genes

MDPI

Article

Global Vectors Representation of Protein Sequences and Its Application for Predicting Self-Interacting Proteins with Multi-Grained Cascade Forest Model

Zhan-Heng Chen [1,2], Zhu-Hong You [1,2,*], Wen-Bo Zhang [1,2], Yan-Bin Wang [1], Li Cheng [1,2] and Daniyal Alghazzawi [3]

[1] The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China; chenzhanheng17@mails.ucas.ac.cn (Z.-H.C.); zhang_wen_bo@foxmail.com (W.-B.Z.); wangyanbin15@mails.ucas.ac.cn (Y.-B.W.); chengli@ms.xjb.ac.cn (L.C.)

[2] University of Chinese Academy of Sciences, Beijing 100049, China

[3] Department of Information Systems, King Abdulaziz University, Jeddah 21589, Saudi Arabia; dghazzawi@kau.edu.sa

* Correspondence: zhuhongyou@ms.xjb.ac.cn or zhuhongyou@gmail.com; Tel.: +86-991-3835-823

Received: 31 August 2019; Accepted: 6 November 2019; Published: 12 November 2019

Abstract: Self-interacting proteins (SIPs) is of paramount importance in current molecular biology. There have been developed a number of traditional biological experiment methods for predicting SIPs in the past few years. However, these methods are costly, time-consuming and inefficient, and often limit their usage for predicting SIPs. Therefore, the development of computational method emerges at the times require. In this paper, we for the first time proposed a novel deep learning model which combined natural language processing (NLP) method for potential SIPs prediction from the protein sequence information. More specifically, the protein sequence is de novo assembled by *k-mers*. Then, we obtained the global vectors representation for each protein sequences by using natural language processing (NLP) technique. Finally, based on the knowledge of known self-interacting and non-interacting proteins, a multi-grained cascade forest model is trained to predict SIPs. Comprehensive experiments were performed on *yeast* and *human* datasets, which obtained an accuracy rate of 91.45% and 93.12%, respectively. From our evaluations, the experimental results show that the use of amino acid semantics information is very helpful for addressing the problem of sequences containing both self-interacting and non-interacting pairs of proteins. This work would have potential applications for various biological classification problems.

Keywords: self-interacting proteins; de novo protein sequence; global vector representation; multi-grained cascade forest

1. Introduction

Proteins perform a vast array of functions within organisms. Their self-interaction needs to be considered for the full understanding of cell functions and biological phenomena. However, it is always an important task to identify the interaction between proteins because of the large data it contains in the post-genome era. The prediction of self-interacting proteins (SIPs) will offer a wide understanding to drug target detection [1], drug discovery [2,3], and even further biological processes [4]. According to investigation, the previous biological experimental studies [5,6] have many disadvantages such as high cost, time-consuming, low efficiency and so on. In order to efficiently predict SIPs, many researchers try their best to draw attention to develop new strategies.

From the past years, several researchers have implemented a tremendous work to generate the protein–protein interactions (PPIs) data, which will provide help for discovering the SIPs. Salwinski et al.

established easily accessible online database of interacting proteins, which can be utilized to identify the most reliable subset of the interactions [7]. Chart-Aryamontri et al. updated the biological general repository for interaction datasets that stored the important information of protein, genetic and chemical interactions for humans and organism species [8]. Szklarczyk et al. collected and integrated the functional interactions between expressed proteins, and constructed the STRING database by consolidating known PPIs data [9]. Based on these widely known PPIs datasets, Liu et al. integrated and built *human* and *yeast* datasets for SIPs detection [10].

Currently, a large scale of methods have been exploited to predict PPIs [11–14]. Jansen et al. developed an approach applying Bayesian networks to predict PPIs from genomic data, which can naturally weight and combine into reliable predictions genomic features only weakly associated with interaction [15]. Ofran and Rost predict directly from the sequence of a single protein which residues are interaction hotspots without known of their partner, this research makes it possible to annotate and analyze the hotspots of PPIs in the whole organism, which is conducive to functional prediction and drug development [16]. Zhang et al. combined three-dimensional structural information with other functional clues to predict PPIs on a genome-wide scale, which was comparable in accuracy to high-throughput experiments [12]. Sun et al. studied the sequence-based PPI prediction by applying a stacked autoencoder method, which was the first to use deep-learning algorithm to sequenced-based PPI prediction [17]. Kovács et al. studied PPIs prediction methods on the basis of biological or network-based similarity, and discovered that proteins interact not if they are similar to each other, but if one of them is similar to the other's partners [18]. However, these approaches could be applied to detect PPIs well [19,20], but they are not good enough to predict SIPs. They mainly exist in the following points: (1) In essence, they also have certain limitations that take the correlation between protein pairs into account for SIPs detection, for example co-expression, co-localization, and co-evolution. Nevertheless, this information has no use for SIPs. (2) In addition, the datasets applied to predict PPIs are different from those of SIPs, the datasets of the former are balanced and those of the latter are unbalanced. (3) Besides, there is no PPIs between same partners in the datasets. In virtue of reasons, these computational methods are not suitable for detecting SIPs.

Recently, the researchers found that some similarities between human language and biological language. Nevertheless, it is quite difficult for people to discover the true meaning of biological patterns different from human language. Some researchers have attempted to introduce natural language processing (NLP) technology to the field of bioinformatics. Anon George et al. used NLP technique to extract features and give appropriate representation for the protein sequences, which can better understand the semantics of protein sequences [21]. Wang et al. developed a biological language processing model called bio-to-vector (Bio2Vec) for PPIs detection based on convolution neural network (CNN) [22]. Wan et al. proposed a new scheme for predicting compound–protein interactions by combining feature embedding with deep learning, which used several NLP techniques to extract important features from proteins and compounds [23]. However, seldom do researchers introduce NLP technique to predict SIPs.

In our study, inspired by recent work in NLP technique and deep learning [24–26], we put forward a multi-grained cascade model for SIPs prediction base on global vectors representation of de novo protein sequence. Furthermore, the major advantages of our method include the following three aspects: (1) *k-mers* method was exploited to de novo assemble protein sequence; (2) we employed global vectors (GloVe) representation learning method to generate feature vector of each *mer* from de novo protein sequence, a 100-Dimensional feature vector from the numerical series was achieved by this method; and (3) multi-grained cascade model was applied to optimize the characteristics and predict SIPs. In detail, we first used *3-mers* to de novo assemble every protein sequence from the corresponding datasets, and each protein sequence was regarded as a "sentence", every *mer* in a "sentence" was treated as a "word". Then, GloVe model was applied for emerging 100-dimensional feature vectors of each *mer*, which can contain, as much as possible, the semantic information between *mers*. Finally, the feature vectors of all protein sequences were fed into a multi-grained cascade forest classifier to predict SIPs. We have tested our model on *yeast* and *human* datasets. The experimental results demonstrated that the superior performance of our model than the other previous methods in predicting new SIPs. It is revealed that the presented method is suitable and perform well for detecting SIPs.

2. Materials and Methods

2.1. Benchmark Datasets Preparation

As we all know that the PPIs related information can be achieved from many different types of resources, including DIP [7], InnateDB [27], IntAct [28], BioGRID [8] and MatrixDB [29]. In this study, the datasets constructed in the experiment which contains 20,199 curated *human* protein sequences mainly derived from the UniProt database [30]. We mainly set up the SIPs datasets for the experiment which embodies two identical interacting protein sequences and whose type of interaction was characterized as "direct interaction" in relational databases. On this foundation, we can obtain 2994 *human* self-interacting protein sequences which would be employed to construct the datasets for the experiment.

We need to select the datasets from the 2994 *human* SIPs for the experiment to measure the performance of our prediction model, which mainly includes three steps [10]: (1) We removed the protein sequences which may be fragments, and retained the length of protein sequences more than 50 residues and less than 5000 residues from all the *human* proteome; (2) to build the *human* positive dataset, we chose a high quality SIPs data which should conform to one of the following conditions: (a) The self-interactions were revealed by at least one small-scale experiment or two sorts of large-scale experiments; (b) the protein has been announced as homo-oligomer (containing homodimer and homotrimer) in UniProt; (c) it has been reported by more than two publications for the self-interactions; and (3) for the *human* negative dataset, we removed all the types of SIPs from the whole *human* proteome (including proteins annotated as 'direct interaction' and more extensive 'physical association') and SIPs detection in UniProt database. Eventually, the ultimate *human* dataset for the experiment was consisted of 1441 SIPs and 15,938 non-SIPs [10].

In the same way, we also generated the *yeast* dataset to further measure the cross-species capacity of our multi-grained cascade forest model by repeating the same strategy mentioned above. Finally, the *yeast* benchmark dataset was consisted of 710 positive sample and 5511 negative sample [10].

2.2. De Novo Assembly Protein Sequences

Actually, there is a close relation between biological language and human natural language. De novo assembly protein sequence is a helpful tool for understanding the relationship between amino acids in biological language. The protein sequences could be regarded as sentences, and the monomeric units (*mer*) were treated as words. The *k-mer* is a de novo protein sequence assembly method in the present study, which could be employed to; divide a sequence into many sets of *k* amino acid residues. In the Figure 1, the protein sequences were de novo assembled by *3-mers* composition [31]. A series of new amino acid compositions were created for each protein sequence [32]. For example, there is one protein sequence with length *m*, and every *3-mers* composition amino acids were regarded as a "word". Then, the protein sequence will be de novo assembled by m-2 *3-mers*.

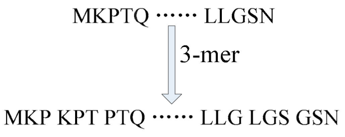

Figure 1. De novo assembled protein sequences by *3-mer*.

In conclusion, *k-mer* has become essential to de novo assembly protein sequences for predicting self-interacting proteins. Each protein sequence was de novo assembled by *k-mer* algorithm. Then, these entire *mers* were saved into a text file, which can be employed for building co-occurrence matrix and obtaining the feature vectors.

2.3. Global Vectors Representation of Protein Sequences

In natural language processing (NLP), global vectors representation (i.e., GloVe) is an ideal tool for obtaining vectors corresponding to each word in a corpus [33]. GloVe used global and local statistical information of words to generate language model and word vectors. At present, there are two main types of vector representations: (1) Global matrix factorization methods, such as latent semantic analysis (LSA) [34]; (2) local context window methods, such as skip-gram model [35]; GloVe combines the advantages of these two methods, it can not only accelerate the training speed of the model, but also control the relative weight of words. This model consists of two steps: Construction of co-occurrence matrix and generation of global vectors.

The statistics of *mer–mer* co-occurrences in a protein sequence is the key source of information. First of all, let the co-occurrence matrix be X, where $X_{i,j}$ represents the frequency of i and j appear together in a window (the size of window is 7). There was one sentence with several words as follow:

MKP KPT PTQ TQD QDS DSQ SQE QEK EKV ...... LLG LGS GSN

Supposing that the central *mer* is TQD, while the contextual *mers* are MKP, KPT, PTQ, QDS, DSQ, and SQE. The number of *mer–mer* co-occurrences is calculated as follows:

$$X_{TQD,MKP} += 1 \tag{1}$$

$$X_{TQD,KPT} += 1 \tag{2}$$

$$X_{TQD,PTQ} += 1 \tag{3}$$

$$X_{TQD,QDS} += 1 \tag{4}$$

$$X_{TQD,DSQ} += 1 \tag{5}$$

$$X_{TQD,SQE} += 1 \tag{6}$$

The co-occurrence matrix X will be obtained by traversing the whole sequences through a sliding window.

Next, we constructed the approximate relationship between word vector and the co-occurrence matrix as follow:

$$v_i^T v_j + b_i + b_j = log(X_{ij}) \tag{7}$$

where v_i, v_j are the vectors of *mer* i and j respectively; b_i, b_j are two additional scalars (biases). Then, the cost model will be built as Formula (8) [33].

$$J = \sum_{i,j}^{N} f(X_{ij})(v_i^T v_j + b_i + b_j - \log(X_{ij}))^2 \tag{8}$$

where, N is the size of dataset, and the size of co-occurrence matrix is $N*N$; $f(X_{ij})$ is the weighting function which should obey the following condition:

$$f(x) = \begin{cases} (x/x_{max})^{\alpha} & if\ x < x_{max} \\ 1 & otherwise \end{cases} \tag{9}$$

In our research, while training the model, most hyper-parameters were set by default, but a few parameters still need to be set. The gradient descent algorithm based on AdaGrad was used to randomly sample all non-zero elements of matrix X. We also set *learn_rate* = 0.05, *vector_size* = 100, *iterations* = 25, *window_size* = 7 and *min_count* = 0. Hence, each protein sequence of the corresponding dataset can be obtained a 100-dimensional vector by applying GloVe model.

2.4. Multi-Grained Cascade Forest

In 2017, Zhou et al. proposed a novel decision tree ensemble method called multi-grained cascade forest [36,37], with less hyper-parameters than deep neural networks (DNN) [38,39]. There are two stages in the training process of multi-grained cascade forest model: Multi-grained scanning and cascade forest. The multi-grained scanning was used to generate feature values, cascade forest was applied to obtain the prediction results through multiple forest cascades.

2.4.1. Multi-Grained Scanning

In our model, in order to improve feature representation, we used sliding window to scan the raw feature vectors and input them into forest to generate new features. Thereby, the cascade forest could be enhanced by the multi-grained scanning method. The scanning process is shown Figure 2.

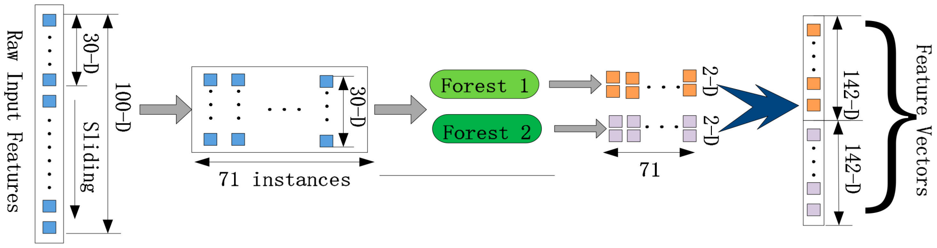

Figure 2. Process of multi-grained scanning.

There are 2 classes in our experiment, and the raw feature vectors were obtained by GloVe method, whose size was 100. We process these feature vectors by using a 30-dimensional sliding window. And then, in total 71 instances were produced. These instances extracted from the same windows were employed to train a random forest and a completely-random tree forest. Subsequently, each forest will achieve 142-dimensional derived vectors. As shown in Figure 2, the final class vectors were obtained and concatenated by these derived vectors, which will be input into the cascade forest.

2.4.2. Cascade Forest

On the basis of representation learning in DNN, we employed cascade forest to predict the SIPs based on layer-by-layer processing of input feature vectors. The main thought is to conjecture better results through multilayer cascade forest. This method also called ensemble of ensembles. Every layer of cascade forest used a set of forests to simulate the representation, and each forest was consisted of many decision trees, and each layer of forest applied the information of the upper layer as its own input, while its output provides input information for the next layer. The cascade forest model is shown in Figure 3.

As is shown in Figure 3, every layer of cascade forest was composed of 2 random forests and 2 completely-random tree forests. The number of trees in these forests were set with default parameters. For the completely-random tree forests, each node of every decision tree randomly chose one feature to split until all the instances of each leaf node belongs to the same class or the number of instances is less than 10. For the random forests, each decision tree was generated by randomly selecting *sqrt(d)* features (*d* is the total input features), and then chose the maximum feature of *gini* coefficient as the condition of node partition. In the experiment, after layer-by-layer processing of input feature vectors, each forest will generate a 2-dimensional predictive probability distribution vector. Then, to average the four 2-dimensional vectors at the last level. Finally, we can achieve the final prediction results with the maximum aggregated value.

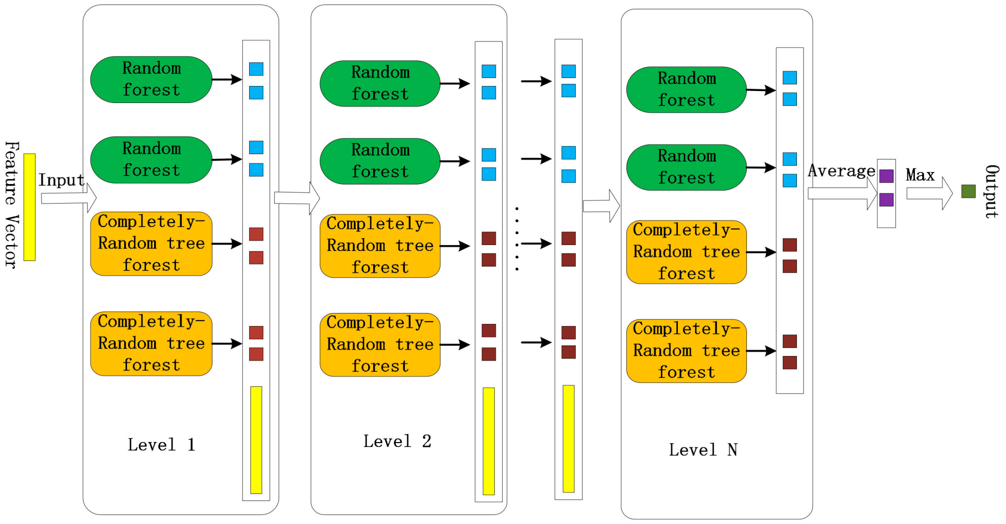

Figure 3. Cascade forest model.

2.5. Performance Evaluation Indicators

As to classification model, there are 3 main evaluation indicators: Confusion matrix (CM), receiver operating characteristic (ROC) curve and area under ROC curve (AUC). CM is an important performance assessment tool for our proposed classification model. We can establish a table contained four underlying indexes as follow:

From the Table 1, each row of the matrix is the situation of true samples and each column of matrix represents the situation of predicted samples. Meaning of the four underlying indexed (also called primary indices) are as follow:

(1) TN: True negative, the number of true non-interacting pairs correctly predicted;
(2) FN: False negative, the quantity of true non-interacting pairs falsely predicted;
(3) FP: False positive, the count of true interacting pairs falsely predicted;
(4) TP: True positive, the quantity of true interacting pairs correctly predicted.

Table 1. Confusion matrix. TN: true negative, FN: false negative, FP: false positive, TP: true positive.

		Predict	
		Negative	**Positive**
Actual	Negative	TN	FN
	Positive	FP	TP

However, in the face of a large number of data, CM is hard to be used to measure the quality of the model only by the number of statistic samples. On the basis of those parameters of CM, we calculated the four values which called secondary indicator as follow:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

$$TNR = \frac{TN}{FP + TN} \tag{11}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{12}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{13}$$

where,

(1) Acc: Accuracy, the proportion of all judged correctly samples in the total observation values from the classification model;
(2) TNR: True negative rate or specificity, the proportion of predicted correctly samples in the result with actual negative value;
(3) F1-score: Measuring the overall performance of the classification model;
(4) MCC: Matthews correlation coefficient, the geometric mean of the problem and dual regression coefficients; It is a better indicator for measuring unbalanced dataset and the most informative single fraction for assessing the quality of binary classifier from the CM.

Moreover, a receiver operating curve (ROC) was plotted to evaluate the performance of random projection method. And then, we can compute the area under curve (AUC) to evaluate the quality of the classifier.

3. Results and Discussion

3.1. Performance Evaluation on Protein Self-Interaction

We first assessed the proposed method on the SIPs extracted from *yeast* dataset. In order to avoid the risk of over-fitting, k-fold cross validation was applied to the class vectors generated by each forest in our model. More concretely, we only separated the datasets which were mainly composed of characteristic values into k non-overlapping pieces, and each training sample was used $k-1$ times in forest to generate $k-1$ class categories list, and then averaged them to generate the final result as the enhancement feature of the next level in the cascade forest. Because cross validation has been done in the process of multi-grained cascade forest model building, it is not necessary to introduce it again when using our proposed model to realize classification prediction. To illustrate the rationality, toughness and stability of our algorithm, we also implemented the method of multi-grained cascade forest on the *human* dataset.

To guarantee impartiality and objectivity of the test, the parameters for *human* and *yeast* datasets should be set in the same way. In our task, compared with DNN, multi-grained cascade forest model has fewer hyper-parameters and much easier to train. When processing various of data in different fields, it can achieve excellent performance under almost identical hyper-parameter settings, in other words, it has high robustness for our model to set hyper-parameters. Because the model is insensitive to the process parameter change, and a set of hyper-parameters can be applied to different datasets. Hence, apart from a few parameters, most of them were set by default. In the experiment, we set *shape_1X* = 30 (shape of a single sample unit. Required when using multi-grained scanning), *window* = 30 (the size of sliding window during multi-grained scanning), *tolerance* = 5.0 (accuracy tolerance for the cascade growth). If the improvement in accuracy is not better than the tolerance, the construction is stopped.

Afterwards, we test our prediction model on *yeast* and *human* benchmark datasets, and the results are shown in Table 2. From the data, it is revealed that our proposed model exhibited the outcomes of Acc, TNR (Specificity), F1-score and MCC of 91.45%, 99.71%, 37.56%, and 0.4389 respectively on *yeast* dataset. Similarly, we can obtain the results by running experiment on *human* dataset, the Acc is 93.12%, TNP is 99.57%, F1-score is 39.10%, and MCC is 0.4421.

Meanwhile, receiver operating characteristic (ROC) curves was widely applied in many fields, such as machine learning, data mining, and so on. We also used ROC curves to measure the comprehensive index between false positive rate and true positive rate continuous variable. The area under curves (AUC) could be shown the prediction accuracy of the classifier. The larger the AUC, the higher the accuracy.

The ROC curve of our proposed model on *yeast* dataset is shown in Figure 4, it is obvious that the AUC is 0.7881. The ROC curve of the proposed model on *human* dataset is shown in Figure 5, it is clear that the AUC is 0.8524. That is to say, our model performs better on large scale dataset than the small dataset. Overall, the presented model is an accurate and robust classifier for predicting SIPs.

Table 2. Performance of our proposed model on the two benchmark datasets. Acc: Accuracy; TNR: True negative rate; F1-score: Measuring the overall performance of the classification model; MCC: Matthews correlation.

Datasets	Acc (%)	TNR (%)	F1-Score (%)	MCC
yeast	91.45	99.71	37.56	0.4389
human	93.12	99.57	39.10	0.4421

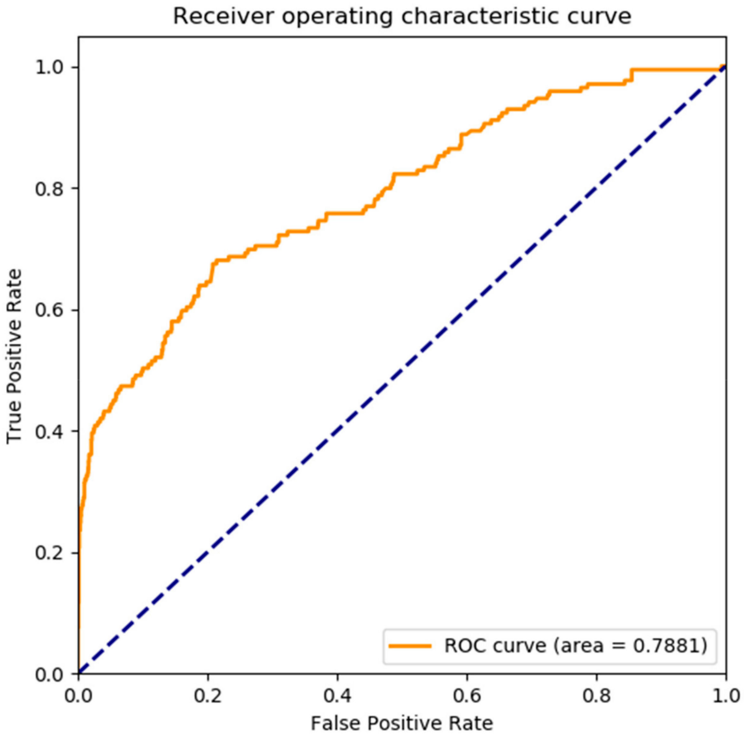

Figure 4. The receiver operating characteristic (ROC) curve of proposed model on *yeast* dataset.

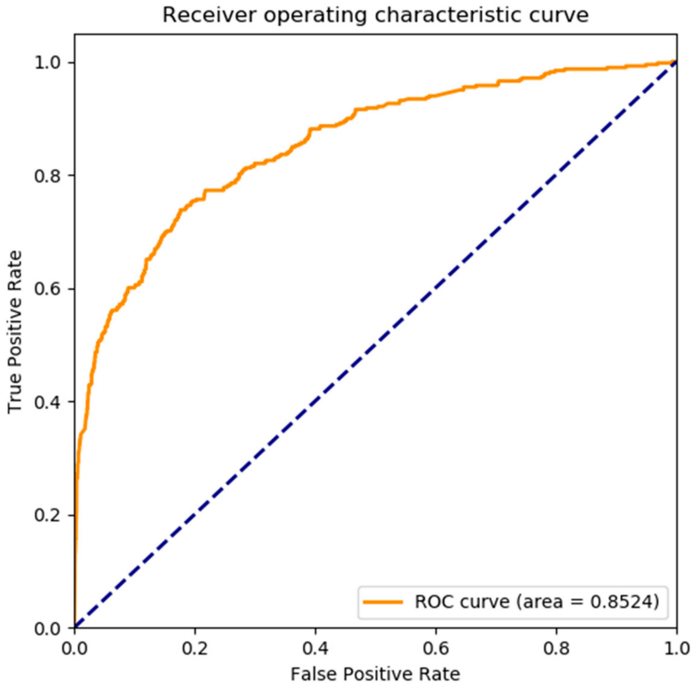

Figure 5. The ROC curve of proposed model on *human* dataset.

3.2. Comparison with Other Existing Methods for Predicting SIPs

To further measure the quality of our proposed model, we compared the proposed model with other previous methods based on the two benchmark datasets. The comparison results were listed a clear statement of account in Tables 3 and 4. From Table 3, it is obvious that the multi-grained cascade forest model obtained the highest accuracy of 91.45% than the other six methods (range from 66.28% to 87.46%) on *yeast* dataset. At the same instant, it is clear to see that the other six methods got lower MCC (range from 0.1577 to 0.2842) than our proposed model of 0.4389 on the same dataset. In exactly the same way, from Table 4, the overall results of our prediction approach is also significantly better than the other six methods on *human* dataset. To make a summary, we assessed our multi-grained cascade forest model against with the other six approaches on both *yeast* and *human* datasets, so that the prediction accuracy of the overall experimental results could be improved. This fully illustrates that a reasonable feature representation method and a suitable classifier are very significant for predicting SIPs. It is further illustrated that the proposed method is superior to the other six approaches and quite suitable for predicting SIPs.

Table 3. Performance of our proposed model and other previous methods on *yeast* dataset. AUC: Area under curve.

Model	Acc (%)	TNR (%)	F1-Score (%)	MCC	AUC
SLIPPER [40]	71.90	72.18	36.16	0.2842	0.7723
DXECPPI [41]	87.46	94.93	34.89	0.2825	0.6934
PPIevo [42]	66.28	87.46	28.92	0.1801	0.6728
LocFuse [43]	66.66	68.10	27.53	0.1577	0.7087
CRS [10]	72.69	74.37	33.05	0.2368	0.7115
SPAR [10]	76.96	80.02	34.54	0.2484	0.7455
Proposed method	**91.45**	**99.71**	**37.56**	**0.4389**	**0.7881**

Table 4. Performance of our proposed model and other previous methods on *human* dataset.

Model	Acc (%)	TNR (%)	F1-score (%)	MCC	AUC
SLIPPER [40]	91.10	95.06	**46.82**	0.4197	**0.8723**
DXECPPI [41]	30.90	25.83	17.28	0.0825	0.5806
PPIevo [42]	78.04	25.82	27.73	0.2082	0.7329
LocFuse [43]	80.66	80.50	27.65	0.2026	0.7087
CRS [10]	91.54	96.72	36.83	0.3633	0.8196
SPAR [10]	92.09	97.40	41.13	0.3836	0.8229
Proposed method	**93.12**	**99.57**	39.10	**0.4421**	0.8524

As mentioned above, it is apparent that our method can receive good effect of SIPs detection because of the appropriate feature representation and classifier. The presented feature representation technique plays a critical part in enhancing the prediction accuracy. The specific reasons can be summed up in the following three aspects: (1) *k-mer* method was exploited to de novo assemble protein sequence. Not only can it represents the information of protein sequence, but also it preserves useful enough information as much as possible; (2) we employed global vectors (GloVe) representation learning method to generate feature vector of each *mer* from de novo protein sequence, a 100-Dimensional feature vector from the numerical series was achieved by this method. Hence, the protein sequences can be described in the form of numerical values; and (3) multi-grained cascade forest model was applied to optimize the characteristics and predict SIPs. In a few words, experimental results revealed that our presented model is extreme fit for SIPs prediction.

4. Conclusions

In this study, a multi-grained cascade forest-based model was developed for predicting SIPs based on protein primary sequence. To better understand the aggregation relationship among amino acids and discover the semantic information of proteins, we proposed an improved global vectors representation learning scheme from the de novo assembled protein sequence based on natural language processing technique. We implement our model on *yeast* and *human* SIPs datasets, each protein sequence can be de novo assembled by *3-mers* technique and obtained a 100-dimensional feature vector. Afterwards, we evaluated the performance of our proposed model on the two benchmark datasets and also compared with other popular methods, which achieved an accuracy rate of 91.45% and 93.12% respectively. Experimental results revealed that our method has better performance than other existing approaches. We conjecture that de novo assembly protein sequence combined with GloVe representation may play an important role in the SIPs prediction and help to increase efforts in discovering amino acid words. For the future work, there will be more effective NLP methods and deep learning techniques introduced for detecting SIPs.

5. Patents

This work has been applied for the national invention patent of China.

Author Contributions: Z.-H.C. and Z.-H.Y. conceived the algorithm, carried out analyses, prepared the data sets, carried out experiments, and wrote the manuscript; W.-B.Z. and Y.-B.W. designed, performed and analyzed experiments; L.C. and D.A. reviewed the manuscript; All authors have read and approved final manuscript.

Funding: This research was funded by NATIONAL NATURAL SCIENCE FOUNDATION OF CHINA (grant number 61373086 and 61572506), the West Light Foundation of The Chinese Academy of Sciences (grant number 2017-XBZG-BR-001).

Acknowledgments: The authors would like to thank all the guest editors and anonymous reviewers for their constructive advices.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yıldırım, M.A.; Goh, K.-I.; Cusick, M.E.; Barabási, A.-L.; Vidal, M. Drug—Target network. *Nat. Biotechnol.* **2007**, *25*, 1119. [CrossRef] [PubMed]
2. Radivojac, P.; Clark, W.T.; Oron, T.R.; Schnoes, A.M.; Wittkop, T.; Sokolov, A.; Graim, K.; Funk, C.; Verspoor, K.; Ben-Hur, A.; et al. A large-scale evaluation of computational protein function prediction. *Nat. Methods* **2013**, *10*, 221. [CrossRef] [PubMed]
3. Cao, R.; Cheng, J. Integrated protein function prediction by mining function associations, sequences, and protein–protein and gene–gene interaction networks. *Methods* **2016**, *93*, 84–91. [CrossRef] [PubMed]
4. Ispolatov, I.; Yuryev, A.; Mazo, I.; Maslov, S. Binding properties and evolution of homodimers in protein–protein interaction networks. *Nucleic Acids Res.* **2005**, *33*, 3629–3635. [CrossRef] [PubMed]
5. Shoemaker, B.; Panchenko, A. Deciphering protein-protein interactions. *PLoS Comput. Biol.* **2006**, *3*, e43.
6. Reguly, T.; Breitkreutz, A.; Boucher, L.; Breitkreutz, B.-J.; Hon, G.C.; Myers, C.L.; Parsons, A.; Friesen, H.; Oughtred, R.; Tong, A.; et al. Comprehensive curation and analysis of global interaction networks in saccharomyces cerevisiae. *J. Biol.* **2006**, *5*, 11. [CrossRef] [PubMed]
7. Salwinski, L.; Miller, C.S.; Smith, A.J.; Pettit, F.K.; Bowie, J.U.; Eisenberg, D. The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **2004**, *32*, D449–D451. [CrossRef] [PubMed]
8. Chatr-Aryamontri, A.; Oughtred, R.; Boucher, L.; Rust, J.; Chang, C.; Kolas, N.K.; O'Donnell, L.; Oster, S.; Theesfeld, C.; Sellam, A.; et al. The biogrid interaction database: 2017 update. *Nucleic Acids Res.* **2017**, *45*, D369–D379. [CrossRef] [PubMed]
9. Szklarczyk, D.; Morris, J.H.; Cook, H.; Kuhn, M.; Wyder, S.; Simonovic, M.; Santos, A.; Doncheva, N.T.; Roth, A.; Bork, P.; et al. The string database in 2017: Quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **2017**, *45*, D362–D368. [CrossRef] [PubMed]

10. Liu, X.; Yang, S.; Li, C.; Zhang, Z.; Song, J. Spar: A random forest-based predictor for self-interacting proteins with fine-grained domain information. *Amino Acids* **2016**, *48*, 1655–1665. [CrossRef] [PubMed]
11. Zhu, L.; Deng, S.-P.; You, Z.-H.; Huang, D.-S. Identifying spurious interactions in the protein-protein interaction networks using local similarity preserving embedding. *IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB)* **2017**, *14*, 345–352. [CrossRef] [PubMed]
12. Zhang, Q.C.; Petrey, D.; Deng, L.; Qiang, L.; Shi, Y.; Thu, C.A.; Bisikirska, B.; Lefebvre, C.; Accili, D.; Hunter, T.; et al. Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature* **2012**, *490*, 556. [CrossRef] [PubMed]
13. You, Z.-H.; Zhou, M.; Luo, X.; Li, S. Highly efficient framework for predicting interactions between proteins. *IEEE Trans. Cybern.* **2017**, *47*, 731–743. [CrossRef] [PubMed]
14. You, Z.-H.; Lei, Y.-K.; Gui, J.; Huang, D.-S.; Zhou, X. Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioinformatics* **2010**, *26*, 2744–2751. [CrossRef] [PubMed]
15. Jansen, R.; Yu, H.; Greenbaum, D.; Kluger, Y.; Krogan, N.J.; Chung, S.; Emili, A.; Snyder, M.; Greenblatt, J.F.; Gerstein, M. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **2003**, *302*, 449–453. [CrossRef] [PubMed]
16. Ofran, Y.; Rost, B. Protein–protein interaction hotspots carved into sequences. *PLoS Comput. Biol.* **2007**, *3*, e119. [CrossRef] [PubMed]
17. Sun, T.; Zhou, B.; Lai, L.; Pei, J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinform.* **2017**, *18*, 277. [CrossRef] [PubMed]
18. Kovács, I.A.; Luck, K.; Spirohn, K.; Wang, Y.; Pollis, C.; Schlabach, S.; Bian, W.; Kim, D.-K.; Kishore, N.; Hao, T.; et al. Network-based prediction of protein interactions. *Nat. Commun.* **2019**, *10*, 1240. [CrossRef] [PubMed]
19. Wang, Y.-B.; You, Z.-H.; Li, X.; Jiang, T.-H.; Cheng, L.; Chen, Z.-H. Prediction of protein self-interactions using stacked long short-term memory from protein sequences information. *BMC Syst. Biol.* **2018**, *12*, 129. [CrossRef] [PubMed]
20. Chen, Z.-H.; You, Z.-H.; Li, L.-P.; Wang, Y.-B.; Wong, L.; Yi, H.-C. Prediction of self-interacting proteins from protein sequence information based on random projection model and fast fourier transform. *Int. J. Mol. Sci.* **2019**, *20*, 930. [CrossRef] [PubMed]
21. George, A.; Ganesh, H.B.; Kumar, M.A.; Soman, K. Significance of Global Vectors Representation in Protein Sequences Analysis. In *Computer Aided Intervention and Diagnostics in Clinical and Medical Images*; Springer: Cham, Switzerland, 2019; pp. 261–269.
22. Wang, Y.; You, Z.-H.; Yang, S.; Li, X.; Jiang, T.-H.; Zhou, X. A high efficient biological language model for predicting protein–protein interactions. *Cells* **2019**, *8*, 122. [CrossRef] [PubMed]
23. Wan, F.; Zeng, J. Deep learning with feature embedding for compound-protein interaction prediction. *bioRxiv* **2016**, 086033.
24. Luo, X.; Zhou, M.; Xia, Y.; Zhu, Q. An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. *IEEE Trans. Ind. Inform.* **2014**, *10*, 1273–1284.
25. Jin, L.; Li, S.; La, H.M.; Luo, X. Manipulability optimization of redundant manipulators using dynamic neural networks. *IEEE Trans. Ind. Electron.* **2017**, *64*, 4710–4720. [CrossRef]
26. Liu, W.; Wang, Z.; Liu, X.; Zeng, N.; Liu, Y.; Alsaadi, F.E. A survey of deep neural network architectures and their applications. *Neurocomputing* **2017**, *234*, 11–26. [CrossRef]
27. Breuer, K.; Foroushani, A.K.; Laird, M.R.; Chen, C.; Sribnaia, A.; Lo, R.; Winsor, G.L.; Hancock, R.E.; Brinkman, F.S.; Lynn, D.J. Innatedb: Systems biology of innate immunity and beyond—Recent updates and continuing curation. *Nucleic Acids Res.* **2012**, *41*, D1228–D1233. [CrossRef] [PubMed]
28. Orchard, S.; Ammari, M.; Aranda, B.; Breuza, L.; Briganti, L.; Broackes-Carter, F.; Campbell, N.H.; Chavali, G.; Chen, C.; Del-Toro, N.; et al. The mintact project—Intact as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **2013**, *42*, D358–D363. [CrossRef] [PubMed]
29. Clerc, O.; Deniaud, M.; Vallet, S.D.; Naba, A.; Rivet, A.; Perez, S.; Thierry-Mieg, N.; Ricard-Blum, S. Matrixdb: Integration of new data with a focus on glycosaminoglycan interactions. *Nucleic Acids Res.* **2018**, *47*, D376–D381. [CrossRef] [PubMed]
30. Uniprot: The universal protein knowledgebase. *Nucleic Acids Res.* **2016**, *45*, D158–D169.

31. Muppirala, U.K.; Honavar, V.G.; Dobbs, D. Predicting rna-protein interactions using only sequence information. *BMC Bioinform.* **2011**, *12*, 489. [CrossRef] [PubMed]
32. Asgari, E.; Mofrad, M.R. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS ONE* **2015**, *10*, e0141287. [CrossRef] [PubMed]
33. Pennington, J.; Socher, R.; Manning, C. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
34. Merchant, K.; Pande, Y. Nlp based latent semantic analysis for legal text summarization. In Proceedings of the 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, India, 19–22 September 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1803–1807.
35. Liu, P.; Qiu, X.; Huang, X. Learning context-sensitive word embeddings with neural tensor skip-gram model. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015.
36. Zhou, Z.-H.; Feng, J. Deep forest: Towards an alternative to deep neural networks. *arXiv* **2017**, arXiv:1702.08835.
37. Chen, Z.-H.; Li, L.-P.; He, Z.; Zhou, J.-R.; Li, Y.; Wong, L. An improved deep forest model for predicting self-interacting proteins from protein sequence using wavelet transformation. *Front. Genet.* **2019**, *10*, 90. [CrossRef] [PubMed]
38. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.; Mohamed, A.-r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Kingsbury, B.; et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
39. Liu, X.; Liu, W.; Ma, H.; Fu, H. Large-scale vehicle re-identification in urban surveillance videos. In Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, USA, 11–15 July 2016; pp. 1–6.
40. Liu, Z.; Guo, F.; Zhang, J.; Wang, J.; Lu, L.; Li, D.; He, F. Proteome-wide prediction of self-interacting proteins based on multiple properties. *Mol. Cell. Proteom.* **2013**, *12*, 1689–1700. [CrossRef] [PubMed]
41. Du, X.; Cheng, J.; Zheng, T.; Duan, Z.; Qian, F. A novel feature extraction scheme with ensemble coding for protein–protein interaction prediction. *Int. J. Mol. Sci.* **2014**, *15*, 12731–12749. [CrossRef] [PubMed]
42. Zahiri, J.; Yaghoubi, O.; Mohammad-Noori, M.; Ebrahimpour, R.; Masoudi-Nejad, A. Ppievo: Protein–protein interaction prediction from pssm based evolutionary information. *Genomics* **2013**, *102*, 237–242. [CrossRef] [PubMed]
43. Zahiri, J.; Mohammad-Noori, M.; Ebrahimpour, R.; Saadat, S.; Bozorgmehr, J.H.; Goldberg, T.; Masoudi-Nejad, A. Locfuse: Human protein–protein interaction prediction via classifier fusion using protein localization information. *Genomics* **2014**, *104*, 496–503. [CrossRef] [PubMed]