2021

Towards A More Effective Bidirectional LSTM-based Learning Model for Human-Bacterium Protein-Protein Interactions

Huaming Chen
hc007@uowmail.edu.au

Jun Shen
University of Wollongong, jshen@uow.edu.au

Lei Wang
University of Wollongong, leiw@uow.edu.au

Yaochu Jin
yaochu.jin@surrey.ac.uk

Towards A More Effective Bidirectional LSTM-based Learning Model for Human-Bacterium Protein-Protein Interactions

Abstract

The identification of protein-protein interaction (PPI) is one of the most important tasks to understand the biological functions and disease mechanisms. Although numerous databases of biological interactions have been published in debt to advanced high-throughput technology, the study of inter-species protein-protein interactions, especially between human and bacterium pathogens, remains an active yet challenging topic to harness computational models tackling the complex analysis and prediction tasks. In this paper, we comprehensively revisit the prediction task of human-bacterium protein-protein interactions (HB-PPI), which is a first ever endeavour to report an empirical evaluation in learning and predicting HB-PPI based on machine learning models. Firstly, we summarise the literature review of human-bacterium interaction (HBI) study, meanwhile a vast number of databases published in the last decades are carefully examined. Secondly, a broader and deeper experimental framework is designed for HB-PPI prediction task, which explores a variety of feature representation algorithms and different computational models to learn from the curated HB-PPI dataset and perform predictions. Furthermore, a bidirectional LSTM-based model is proposed for the prediction task, which demonstrates a more effective performance in comparison with the others. Finally, opportunities for improving the performance and robustness of machine learning models for HP-PPI prediction are also discussed, laying a foundation for future work.

Keywords

lstm-based, model, interactions, bidirectional, effective, learning, protein-protein, towards, more, human-bacterium

Disciplines

Engineering | Science and Technology Studies

Publication Details

Chen, H., Shen, J., Wang, L. & Jin, Y. (2021). Towards A More Effective Bidirectional LSTM-based Learning Model for Human-Bacterium Protein-Protein Interactions. 14th International Conference on Practical Applications of Computational Biology & Bioinformatics (pp. 91-101). Switzerland: Springer.

Towards A More Effective Bidirectional LSTM-based Learning Model for Human-Bacterium Protein-Protein Interactions

Huaming Chen[1], Jun Shen[1] Lei Wang[1], and Yaochu Jin[2]

[1] University of Wollongong, Wollongong, NSW 2500, Australia
hc007@uowmail.edu.au,{jshen, leiw}@uow.edu.au
[2] University of Surrey, Guildford GU2 7XH, UK
yaochu.jin@surrey.ac.uk

Abstract. The identification of protein-protein interaction (PPI) is one of the most important tasks to understand the biological functions and disease mechanisms. Although numerous databases of biological interactions have been published in debt to advanced high-throughput technology, the study of inter-species protein-protein interactions, especially between human and bacterium pathogens, remains an active yet challenging topic to harness computational models tackling the complex analysis and prediction tasks. In this paper, we comprehensively revisit the prediction task of human-bacterium protein-protein interactions (HB-PPI), which is a first ever endeavour to report an empirical evaluation in learning and predicting HB-PPI based on machine learning models. Firstly, we summarise the literature review of human-bacterium interaction (HBI) study, meanwhile a vast number of databases published in the last decades are carefully examined. Secondly, a broader and deeper experimental framework is designed for HB-PPI prediction task, which explores a variety of feature representation algorithms and different computational models to learn from the curated HB-PPI dataset and perform predictions. Furthermore, a bidirectional LSTM-based model is proposed for the prediction task, which demonstrates a more effective performance in comparison with the others. Finally, opportunities for improving the performance and robustness of machine learning models for HP-PPI prediction are also discussed, laying a foundation for future work.

Keywords: Human-bacterium interactions, protein-protein interactions, machine learning, computational model

1 Introduction

Monitoring and curing the infectious diseases for human are still prevalent and intractable problems, while there have been substantial researches focusing on the understanding of infectious mechanisms and the development of novel therapeutic solutions. This solicits great efforts in revealing the biological interactions between human and different pathogens [1, 12, 22]. However, research on identification of interactions is still in its early stage. Some published data may focus on particular human-pathogen interactions (HPI) system, for example between human and HIV virus, which may be of special interest to a small group of researchers. Meanwhile, the identification of interactions takes huge amount of experimental resources and consumes lots of time. This has significantly limited the progress in studying different HPI systems.

As a cost-effective approach, computational models for analysis and predictions of HPI systems have been investigated. Although several literature reviews have been published by introducing the machine learning-based methods and some applications in the HPI domain, little research on empirical evaluations of the performance of HBI predictions based on machine learning models has been ever conducted [27, 31], and no work focusing on the prediction of human-bacterium interactions has been reported. Meanwhile, most studies of PPI predictions have been conducted based on a hypothesis on evaluating the predictor with a balanced and small dataset, in which the numbers of positive and negative PPIs are the same. In order to learn the HBI data in a comprehensive manner, a dedicated experiment setting is desired. Moreover, the prediction performance may vary a lot on HB-PPI dataset using different machine learning models.

To achieve an extensive empirical evaluation of predictions of HB-PPI based on machine learning models, we firstly build human-bacterium protein-protein interaction dataset. Our dataset was curated based on our dedicated and comprehensive review of published databases for the last two decades. We specify our data with either expert annotated interactions or directly experiments outcomes, to build a trustable positive protein interactions dataset. Furthermore, we collected the unlabelled protein interaction data by accessing UnitProtKB database [8], among which the human proteins (taxonomy ID: 9606) were downloaded and the corresponding proteins for the bacterium specie were also acquired. We constructed the typical protein interaction data curation process by following [9,37], meanwhile an extensively dataset curation strategy on top of [13] was included. The details for data curation will be discussed in Section 3.

By building the human-bacterium protein-protein interactions dataset, we found that prediction of HB-PPI posed three challenges for machine learning methods to build robust and efficient models: (C1) Given the data availability and experiments design, a curated HB-PPI dataset for machine learning model is required; (C2) The protein sequence information, which is the preliminary information determining subsequent levels of protein structure, still requires an effective feature representation algorithm to retain their identities; (C3) Different machine learning methods exhibit various performances regarding C1. How to design a robust and effective model remains challenging.

To evaluate solutions tackling challenges C1 and C2, the experiment settings to build the HPI datasets are designed by including, firstly different ratios of positive HB-PPI to negative interactions, and secondly two different categories of sequence feature representation algorithms. Our evaluations of various traditional machine learning methods and models found in the literature review have revealed that, current techniques could not render a robust performance and could not generalise well for the HB-PPI dataset. Fig. 1 illustrates the detail of the overall experimental framework.

Thus, to tackle C3, we have subsequently proposed a bidirectional long short-term memory-based model, jointly learning with the designed multi-channel feature representation algorithm, tree-based feature selection algorithm and synthetic minority over-sampling technique (SMOTE), for the prediction of HB-

PPI dataset. The proposed model demonstrates a superior performance over the others. The details of design will be discussed in Section 4.

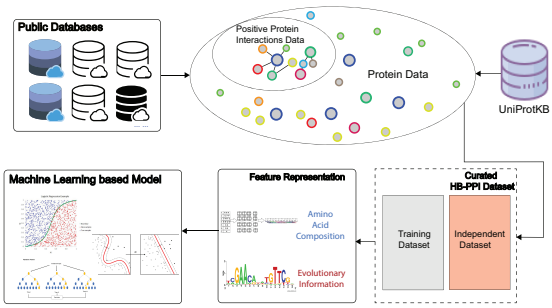

Fig. 1. The Overall Experimental Framework

The contributions of this paper can be summarised as follows:

- (1) A comprehensive and systematic HB-PPI review is achieved, and we have collectively presented different feature representation algorithms and machine learning models for prediction of HB-PPI; (Section 2)
- (2) To address the challenges of C1-C3, we have implemented a broader and deeper experimental framework to revisit the learning task of HB-PPI dataset. The extensively empirical evaluations considering different categories of sequence feature representation algorithms and traditional machine learning methods show that, there is still plenty of room for improvements to achieve a robust and efficient machine learning based method for prediction of HB-PPI; (Section 3)
- (3) We have proposed a model achieving a more robust and effective performance on the HB-PPI datasets of three different HBI systems, based on bidirectional LSTM model with the designed multi-channel feature. The proposed model indicates a promising research direction of studying big HB-PPI dataset with deep learning model. (Section 4)

2 A Comprehensive Re-examination of Host-pathogen Interactions

There have been substantial research interests in applying machine learning methods for prediction of protein-protein interactions [3,13,16,26,31,32,34,37]. A similarity among these works was to have successfully applied machine learning methods in a given positive protein interactions data, whilst their work focused on a balanced protein interactions dataset by building negative protein interactions data with a same number of the positives.

In our work, we will explicitly characterise the prediction tasks of HBI systems from the identified challenges. We formulate our empirical evaluation from two different aspects, which were somehow scarcely investigated in the past.

2.1 Host-Pathogen Protein-Protein Interactions

Prior to conduct the empirical evaluation for HB-PPI prediction, we have carefully reviewed the existing literature reviews. Since there is currently no single review dedicated to HB-PPI, several up-to-date reviews of broad topics on HP-PPI

are evaluated. A wide coverage of HPI study can be found in [31], [10] and [33], which includes the prediction as well as analysis, while research on computational prediction of HPI was discussed in [27] and [41]. Since these reviews aimed at describing the progress in prediction of host-pathogen interactions without anchors of naming pathogens, they have collectively listed potential computational methods. The computational methods include a homology-based approach, a structure-based approach, and a motif interaction-based approach and machine learning-based approach. Furthermore, no systematic evaluation with sufficient details has been implemented and reported in these reviews.

2.2 Variety of Host-pathogen Databases

An systematic literature review has been conducted to screen the abundant HPI resources. There are over 4,000 returning items according to the keywords search of 'pathogen' and 'database' by NCBI PubMed search engine. The first 400 results ranking with best relevance are manually examined with the 'Abstract'. 45 databases have been evaluated by their availability and contents. Eventually, there are 11 databases chosen to curate our dataset of different HPI systems. We focus on those in which human is the host (taxonomy ID: 9606) and the bacterium is the pathogen. These 11 databases are DIP [29], Reactome [20], APID [28], IntAct [21], MINT [24], InnateDB [4], PHISTO [11], PATRIC [35], Mentha [5], HPIDB [2] and BioGRID [6]. All the data sources from 11 databases are collected via literature and domain expert manual verification, which are of high fidelity and confidence.

After cleansing the databases, 90 different bacterium pathogens are identified having interactions with other hosts. In this study, we have dedicated the study between three bacteria and human host, for the reason of their sufficiently available protein information to constitute big datasets for the evaluation and comparison with the proposed model. The others could be used for further repeated verification and research, but are not within the scope of this paper.

3 Evaluation Design for HB-PPI Dataset

As mentioned, although several reviews have discussed the research challenges in HPI prediction, there is neither curated datasets available nor evaluation results presented in the research papers. In this section, we introduce the evaluation design for HB-PPI dataset, as illustrated in Fig. 1.

3.1 The HB-PPI Dataset

Since only a small number of positive protein interaction data are catalogued in public databases and the scale of remaining unknown protein interactions relationships are very huge, most studies of intra-species PPI in the literature adopted random sampling scheme to select protein interactions from the unknown data as the negative protein interactions data to constitute a discriminative dataset for model learning [14, 15, 32, 37]. A balanced protein interactions dataset, which assumed that positive and negative protein interactions data were with the same amount, is normally curated for evaluation.

However, considering HPI systems, which is concerned with the inter-species interactions, the interaction ratio (i.e. the number of positives in a large set

of protein pairs between species) is expected to be very low, which in practice may be set as 25, 50 even 100 times as many negative interactions as positive interactions [13, 23]. In other words, it could be a highly imbalanced dataset. Given the hypothesis, our work strives to evaluate the impact of amount of negative data by reproducing different ratios to generate HB-PPI datasets.

The details of our curated HB-PPI dataset are shown in Table. 1. The taxonomy IDs are listed as the specific bacterium pathogens selected after data pre-processing. They correspond to three different bacterium pathogens actively interacting with human host. To alleviate the impact of randomness in sampling, we have repeated this process for five times, which resulted in a five-fold independent tests for our evaluation.

Table 1. DATASETS STATISTICS

Taxonomy ID[a]	Positive Interactions Number	Ratio 1:25		Ratio 1:50		Ratio 1:100	
		Training	Independent Testing	Training	Independent Testing	Training	Independent Testing
1491	57	1185	297	2325	582	4605	1151
177419	1207	25105	6277	49245	12312	97525	24382
1392	2810	58448	14612	114648	28662	227048	56762

[a] '1491' represents Clostridium botulinum, '177419' is Francisella tularensis subsp. tularensis (strain SCHU S4 / Schu 4), and '1392' is Bacillus anthracis bacterium

3.2 Interpreting the Sequence Information

Since utilizing protein sequence information has become a research trend due to its availability of abundant information, it also solicits novel feature representation algorithms to the ongoing protein researches to improve the prediction performance [9, 39, 40]. In our work, we will focus on sequence information. We anticipate the study can be potentially extended to other related research topics. Thus, mapping the sequence information according to the selected feature representation algorithms is the first step.

Because every different protein possesses different length of amino acid combinations, it will be difficult to directly input the sequence information into the machine learning methods. This raises a great interest for us to develop an efficient and powerful algorithm to retain the identity of proteins. Two different categories are included in our work, namely, amino acid composition methods and evolutionary information methods.

Amino acid composition methods consider the feature representation according to the amino acid combination of a given protein sequence information in different ways, such as their grouping based on different physicochemical characteristics and their order of sequence information. This results in two different popular algorithms, namely the conjoint triad method [32] and auto covariance algorithm [17]. The conjoint triad method categorises twenty types of amino acids into seven groups according to their physicochemical characteristics. The auto covariance algorithm calculates the auto covariance relationship using the order of amino acids in sequence information.

Evolutionary information methods involve a protein alignment process against a reference protein sequence database, which produces a position-specific

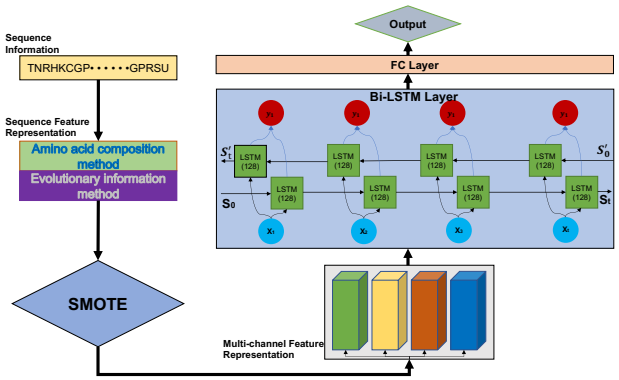

Fig. 2. The Overall Experimental Framework

scoring matrix (PSSM) to indicate the probability of each amino acid type for corresponding position. PSSM is a T*20 matrix for a protein sequence by PSI-BLAST. T denotes the length of protein sequence. In our work, we apply two different methods, which are Pseudo Position-Specific Score Matrix (Pse-PSSM) [7] and Block-PSSM [19]. Pse-PSSM is a 40-dimensional vector, which represents a direct and joint amino acids relationship from the original PSSM. Block-PSSM divides PSSM profile and protein sequence into 20 equal blocks. For each block, a 20-dimensional vector is calculated according to amino acid information.

3.3 Machine Learning based Methods

It is crucial to select feasible machine learning methods to perform HBI prediction task, in which challenges C1 and C2 are inherent. In this paper, we evaluate several popular machine learning models, including support vector machine (SVM), random forests (RF), logistic regression (LR), naïve Bayes model (GNB), decision tree (DT) and gradient boosting machine (GBM). These machine learning models are still more predominant than deep learning methods in protein interaction studies, because they usually require less data and have a simpler architecture, yet achieving a reasonable performance, in contrast to computer vision or other AI problems. For the hyperparameters optimization, five-fold cross-validation test was adopted to select the best parameters. Meanwhile, two sequence-based machine learning models are included [9,37], for comparisons. The two methods [9,37] have applied SVM and RF model accordingly by involving different feature representation algorithms as the learning models.

4 Proposed Bi-LSTM-based Model and Evaluation

4.1 Our Model

Fig. 2 illustrates the novel model we proposed, including different components[1]. **Bidirectional LSTM** model (Bi-LSTM) is the critical component of the model, which is a variant deep learning model of LSTM proposed by [18, 30]. LSTM model and its variant version Bi-LSTM have demonstrated superior performance in domainss such as natural language processing , transportation and action recognition [36, 38]. In Bi-LSTM model, two layers, namely forward and backward layers, are designed to converge into a single layer.

[1] The code and data are available on: https://huaming-chen.com/Bi-LSTM-Predictor/

However, the Bi-LSTM model explicitly suffers from the conventional vanishing gradient problem for the prediction of the highly skewed HB-PPI data. To resolve the problem, we firstly introduce the focal loss function [25] as the cost function Δ in Bi-LSTM model, which is defined in Eq. 1. Normally, cross entropy loss is applied for binary classification, which could be defined as $\Delta(p, y) = \Delta(p_t) = -log(p_t)$. Alternatively, Equa. 1 is defined in our model, where p_t defines the estimated output probability and α_t and γ are the parameters. In this study, $\alpha_t = 0.5$ and $\gamma = 2$ for all the experiments.

$$\Delta(p_t) = -\alpha_t(1 - p_t)^{\gamma} log(p_t) \tag{1}$$

Additionally, we designed a novel three-dimension tensor data as the feature representation algorithm, which is a multi-channel feature in this study. The design of the multi-channel feature benefits from the sequence-based feature representation algorithms. The tree-based feature selection algorithm is employed at first to unify the features to be transformed as multi-channel feature. Once the features are processed, the data will be learnt by SMOTE technique to ease the imbalanced ratio. The output of the SMOTE model will be subsequently stacked horizontally to build the multi-channel feature data, which is then input to Bi-LSTM.

Table 2. Results of F1-score for Pathogen Taxonomy ID '1491' and '1392'

Model		'1491'[a]			'1392'		
		1:25	1:50	1:100	1:25	1:50	1:100
RF	$\Re_1$[b]	0.957±0.000	**0.992±0.016**	0.984±0.020	0.170±0.010	0.140±0.009	0.068±0.007
	$\Re_2$	0.941±0.075	0.959±0.024	0.925±0.083	0.103±0.020	0.079±0.006	0.056±0.013
	$\Re_3$	0.985±0.031	0.969±0.029	0.983±0.021	0.207±0.009	0.166±0.004	0.092±0.014
	$\Re_4$	0.955±0.052	0.992±0.016	**1.000±0.000**	0.198±0.016	0.174±0.008	0.104±0.003
SVM	$\Re_1$	**1.000±0.000**	0.992±0.016	0.984±0.020	0.000±0.000	0.000±0.000	0.000±0.000
	$\Re_2$	0.969±0.029	0.991±0.017	1.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000
	$\Re_3$	1.000±0.000	0.984±0.020	0.957±0.000	0.000±0.000	0.000±0.000	0.000±0.000
	$\Re_4$	1.000±0.000	0.984±0.020	0.957±0.000	0.048±0.029	0.000±0.000	0.003±0.006
LR	$\Re_1$	0.667±0.000	0.406±0.071	0.278±0.009	0.021±0.006	0.000±0.000	0.007±0.000
	$\Re_2$	0.969±0.029	0.992±0.016	0.957±0.000	0.051±0.006	0.012±0.003	0.007±0.003
	$\Re_3$	0.954±0.038	0.939±0.038	0.832±0.100	0.031±0.003	0.000±0.000	0.000±0.000
	$\Re_4$	0.985±0.031	0.985±0.031	0.984±0.020	0.108±0.004	0.042±0.005	0.016±0.003
Naïve Bayes	$\Re_1$	0.883±0.025	0.759±0.083	0.649±0.071	0.105±0.002	0.057±0.000	0.030±0.000
	$\Re_2$	0.911±0.043	0.859±0.038	0.772±0.076	0.109±0.000	0.067±0.001	0.030±0.000
	$\Re_3$	0.852±0.030	0.710±0.093	0.509±0.072	0.115±0.001	0.060±0.001	0.038±0.000
	$\Re_4$	0.852±0.029	0.708±0.099	0.535±0.071	0.117±0.001	0.063±0.000	0.034±0.000
GBM	$\Re_1$	0.941±0.020	0.955±0.052	0.911±0.044	0.158±0.005	0.118±0.004	**0.142±0.017**
	$\Re_2$	0.921±0.052	0.984±0.020	0.829±0.120	0.152±0.007	0.119±0.011	0.093±0.012
	$\Re_3$	0.938±0.055	0.939±0.048	0.876±0.051	0.115±0.009	0.096±0.023	0.091±0.009
	$\Re_4$	0.915±0.091	0.961±0.034	0.856±0.057	0.156±0.013	0.114±0.012	0.101±0.018
DT	$\Re_1$	0.870±0.016	0.867±0.076	0.860±0.070	**0.238±0.014**	0.039±0.013	0.011±0.016
	$\Re_2$	0.768±0.096	0.885±0.082	0.804±0.063	0.085±0.022	0.035±0.007	0.017±0.007
	$\Re_3$	0.935±0.065	0.902±0.063	0.891±0.028	0.235±0.016	0.073±0.009	0.006±0.011
	$\Re_4$	0.893±0.075	0.933±0.054	0.955±0.052	0.035±0.034	**0.187±0.014**	0.080±0.018
Model$_1$[c]		0.693±0.066	0.928±0.023	0.604±0.031	0.046±0.005	0.052±0.004	0.017±0.002
Model$_2$[b]		0.950±0.039	0.976±0.020	0.978±0.044	0.199±0.012	0.152±0.005	0.123±0.015
Proposed Model		0.939±0.038	0.925±0.044	0.969±0.029	**0.281±0.011**	**0.243±0.016**	**0.194±0.011**

[a] '1491' and '1392' represent the taxonomy IDs for the related bacterium pathogen species, details can be found in Section 3.1;
[b] $\Re_1$–$\Re_4$ are the different feature representations algorithms, representing ACC, CTM, PsePSSM and BlockPSSM;
[c] Model$_1$ is the method from [37];[d] Model$_2$ is the method from [9]

As illustrated in Fig. 2, the proposed model is designed with the consideration of the distinct feature of protein sequence information and the imbalanced issue of the HB-PPI datasets. In next section, we will present the complete evaluation performance as well as the proposed model results with regard to F1-score, which is a suitable measurement for our primary evaluation in this research.

4.2 Evaluation and Discussion

For the evaluation, all the data used in the evaluation have been preprocessed with the same protocol according to the relevant literature. Due to the space

limit, the results of pathogens with taxonomy ID '1491' and '1392' are collectively included in Table. 2 and the result of '177419' is included in Table. 3. The first two best performances of each column are indicated by bold font. We can observe that, the performances of different machine learning models for the different dataset vary a lot. ParIt is not easy to identify which one would achieve the best in a combination with an appropriate feature representation algorithm. In Table 2, the overall performance of Model_2 is better than the results from Model_1. However, they are neither the best nor the second best. For different column, the traditional models present different capabilities of the performance.

Table 3. Results of F1-score for Pathogen Taxonomy ID '177419'

Model		'177419'[a]		
		1:25	1:50	1:100
RF	$\Re_1$[b]	0.040±0.014	0.003±0.004	0.000±0.000
	$\Re_2$	0.029±0.015	0.007±0.003	0.008±0.005
	$\Re_3$	0.069±0.014	0.015±0.006	0.005±0.004
	$\Re_4$	0.043±0.013	0.008±0.009	0.002±0.003
SVM	$\Re_1$	0.127±0.014	0.052±0.006	0.027±0.006
	$\Re_2$	0.023±0.006	0.041±0.013	0.052±0.010
	$\Re_3$	0.122±0.011	0.040±0.008	0.000±0.000
	$\Re_4$	0.106±0.014	0.020±0.006	0.000±0.000
LR	$\Re_1$	0.008±0.000	0.000±0.000	0.000±0.000
	$\Re_2$	0.062±0.007	0.011±0.004	0.000±0.000
	$\Re_3$	0.000±0.000	0.000±0.000	0.000±0.000
	$\Re_4$	0.145±0.010	0.082±0.010	**0.056±0.005**
Naïve Bayes	$\Re_1$	0.116±0.001	0.063±0.001	0.036±0.000
	$\Re_2$	0.113±0.001	0.056±0.001	0.029±0.000
	$\Re_3$	0.123±0.003	0.076±0.002	0.040±0.000
	$\Re_4$	0.119±0.001	0.067±0.000	0.035±0.000
GBM	$\Re_1$	0.076±0.009	0.074±0.025	0.041±0.007
	$\Re_2$	0.103±0.024	0.045±0.006	0.037±0.008
	$\Re_3$	0.111±0.007	0.092±0.009	0.048±0.007
	$\Re_4$	0.122±0.017	0.082±0.007	0.051±0.012
DT	$\Re_1$	0.153±0.023	0.017±0.012	0.000±0.000
	$\Re_2$	0.036±0.036	0.020±0.015	0.006±0.006
	$\Re_3$	**0.164±0.017**	0.049±0.006	0.014±0.012
	$\Re_4$	0.002±0.003	**0.106±0.010**	0.020±0.014
Model_1[c]		0.029±0.011	0.005±0.004	0.000±0.000
Model_2		0.109±0.016	0.068±0.011	0.052±0.013
Proposed Model		**0.244±0.012**	**0.186±0.015**	**0.135±0.015**

[a] '177419' represents the taxonomy ID for the related bacterium pathogen specie, details can be found in Section 3.1;
[b] $\Re_1 - -\Re_4$ are the different feature representations algorithms, representing ACC, CTM, PsePSSM and BlockPSSM;
[c] Model_1 is the method from [37];[d] Model_2 is the method from [9]

For our proposed Bi-LSTM-based model, it has achieved a more stable and better performance than the others for HBI systems of ID '1392' and '177419'. These two datasets are much bigger than the one of ID '1491', for which Bi-LSTM-based model has not been the best. However, it still yields results quite smoothly when the ratio changes. Meanwhile, Bi-LSTM-based model also shows a strong capability in dealing with the imbalanced issue. In comparison with the evaluation models and the literature methods, Bi-LSTM-based model has demonstrated a better performance in our study.

5 Conclusion

In this study, our extensive evaluation of HB-PPI is presented. We anticipate in delivering this research work as a first attempt to systematically evaluate machine learning methods for HB-PPI prediction. Three challenges were identified as causing the performance fluctuation in the HBI datasets. Thus, a complete experimental framework in different HBI systems was established to learn and predict from positive and unlabeled protein interactions data.

We have also proposed a Bi-LSTM-based model achieving a more robust and effective performance. Although the performance is better than the others, it is

expected to propose a more powerful approach to harness the protein information and design a sophisticated machine learning models for prediction in the future.

References

1. Ahmed, H.R., et al.: Pattern discovery in protein networks reveals high-confidence predictions of novel interactions. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. pp. 2938–2945 (2014)
2. Ammari, M.G., et al.: Hpidb 2.0: a curated database for host–pathogen interactions. Database 2016 (2016)
3. Ben-Hur, A., et al.: Kernel methods for predicting protein–protein interactions. Bioinformatics 21(suppl_1), i38–i46 (2005)
4. Breuer, K., et al.: Innatedb: systems biology of innate immunity and beyond—recent updates and continuing curation. Nucleic acids research 41(D1), D1228–D1233 (2013)
5. Calderone, A., et al.: Mentha: a resource for browsing integrated protein-interaction networks. Nature methods 10(8), 690–691 (2013)
6. Chatr-Aryamontri, A., et al.: The biogrid interaction database: 2017 update. Nucleic acids research 45(D1), D369–D379 (2017)
7. Chou, K.C., et al.: Memtype-2l: a web server for predicting membrane proteins and their types by incorporating evolution information through pse-pssm. Biochemical and biophysical research communications 360(2), 339–345 (2007)
8. Consortium, U., et al.: Uniprot: the universal protein knowledgebase. Nucleic acids research 46(5), 2699 (2018)
9. Cui, G., et al.: Prediction of protein-protein interactions between viruses and human by an svm model. In: BMC bioinformatics. vol. 13, p. S5. Springer (2012)
10. Durmuş, S., et al.: A review on computational systems biology of pathogen–host interactions. Frontiers in microbiology 6, 235 (2015)
11. Durmuş Tekir, S., et al.: Phisto: pathogen–host interaction search tool. Bioinformatics 29(10), 1357–1358 (2013)
12. Durmus Tekir, S., et al.: Infection strategies of bacterial and viral pathogens through pathogen–human protein–protein interactions. Frontiers in Microbiology 3, 46 (2012)
13. Dyer, M.D., et al.: Supervised learning and prediction of physical interactions between human and hiv proteins. Infection, Genetics and Evolution 11(5), 917–923 (2011)
14. Eid, F.E., et al.: Denovo: virus-host sequence-based protein–protein interaction prediction. Bioinformatics 32(8), 1144–1150 (2016)
15. Emamjomeh, A., et al.: Predicting protein–protein interactions between human and hepatitis c virus via an ensemble learning method. Molecular Biosystems 10(12), 3147–3154 (2014)
16. Gomez, S.M., et al.: Learning to predict protein–protein interactions from protein sequences. Bioinformatics 19(15), 1875–1881 (2003)
17. Guo, Y., et al.: Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic acids research 36(9), 3025–3030 (2008)
18. Hochreiter, S., et al.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
19. cheol Jeong, J., et al.: On position-specific scoring matrix for protein function prediction. IEEE/ACM transactions on computational biology and bioinformatics 8(2), 308–315 (2010)

20. Joshi-Tope, G., et al.: Reactome: a knowledgebase of biological pathways. Nucleic acids research 33(suppl_1), D428–D432 (2005)
21. Kerrien, S., et al.: The intact molecular interaction database in 2012. Nucleic acids research 40(D1), D841–D846 (2012)
22. König, R., et al.: Global analysis of host-pathogen interactions that regulate early-stage hiv-1 replication. Cell 135(1), 49–60 (2008)
23. Kshirsagar, M., et al.: Multitask learning for host–pathogen protein interactions. Bioinformatics 29(13), i217–i226 (2013)
24. Licata, L., et al.: Mint, the molecular interaction database: 2012 update. Nucleic acids research 40(D1), D857–D861 (2012)
25. Lin, T.Y., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
26. Nanni, L., et al.: An empirical study of different approaches for protein classification. The Scientific World Journal 2014 (2014)
27. Nourani, E., et al.: Computational approaches for prediction of pathogen-host protein-protein interactions. Frontiers in microbiology 6, 94 (2015)
28. Prieto, C., et al.: Apid: agile protein interaction dataanalyzer. Nucleic acids research 34(suppl_2), W298–W302 (2006)
29. Salwinski, L., et al.: The database of interacting proteins: 2004 update. Nucleic acids research 32(suppl_1), D449–D451 (2004)
30. Schuster, M., et al.: Bidirectional recurrent neural networks. IEEE transactions on Signal Processing 45(11), 2673–2681 (1997)
31. Sen, R., et al.: A review on host–pathogen interactions: classification and prediction. European Journal of Clinical Microbiology & Infectious Diseases 35(10), 1581–1599 (2016)
32. Shen, J., et al.: Predicting protein–protein interactions based only on sequences information. Proceedings of the National Academy of Sciences 104(11), 4337–4341 (2007)
33. Soyemi, J., et al.: Inter-species/host-parasite protein interaction predictions reviewed. Current bioinformatics 13(4), 396–406 (2018)
34. Wang, X., et al.: A novel matrix of sequence descriptors for predicting protein-protein interactions from amino acid sequences. PloS one 14(6) (2019)
35. Wattam, A.R., et al.: Patric, the bacterial bioinformatics database and analysis resource. Nucleic acids research 42(D1), D581–D591 (2014)
36. Wu, J., et al.: Towards a general prediction system for the primary delay in urban railways. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). pp. 3482–3487. IEEE (2019)
37. Wuchty, S.: Computational prediction of host-parasite protein interactions between p. falciparum and h. sapiens. PLoS One 6(11) (2011)
38. Yao, Y., et al.: Bi-directional lstm recurrent neural network for chinese word segmentation. In: International Conference on Neural Information Processing. pp. 345–353. Springer (2016)
39. Zhang, J., et al.: Review and comparative assessment of sequence-based predictors of protein-binding residues. Briefings in bioinformatics 19(5), 821–837 (2018)
40. Zhang, L.: Sequence-based prediction of protein-protein interactions using random tree and genetic algorithm. In: International Conference on Intelligent Computing. pp. 334–341. Springer (2012)
41. Zhou, H., et al.: Progress in computational studies of host–pathogen interactions. Journal of bioinformatics and computational biology 11(02), 1230001 (2013)