer	GO Input (768, N)		Protein Sequence Input (20, L)	
Feature	inception CNN:	Bi-GRU (64)	inception CNN	Bi-GRU (64)
Extraction layer	(1) Conv1D(32,1)		(1) Conv1D(32,1)	
	(2) Conv1D(32,3)		(2) Conv1D(32,3)	
	(3) Conv1D(32,1) stacked by Conv1D(32,3)		(3) Conv1D(32,1) stacked by Conv1D(32,3)	
	(4) Conv1D(32,1) stacked by Conv1D(32,5)		(4) Conv1D(32,1) stacked by Conv1D(32,5)	
Fully connected layer	(1) FC(1024) stacked by dropout(0.1) (2) FC(1024) stacked by dropout(0.1) (3) FC(512) stacked by dropout(0.1) (4) FC(1)			

attention matrix A synthesizes the influence of all other locations on that position. The final output Z of the attention layer based on the attention matrix A is $Z = A \times V$, where V is a value matrix, which defined as $V = W_V^T X$ using the associated weight matrix W_V.

Global Max/Average Pooling

The GMP takes the output of inception CNN or Bi-GRU as input and computes the maximum of all of the values for each of the input channels. The GAP takes the output of inception CNN or Bi-GRU as input and computes the mean of all of the values for each of the input channels.

Classification Module

The PPI or PPI type prediction is performed by four fully connected feed-forward neural networks, which comprise 1,024, 1,024, 512, and 1 neural nodes at each fully connected layer. Given a training set $\theta = \{(x1, y1), ..., (xN, yN)\}$ of N protein pairs with true interaction or type label, the parameters of our neural network were optimized by cross entropy on the training samples, which is defined as the equation$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_i^j \log(o_i^j)$, where C is the number of categories, which depends on the specific task (for binary PPI prediction, C is 2, and for PPI type prediction, C is 7), and o_i^j is the j-th network output when the i-th protein pair is provided as the input of the network. Finally, the weights of our DNN are updated by the Ranger optimizer, which combines the advantages of two new optimizer developments, RAdam and Lookahead.

The parameters in our neural network are listed in Table 7, where the convolutional layer parameters are denoted as “conv1D(no. of channels, kernel size).” Parameters of the Bi-GRU are denoted as “Bi-GRU (no. of units).” The FC is an abbreviation for fully connected layer. “FC(m)” means fully connected layer with m as the hidden units, and dropout(p) means the dropout layer with rate of p.

The Evaluation Method

To validate the effectiveness of the proposed GOSeqPPI, we present three experimental tests on the PPI prediction task from different evaluation perspectives, the evaluation on the influence of joint feature representation; the performance comparison on the datasets with balanced or unbalanced and large or small samples; and the PPI network prediction. To ensure an unbiased evaluation on the GOSeqPPI, we trained and tested GOSeqPPI on the same datasets and validation strategies as the baseline approaches selected for comparison. The following seven typical metrics were selected to evaluate the model prediction performance: Prec, Accu, Sen (also called the true positive rate [TPR]), Spec, F-score, MCC, and AUC. In addition, we applied GOSeqPPI in the analysis of protein interactions and disease to provide a latent application to predict the protein-related diseases according to PPI values.

AUTHOR CONTRIBUTIONS

L.Z. and L.C. substantially contributed to the conception and design of the study. J.W. analyzed and interpreted the data. L.Z., J.W., and Y.H. drafted the article. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

This work was supported by the Tou-Yan Innovation Team Program of the Heilongjiang Province (2019-15), the National Natural Science Foundation of China (61871160), and by the Heilongjiang Province Postdoctoral Fund (LBH-TZ20). We thank all of the reviewers for their valuable comments and suggestions.

REFERENCES

1. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. Nucleic Acids Res. *32* (*Suppl 1*), D449–D451.

2. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., et al. (2004). IntAct: an open source molecular interaction database. Nucleic Acids Res. *32* (*Suppl 1*), D452–D455.

3. Chatr-Aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L., and Cesareni, G. (2007). MINT: the Molecular INTeraction database. Nucleic Acids Res. *35* (*Suppl 1*), D572–D574.

4. Zeng, J., Li, D., Wu, Y., Zou, Q., and Liu, X. (2016). An empirical study of features fusion techniques for protein-protein interaction prediction. Curr. Bioinform. *11*, 4–12.

5. Guo, Y., Yu, L., Wen, Z., and Li, M. (2008). Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res. *36*, 3025–3030.

6. Yang, L., Xia, J.F., and Gui, J. (2010). Prediction of protein-protein interactions from protein sequence using local descriptors. Protein Pept. Lett. *17*, 1085–1090.

7. Davies, M.N., Secker, A., Freitas, A.A., Clark, E., Timmis, J., and Flower, D.R. (2008). Optimizing amino acid groupings for GPCR classification. Bioinformatics *24*, 1980–1986.

8. Chen, C., Zhang, Q., Ma, Q., and Yu, B. (2019). LightGBM-PPI: predicting protein-protein interactions through LightGBM with multi-information fusion. Chemom. Intell. Lab. Syst. *191*, 54–64.

9. Zhou, C., Yu, H., Ding, Y., Guo, F., and Gong, X.-J. (2017). Multi-scale encoding of amino acid sequences for predicting protein interactions using gradient boosting decision tree. PLoS ONE *12*, e0181426.

10. Zhang, L., Yu, G., Xia, D., and Wang, J. (2019). Protein-protein interactions prediction based on ensemble deep neural networks. Neurocomputing *324*, 10–19.

11. Du, X., Sun, S., Hu, C., Yao, Y., Yan, Y., and Zhang, Y. (2017). DeepPPI: boosting prediction of protein-protein interactions with deep neural networks. J. Chem. Inf. Model. *57*, 1499–1510.

12. Li, H., Gong, X.-J., Yu, H., and Zhou, C. (2018). Deep neural network based predictions of protein interactions using primary sequences. Molecules *23*, 1923.

13. Sun, T., Zhou, B., Lai, L., and Pei, J. (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinformatics *18*, 277.

14. Hashemifar, S., Neyshabur, B., Khan, A.A., and Xu, J. (2018). Predicting protein-protein interactions through sequence-based deep learning. Bioinformatics *34*, i802–i810.

15. Wang, L., Wang, H.-F., Liu, S.-R., Yan, X., and Song, K.-J. (2019). Predicting protein-protein interactions from matrix-based protein sequence using convolution neural network and feature-selective rotation forest. Sci. Rep. *9*, 9848.

16. Chen, K.-H., Wang, T.-F., and Hu, Y.-J. (2019). Protein-protein interaction prediction using a hybrid feature representation and a stacked generalization scheme. BMC Bioinformatics *20*, 308.

17. Chen, M., Ju, C.J.-T., Zhou, G., Chen, X., Zhang, T., Chang, K.-W., Zaniolo, C., and Wang, W. (2019). Multifaceted protein-protein interaction prediction based on Siamese residual RCNN. Bioinformatics *35*, i305–i314.

18. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

19. Cheng, L. (2019). Computational and biological methods for gene therapy. Curr. Gene Ther. *19*, 210–10.

20. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science *302*, 449–453.

21. Maetschke, S.R., Simonsen, M., Davis, M.J., and Ragan, M.A. (2012). Gene Ontology-driven inference of protein-protein interactions using inducers. Bioinformatics *28*, 69–75.

22. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. *26*, 3111–3119.

23. Smaili, F.Z., Gao, X., and Hoehndorf, R. (2018). Onto2Vec: joint vector-based representation of biological entities and their ontology-based annotations. Bioinformatics *34*, i52–i60.

24. Smaili, F.Z., Gao, X., and Hoehndorf, R. (2019). OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. Bioinformatics *35*, 2133–2140.

25. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv, arXiv:1802.05365, https://arxiv.org/abs/1802.05365.

26. Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. arXiv, arXiv:1810.04805, https://arxiv.org/abs/1810.04805.

27. Fraser, K.C., Nejadgholi, I., De Bruijn, B., Li, M., LaPlante, A., and El Abidine, K.Z. (2019). Extracting UMLS concepts from medical text using general and domain-specific deep learning models. arXiv, arXiv:1910.01274, https://arxiv.org/abs/1910.01274.

28. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M.B.A. (2019). Publicly available clinical BERT embeddings. arXiv, arXiv:1904.03323, https://arxiv.org/abs/1904.03323.

29. Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., Han, J., Liu, S., and Jiang, Q. (2019). Computational methods for identifying similar diseases. Mol. Ther. Nucleic Acids *18*, 590–604.

30. Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2020). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. Nucleic Acids Res. *48* (D1), D554–D560.

31. Wei, L., Zou, Q., Liao, M., Lu, H., and Zhao, Y. (2016). A novel machine learning method for cytokine-receptor interaction prediction. Comb. Chem. High Throughput Screen. *19*, 144–152.

32. Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., and Jiang, H. (2007). Predicting proteinprotein interactions based only on sequences information. Proc. Natl. Acad. Sci. USA *104*, 4337–4341.

33. DingTang, J., and Guo, F. (2016). Predicting protein-protein interactions via multivariate mutual information of protein sequences. BMC Bioinf *17*, 398.