Systematic evaluation of machine learning methods for identifying human-pathogen protein-protein interactions

Huaming Chen[1], Fuyi Li[2,3], Lei Wang[1], Yaochu Jin[4], Chi-Hung Chi[5], Lukasz Kurgan[6,]*, Jiangning Song[2,3,7,*], Jun Shen[1,*]

[1]Faculty of Engineering and Information Science, School of Computing and Information Technology, University of Wollongong, NSW 2522, Australia;

[2]Monash Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia;

[3]Monash Centre for Data Science, Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia;

[4]Department of Computer Science, University of Surrey, Guildford, Surrey GU2 7XH, United Kingdom;

[5]Data61, Commonwealth Scientific and Industrial Research Organization (CSIRO), Sandy Bay, TAS 7005, Australia;

[6]Department of Computer Science, Virginia Commonwealth University, 401 West Main Street, Room E4225, Richmond, VA 23284, USA;

[7]ARC Centre of Excellence in Advanced Molecular Imaging, Monash University, Melbourne, VIC 3800, Australia.

[*]To whom correspondence should be addressed: (1) Jun Shen, Faculty of Engineering and Information Science, School of Computing and Information Technology, University of Wollongong, NSW 2522, Australia. Tel: +61-2-4221-3873; Email: jshen@uow.edu.au; (2) Lukasz Kurgan, Department of Computer Science, Virginia Commonwealth University, 401 West Main Street, Room E4225, Richmond, VA 23284, USA. Tel: (804) 827-3986; E-mail: lkurgan@vcu.edu; (3) Jiangning Song, Biomedicine Discovery Institute, Department of Biochemistry and Molecular Biology, and Monash Centre of Data Science, Monash University, Melbourne, VIC 3800, Australia. Tel: +61-3-9902-9304; Email: Jiangning.Song@monash.edu.

Key words: bioinformatics; human-pathogen interactions; protein-protein interactions; systematic evaluation; sequential analysis; machine learning

Running Head: Systematic evaluation of predictors for HP-PPIs

Author Biographies:

Huaming Chen received his B.Eng. and M.Eng. degrees from Lanzhou University, China, in 2012 and 2015, respectively. He is currently a Ph.D. candidate with the University of Wollongong. His research interests are bioinformatics, machine learning, and neural network.

Fuyi Li received his BEng and MEng degrees from Northwest A&F University, China. He is currently a PhD candidate in the Department of Biochemistry and Molecular Biology and Biomedicine Discovery Institute, Monash University, Australia. His research interests are bioinformatics, computational biology, machine learning, and data mining.

Lei Wang received his B.Eng. and M.Eng. degrees from Southeast University, China, in 1996 and 1999, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2004. He is currently an Associate Professor with the Faculty of Informatics, University of Wollongong.

Yaochu Jin received his B.Sc., M.Sc., and Ph.D. degrees from Zhejiang University, Hangzhou, China, in 1988, 1991, and 1996, respectively, and the Dr.-Ing. degree from Ruhr University Bochum, Bochum, Germany, in 2001. He is a Professor in the Department of Computer Science, University of Surrey, United Kingdom. His research interests include computational intelligence, computational systems biology and nature-inspired problem solving.

Chi-Hung Chi received the Ph.D. degree from Purdue University, USA. He is currently a Senior Principal Research Scientist of Data61 in CSIRO (Commonwealth Scientific and Industrial Research Organization), Australia. His research areas include cybersecurity, behaviour modeling, knowledge graph, data engineering and analytics, cloud and service computing, social computing, Internet-of-Things, and distributed computing.

Lukasz Kurgan is a Robert J. Mattauch Endowed Professor of Computer Science at the Virginia Commonwealth University. He is a Fellow of American Institute for Medical and Biological Engineering (AIMBE) and has published close to 150 peer-reviewed journal articles that focus on structural and functional characterization of proteins and small RNAs. More details about his research group are available at http://biomine.cs.vcu.edu/.

Jiangning Song is an associate professor and group leader in the Biomedicine Discovery Institute, Monash University, Australia. He is also affiliated with the Monash Centre for Data Science, Faculty of Information Technology, Monash University. His research interests include bioinformatics, computational biology, machine learning, data mining, and pattern recognition.

Jun Shen received the Ph.D. degree from Southeast University, China, in 2001. He is currently an Associate Professor with the University of Wollongong, Wollongong, NSW, Australia. His expertise is on cloud computing and big data.

ABSTRACT

In recent years, high-throughput experimental techniques have significantly enhanced the accuracy and coverage of protein-protein interaction identification, including human-pathogen protein-protein interactions (HP-PPIs). Despite this progress, experimental methods are, in general, expensive in terms of both time and labor costs, especially considering that there are enormous amounts of potential protein-interacting partners. Developing computational methods to predict interactions between human and bacteria pathogen has thus become critical and meaningful, in both facilitating the detection of interactions and mining incomplete interaction maps. In this paper, we present a systematic evaluation of machine-learning-based computational methods for human-bacterium protein-protein interactions (HB-PPIs). We first review a vast number of publicly available databases of HP-PPIs, and then critically evaluate the availability of these databases. Benefitting from its well-structured nature, we subsequently preprocess the data and identified six bacterium pathogens that could be used to study bacterium subjects in which a human was the host. Additionally, we thoroughly reviewed the literature on “host-pathogen interactions” whereby existing models were summarized that we used to jointly study the impact of different feature-representation algorithms and evaluate the performance of existing machine-learning computational models. Owing to the abundance of sequence information and the limited scale of other protein-related information, we adopted the primary protocol from the literature and dedicated our analysis to a comprehensive assessment of sequence information and machine-learning models. A systematic evaluation of machine-learning models and a wide range of feature-representation algorithms based on sequence information are presented as a comparison survey towards the prediction performance evaluation of HB-PPIs.

1 INTRODUCTION

Infectious diseases are predominantly caused by many pathogenic species, such as fungi, viruses, bacteria and so on. These infectious species actively interact with their hosts in a variety of ways, which place host-pathogen interactions (HPIs) in a complicated, but also critical, role in the study of infectious-disease mechanisms. In most cases, the host-pathogen system is studied from different perspectives to further our understanding of infectious mechanisms [1]. A major approach is studying the interactions of inter-species proteins, in which one protein is from the host and the other is from the pathogen.

While protein interactions occur extraordinarily between human and bacterium pathogens, one of the earliest studies illustrated the importance of human-bacterium interactions (HBI) in relation to the symptoms caused by anthrax [2]. In this study, *Bacillus anthracis* was conclusively demonstrated as the primary cause of anthrax. Additional studies of *Bacillus anthracis* were conducted, aimed at fully understanding the mechanisms of a complete protein interaction network between *Bacillus anthracis* (the bacterium pathogen) and *Homo sapiens* (the host) [3, 4]. These studies encouraged researchers to study a broad range of infectious diseases by exploring human-bacterium protein-protein interactions (HB-PPIs).

However, the investigation of HBIs consumes lots of time, money and resources in determining the complete interaction network and understanding their mechanisms. Currently, investigations of the interactions between host and pathogens are still very limited. Even though large-scale biomedical technologies, such as yeast two-hybrid assay and the affinity purifications-mass spectrometry (AP-MS) method, have allowed us to detect interactions (positive or negative) in a faster and more accurate way, the amount of possible human-bacterium protein-protein interactions is large. Other small-scale technologies, like nuclear magnetic resonance (NMR), are often labour-intensive and time-consuming. Thus, it is critical to formulate a computational model for the prediction of HB-PPIs.

Several reviews studied current computational approaches [5, 6] as well as researches on applying machine-learning-based models to predict host-pathogen protein-protein interactions (HP-PPIs) [7-11]. In particular, how to deploy machine-learning models as a generic approach in predicting novel HBIs based on sequence information is considered as an important category of research, which involves many challenges and opportunities. However, there is currently no comprehensive evaluation study that has focused on machine-learning models as the primary computational method and further comparatively evaluated their corresponding performances across a wide range of HBI systems.

In this paper, we implemented an evaluation protocol based on literature reviews by first collecting HBI data from a wide range of host-pathogen databases. The systematic evaluation was subsequently achieved from two aspects. The first considered the application of feature-representation algorithms to the protein data, while the other was related to different machine-learning-based models.

The remainder of this paper is organised as follows. Section 2 summaries the literature review from four different perspectives, including the review of host-pathogen interaction studies, the review of available host-pathogen interaction databases, the review of computational methods for host-pathogen protein-protein interaction predictions and the review of sequential-representation algorithms and the machine-learning-based methods for prediction. In Section 3, the materials collected for evaluation and the details of curated datasets are presented. Section 4 discusses the evaluation results in detail, and we conclude the paper in Section 5.

2 LITERATURE REVIEW

Although there has been a long history of research on PPI prediction, so far there are only a small number of publications that have focused on host-pathogen interaction reviews [5, 6, 12, 13]. A broad search has resulted in four major review papers, and **Table 1** summarizes the reviews. The studies by [6] and [12] have a wide coverage of HPIs, which include predictions as well as analyses, while the reviews by [5] and [13] focused on the computational prediction of HPIs. These reviews aimed at describing the progress of HPIs, without anchors of naming pathogens, and they collectively reported on potential computational methods such as homology-based approaches, structure-based approaches, domain and motif interaction-based approaches, and machine-learning-based approaches. Furthermore, no systematic evaluation with details was implemented or reported in these reviews. Recently, [14] conducted a sequence-based predictors review, however they focused on the prediction of protein-binding residues via single-sequence methods.

Adapted from these reviews, we subsequently collected all published predictors that focused on HB-PPIs and HP-PPIs, which are summarized in **Table 2**. The frameworks of the two different types of computational models for predicting HP-PPIs, including machine-learning-based models and template-based models, are shown in **Figure 1**.

[Table 1]

[Figure 1]

A template-based model utilises different types of protein information to build the prediction model, including sequence information, structure information and domain information [15-17]. Template-based models use different protein information to detect high score homology which might yield similar functions. However, template-based models may fail to predict whether the remote homology is interacting with known proteins. Another type of computational model is based on machine-learning models. The protein information is first vectorised as the input to learn their inherent relationships automatically, which are thus used to build the model and predict the interactions. Specifically, for PPIs, the relevant protein information can be sequence information, gene ontology information, domain information, gene expression information and interaction network information.

As indicated in **Table 2**, numerous feature-representation algorithms for sequence information are incorporated with different machine-learning models for predicting HP-PPIs [7, 8, 10, 15, 16, 18-21]. In this regard, we first grouped the sequential feature-representation algorithms into three different types: amino acid composition, pseudo-amino acid composition and evolutionary information. It should be noted that, not only the reported algorithms in **Table 2**, but also the related sequential-representation algorithms from other protein sequence-specific topics, such as protein structure, protein folding topics, are included in this section. The models from [8] and [19], which are shown in **Table 2**, were selected as the representative models regardless of the pathogen species.

[Table 2]

2.1 HPI Databases

There has been continuous effort spent on developing online HPI databases and repositories by many researchers. These developments mostly benefited from the National Institute of Allergy and Infectious Diseases (NIAID), which initialized a strategic plan to focus on biodefense research. Several ‘priority pathogens’ were defined. Several initial developments, including pathogen interaction gateway (PIG [22]), BioHealthBase [23] and the Pathosystems Resource Integration Center (PATRIC [24]), were wholly or partially funded by the NIAID.

The first web-based database with massive annotated records for pathogen research was the Ecological Database of the World’s Insect Pathogens (EDWIP) [25]. EDWIP uses a one-to-one interaction relationship, which records the infection between a single host species and a single pathogenic species. This strategy resulted in 9,400 records between 4,454 host species and 2,285 pathogen species when it was first released in 2003. PIG was designed as a collection of a number of public resources, which focussed on experimentally verified and manually curated HP-PPIs. This centralized database served as an easy-to-use database which transfers search results to the relevant database, such as the UniProt [26] database. Another important host-pathogen interaction database is the Pathogen-Host Interaction Search TOol (PHISTO) [27]. This tool aims to provide researchers with a complete coverage of HPI data via monthly updates. Proteomics Standards Initiative Common Query InterfaCe (PSICQUIC) [28] service was installed to allow access to and extraction of HPI data the other web-based databases.

Although EDWIP is no longer available online, it is still of particular interest for pathologists and ecologists to collect and analyze the HPI data. Concerning the HPI databases, one of the most critical factors in building a trustable database is the data sources. Typically, there are several different sources. One primary approaches is to collect data from literature and domain expert manual verification, such as the Database of Interacting Proteins (DIP) [29], Reactome [30]. Another approach is to collect data submitted from users. Finally, data can also be novel derived/predicted data from computational models, such as the Pathogen-Host Interaction Data Integration and Analysis System (PHIDIAS) [31] and the Penicillium-Crop Protein–Protein Interactions database (PCPPI) [32]. After the development of DIP and EDWIP, the HPI databases have become more interactive for the users. From Table 2, we can see that, the most commonly used databases for HPI study, including DIP [29], IntAct [33], Mentha [34], and the pathogen-host interaction search tool (PHISTO) [27] and so on, have been serving for different purposes. According to the evaluation and data analysis of the databases, Mentha [34] has covered most of Homo sapiens interaction information, while PHISTO focuses on studying human as the host species.

We, herein, have reviewed numerous publicly available databases, whose results were returned by searching specific keywords in the NCBI PubMed search engine. We manually examined the abstracts of the first 400 results ranked by 'best relevance' out of more than 4,000 returned items based on the keywords 'pathogen' and 'database'. As such, in this paper, a selection of 11 databases is reviewed and evaluated based on their contents. Details are provided in the following sections.

2.2 Sequential-Representation Algorithms

To encode proteins as feature vectors, several different features have been included in this study to predict PPIs between *Homo sapiens* and bacterium pathogens, which are: (1) protein amino acid composition information [35-37], (2) protein pseudo-amino acid composition information [38-40], and (3) protein evolutionary information features [41, 42]. We discuss the related feature-encoding algorithms below.

2.2.1 Amino acid composition

** Conjoint Triad Method*

It was proposed by [35] to classify 20 amino acids into seven groups according to the dipole scale and volume scale of each amino acid, which describe their respective electrostatic and hydrophobic properties. Since this proposal, several variations of encoding algorithms for sequence representation have been devised based on this classification scheme. Among these, one popular approach is to consider the relationship of the properties of one amino acid and its vicinal amino acids as a descriptor [35], which is named the conjoint triad method (CTM). The conjoint triad information of several adjacent amino acids makes it easy to represent every single protein sequence as a class-based feature with the same length, which is also called its *k-mer* feature. Each amino acid type is indicated as a number ranging from *1-7* according to its group. The frequency of three conjoint triad data (*3-mer*) of a sequence is calculated. In total, there will be a combination set including *{(1,1,1), (1,2,1), ..., (1,7,1), ..., (1,7,7), ..., (7,7,7)}*. As a result, *3-mer* features will encode a sequence to a vector of 343 dimensions. For other *2-mer*, *4-mer* and *5-mer* features, the features number would be 49, 2401 and 16807, respectively.

** Auto covariance*

The auto covariance (AC) relationship among the amino acids based on the order of the sequence information was utilised in another feature representation algorithm by [36]. It is a popular transformation algorithm used to adopt numerical vectors to uniform matrices by analysing sequences in the auto cross covariance (ACC) information. Between two different vectors, there are two covariance relationships: cross covariance (CC) and ACC. Only ACC variables are calculated [36]. The basic idea is to derive the physicochemical properties of the amino acid, which include its hydrophobicity (H), volume of side chains (VSCs), polarity (P1), polarizability (P2), solvent-accessible surface area (SASA) and the net charge index of the side chains (NCISC).

In the AC method, each single protein sequence is first translated into a numerical value corresponding to seven different physicochemical properties. Because the ranges of these seven physicochemical properties vary a lot from each other, a first step to normalize the numerical values is required. These values were hence normalized to a distribution whose mean is zero and the standard deviation is one. The normalization equation is shown in Eq. 1:

$$\overline{\mathrm{P}_{\mathrm{i,j}}}= \frac{\mathrm{p}_{\mathrm{i,j}}\text{-}\mathrm{mean}_{\mathrm{j}}}{\mathrm{sd}_{\mathrm{j}}}\ (i = 1, 2, 3, \ldots, 20; j = 1, 2, 3, 4, 5, 6, 7) \qquad (1)$$

where $p_{i,j}$ represents the jth property value of the ith amino acid, $mean_j$ is the mean value of the jth property over the 20 amino acids, and sd_j is the standard deviation of the jth property over the 20 amino acids. Via this operation, every protein sequence is translated into an $N * M$ matrix with zero mean and a standard deviation of unity in each column. With a proper range of these numerical values for each single protein sequence, AC can be used to represent them in a uniform matrix. Based on Eq. 2, a matrix of lg$*$ 7 was calculated, where lag is the distance threshold between two amino acids, and $0 <$ lg $\leq lag$.

$$AC(lag, j) = \frac{1}{N-lag} \sum_{i=1}^{N-lag} \left(p_{i,j} - \frac{1}{N}\sum_{i=1}^{N} p_{i,j}\right) * \left(p_{i+lag,j} - \frac{1}{N}\sum_{i=1}^{N} p_{i,j}\right) \qquad (2)$$

For z properties chosen from the seven physicochemical properties, the length of AC is $lag * z.\ p_{i,j}$, which corresponds to the value from $\{p_{i,j}\}$. Here, N is the length of the protein sequence. After ACC transformation, a representation of the PPI is a concatenation of these two AC transform calculation results.

** Local Descriptor*

Another sequence-based feature-representation method is a local descriptor [37]. The most important feature of an HP-PPI is that the interaction often occurs in some specific intermittent fragments. To better extract this continuous or discrete knowledge from sequence information, [37] proposed using region descriptors to first divide a protein sequence into 10 regions via six different methods: quarter regions, half regions (E, F), central 50% region (G), first 75% region (H), last 75% region (I) and the central 75% region (J). With these 10 regions, a local descriptor is utilized to transform the region sequence into three related descriptors [37]: composition (C), transition (T) and distribution (D). C is the composition ratio of each group of amino acid within a separate region, T represents the percentage of which amino acid group is followed by another amino acid group, and D describes the specific location information obtained by selecting the first, 25%, 50%, 75% and last of each amino acid group. When using a local descriptor, the extracted feature vector contains seven C features, 21 T features and 35 D features. When multiplied by 10 different local regions, the local-descriptor method generates 630 features for a single protein sequence. For an HB-PPI pair, this local descriptor contains 1260 features.

There are also some other schemes that can be used to extract different types of features of a protein sequence, for example the Moran autocorrelation score [43] and the amino acid triplet [8]. As protein sequence information is directly linked to PPI, a further novel representation of PPIs, especially for HB-PPIs, might include any other information related to the specific host species and pathogenic species, which may be a better alternative for predicting HP-PPIs [9].

2.2.2 Pseudo-amino acid composition

Directly converting a protein sequence to a vectorised feature according to the amino acid composition (AAC) might result in sequence-order information loss. The pseudo-amino acid composition (PseAAC) method was proposed as a novel protein sequence representation of a discrete model, which has remarkable prediction performance as an important feature-representation algorithm [38, 44-47].

Various modes of PseAAC have been introduced in the literature. The key is to combine the sequence order correlation information from the protein sequence. In the work of [38], the original version of PseAAC was introduced, as shown in Eq. 3:

$$\begin{aligned}
&\theta_1 = \frac{1}{T-1}\sum_{i=1}^{T-1} \ominus\,(S_i, S_{i+1}) \\
&\theta_2 = \tfrac{1}{T-2}\textstyle\sum_{i=1}^{T-2} \ominus\,(S_i, S_{i+2}) \qquad\qquad \lambda < T \\
&\cdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (3) \\
&\theta_\lambda = \frac{1}{T-\lambda}\sum_{i=1}^{T-\lambda} \ominus\,(S_i, S_{i+\lambda})
\end{aligned}$$

Here, the $\ominus$ function is calculated by Eq. 4:

$$\ominus\,(S_i, S_j) = \tfrac{1}{3}\,\{[H_1(R_j) - H_1(R_i)]^2 + [H_2(R_j) - H_2(R_i)]^2 + \ [M(R_j) - M(R_i)]^2\} \qquad (4)$$

where $H_1(R_i)$, $H_2(R_i)$ and $M(R_i)$ are the corresponding physical-chemical properties of the amino acid residue R_i. Equation 4 produces a λ-dimension vector.

2.2.3. Evolutionary Information

** Position-Specific Scoring Matrix (PSSM)*

By scanning a unique sequence against a reference database, the compilation of a set of alignment profiles results in a position-specific scoring matrix PSSM) of the sequence, which indicates the probability of the corresponding positions of the amino acids [41]. The PSSM is returned as a $T *$ 20 matrix for a given protein sequence by position-specific iterated BLAST (PSI-BLAST). Here, T denotes the length of the corresponding protein sequence. Transformation of the PSSM, which involves highly and broadly homologous sequence information, has been widely used in sequence-related studies [42, 48-53]. These studies indicated that including evolutionary information for feature representation helps to improve prediction model performance.

In detail, given a protein sequence $S = S_1S_2S_3 \dots S_T$, where T is the length of the protein sequence, the corresponding PSSM, $P = \{P_{m,n}\}, m = 1, 2, \dots, T; n = 1, \dots, 2$, is calculated based on the amino acid similarity matrix. The matrix used can be either a point-accepted mutation (PAM, such as Dayhoff's mutation matrix [54]) or a position-weight matrix (PWM, such as the block substitution matrix BLOSUM [55]). The PSSM is calculated according to Eq. 5:

$$P_{m,n} = \sum_{k=1}^{20} w(m,k) * \theta(n,k) \tag{5}$$

where $w(m,k)$ is the probability that the k_{th} amino acid appears at position m, and $\theta(n,k)$ is the value of the position of (n,k) in the similarity matrix.

In this study, PSI-BLAST was employed to create PSSMs with three iterations, where the e-value was set to 0.001. Accordingly, the various lengths of the protein sequences resulted in matrices with different dimensions, which introduced different encoding features based on the PSSM profiles. The following parts present several PSSM-based features-representation algorithms.

** Pse-PSSM*

The pseudo position-specific score matrix (Pse-PSSM) was first introduced for the task of predicting whether or not an uncharacterized protein was a membrane protein [44]. Pse-PSSM extends the idea of corrupting the PSSM descriptor vertically as a mean value, as shown in Eq. 7, where the value of the PSSM is first processed by a standardization procedure horizontally by rows in Eq. 6. The concept of the PseAAC is to generate correlation information between different amino acid locations.

$$p'_{m,n} = \frac{p_{m,n} - \frac{1}{20}\sum_{k=1}^{20} p_{m,k}}{\sqrt{\frac{1}{20}\sum_{k=1}^{20}(p_{m,k} - \frac{1}{20}\sum_{k=1}^{20} p_{m,k})^2}} \tag{6}$$

$$\overline{p_n} = \frac{1}{T}\sum_{m=1}^{T} p'_{m,n} \quad (n = 1, 2, \dots, 20) \tag{7}$$

Thus, the original PSSM profile is converted to a 20-dimension vector, $\bar{p} = \{\overline{p_n}, n = 1, 2, \dots, 20\}$. This derived feature focuses on representing the average score of each amino acid type according to the reference database, which loses the sequence-order information of the protein. Thus, [44] proposed considering supplementary information from the PseAAC, which slices the PSSM profile according to Eq. 8:

$$p_{se_n} = \frac{1}{T-c} \sum_{m=1}^{T-c} [p'_{m,n} - p'_{(m+c),n}]^2 \ \ (n = 1, 2, \dots, 20; c < T) \tag{8}$$

This process generates a 40-dimension vector $P_{se} = \{\overline{P_1}, \overline{P_2}, \overline{P_{20}}, \overline{P_{se_1}},\ \overline{P_{se_2}}, \dots ,\ \overline{P_{se_{20}}}\}$ where $0 < c < min\ (T)$. For a given set of protein sequences, the upper bound of c should be smaller than the shortest length of the protein sequences.

* *Block-PSSM*

By considering the PSSM profile in a dimension format of $T * 20$, [56] proposed dividing the whole sequence into 20 equal blocks, where each represents five percent of the total sequence. Each block generates a 20-dimension vector, which is finally combined as a 20*20=400-dimension vector.

The ith block is calculated according to Eq. 9:

$$pblock_{i,j} = \frac{1}{B_i} \sum_{i=1}^{B_i} p_{i,j} \ \ i = 1, 2, ..., 20; j = 1, 2, \dots, 20 \tag{9}$$

where i represents the block number. Since each five percent of a sequence is considered as a block, i ranges from 1 to 20. j is the number of amino acid types. In short, $pblock_{i,j}$ is extracted as a 1*20 vector, thus $pblock = pblock_1, pblock_2, \dots, pblock_{20}$ is calculated as the Block-PSSM feature in the form of 1*400 vector feature.

* *DPC-PSSM*

Another variation of the PSSM-based feature was proposed by [57]. The original PSSM profile is scaled to the range 0 to 1 by following a sigmoid function, as shown in Eq. 10:

$$p''_{m,n} = \frac{1}{1+e^{-p_{m,n}}} \tag{10}$$

where $p''_{m,n}$ is also used in the transition probability composition PSSM. The AAC-PSSM method is used to extract the corresponding AAC information from $p = \{p''_{m,n}, m = 1, \dots, T;\ n = 1, \dots, 20\}$. The vector in Eq. 11 represents the average mutation score of the amino acid types in the protein during the evolution process, namely the AAC-PSSM. This calculation generates a 20-dimension feature vector.

As a supplementary approach, the traditional dipeptide composition (DPC) of the protein sequence was extended by[57], which was named DPC-PSSM. The calculation of DPC-PSSM is based on the covariance between two adjacent amino acid residues, denoted in Eq. 12. This process produces a 400-dimension feature vector.

$$P_{aac_n} = \frac{1}{T}\sum_{m=1}^{T} p''_{m,n} \qquad m = 1, \dots, T; n = 1, \dots, 20 \tag{11}$$

$$P_{dpc_{i,j}} = \frac{1}{T-1}\sum_{k=1}^{T-1} p''_{k,i} * p''_{(k+1),j} \qquad i,j = 1, \dots, 20 \tag{12}$$

2.3 Machine-Learning-Based Methods for Prediction

Applying computational approaches to predict bioinformatics tasks is considered an important supplementary method for identifying specific targets and high-fidelity interactions in experiments. Recently, we have witnessed numerous applications focusing on domains containing an abundance of unknown data, which require hypothesis verification [5, 13, 58, 59].

In **Table 2**, the predictors of [8, 19], which are based on machine-learning methods and protein sequence information, were selected for our study. These machine-learning models include support vector machine (SVM) and random forest (RF). In this section, we will first briefly review most of the potential machine-learning models that can be utilized for HB-PPI prediction, which include logistic regression (LR), the Na ïve Bayes (NB) model and gradient boosting machine (GBM). These models have demonstrated their capability in other applications for protein structure prediction; however, this is the first time they have been presented in an overall performance evaluation in relation to different feature-representation algorithms for HB-PPIs.

** Support Vector Machine* (*SVM*)

SVM is one of the most widely used models in the literature, which was originally developed by [60]. The introduced structural risk minimization theory ensures the performance of SVM to be widely and successfully applied to many classification and regression tasks in computational biology. SVM contains a radial basis function (RBF) kernel, which is given the task of classifying HP-PPI pairs [8, 11]. Given a dataset of HB-PPIs denoted as $\{x_i, y_i\}, i = 1, 2, \ldots, N,$ where $x_i \in R^n$ and $y_i \in \{+1, -1\}$, y_i is calculated as shown in Eq. 17:

$$y(x) = sign\;[\textstyle\sum_{i=1}^{N} y_i \alpha_i * K(x, x_i) + b] \tag{17}$$

where $K(x, x_i) = \exp(-\gamma \left\|x_i - x_j\right\|^2)$ stands for the RBF kernel, and α_i contains the parameters from a convex quadratic programming problem.

* *Decision Tree*

The decision tree (DT) was designed as a non-parametric supervised model [61]. It uses a tree-like graph to predict an incoming instance based on learned decision rules from given data samples and represented features. DTs are simple to understand and interpret, and they are capable of handling both numerical and categorical data.

* *Random Forest* (*RF*)

Derived from the DT model, RF adopts a random learning method to construct a combination of DTs [62]. RF has superior performance compared with other machine-learning algorithms for classification tasks [63, 64], regression tasks and so on. Technically, RF is an ensemble learning model based on the tree bagging method, which builds a bunch of random DTs to avoid the latent problem caused by potentially biased data. In this study, we implemented RF using the scikit-learn toolkit [65] in Python.

* *Logistic Regression* (*LR*)

LR is an important machine-learning model, which targets modelling y_i between 0 and 1 given unseen data x_i [66, 67]. Accordingly, the LR returns results via Eq. 18:

$$P(y_i = 1|x_i) = h_\theta(x_i) = 1/(1 + \exp(-\theta^T * x_i))$$

$$P(y_i = 0|x_i) = 1 - P(y_i = 1|x_i) = 1 - h_\theta(x_i) \tag{18}$$

where θ is the combination of the model parameters, and the optimization of θ is solved with either the cross-entropy function J_1 or the mean square error loss function J_2, as shown in Eq. 19:

$$J_1(\theta) = -\sum_i (y_i \log\big(h_\theta(x_i)\big) + (1 - y_i)\log(1 - h_\theta(x_i)))$$

$$J_2(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h_\theta(x_i))^2 \tag{19}$$

* Na ïve Bayes Model

Based on Bayes' theorem [68, 69], the Na ïve Bayes model consists of a probabilistic classifier and considers features as independent variables when the class label is given. Given $X = (x_1, x_2, \dots, x_n)$, where x_i is the i_{th} feature, the probability of being in category y_k is calculated via Eq. 20:

$$p(y_k|X) = \frac{p(y_k)}{p(X)}\prod_{i=1}^{n} p(x_i|y_k) \tag{20}$$

In this study, we selected the Gaussian Na ïve Bayes (GNB) model to deal with the continuous data produced by the various feature-representation algorithms. The distribution of the data was assumed to be a Gaussian distribution, which follows Eq. 21:

$$p(x_i|y_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x_i-\mu_k)}{2\sigma_k^2}} \tag{21}$$

where μ_k is the mean of X and σ_k^2 is the corresponding variance.

* Gradient Boosting Machine (_GBM_)

GBM was first developed as a greedy optimization model [70] for both regression and classification tasks. Among the variants of GBM, gradient tree boosting is a frequently used model integrated with DTs. Given $X = (x_1, x_2, \dots, x_n)$, in which x_i is related to label y_i, gradient tree boosting builds an ensemble of trees sequentially by distilling the gradient-descent algorithm into the process of new tree construction. A new tree is constructed under the discrepancy between the target function $f(x)$ and current model, in which $f(x_i) = y_i$. The discrepancy between the target function $f(x)$ and the current model is also called the residual of GBM.

3 MATERIALS

3.1 Human-bacterium Interaction Resources

In this section, we first collected and reviewed 11 public databases, as summarized in **Table 3**: the Database of Interacting Proteins (DIP) [29], Reactome [30], the Agile Protein Interaction DataAnalyzer (APID) [71], IntAct [33], the Molecular Interaction Database (MINT) [72], the InnateDB [73], the pathogen-host interaction search tool (PHISTO) [27], the Pathosystems Resource Integration Center (PATRIC) [24], Mentha [34], the Host Pathogen Interaction Database (HPIDB) [74, 75], and the Biological General Repository for Interaction Datasets (BioGRID) [76].

[Table 3]

As humans are one of the primary host species among infectious diseases, the HPI resources are considered as the preliminary investigation subjects from all these databases. The column ‘HPI number’ indicates the corresponding recorded interaction number from the databases, which contain both inter-species interactions and intra-species interactions. These 11 databases were selected because their data sources mainly come from the literature, which have been subjected to expert manual verification, and public archival databases, which also contain high confidence of the presented data.

Taking database PATRIC [24] as an example, the data source was built upon several public archival databases, such as MINT [72], IntAct [33], BioGRID [76] and DIP [29]. The cross-archived databases have extended the availability of HPI resources; however, some duplicates inevitably occur during the combination of these 11 databases. Thus, we followed the traditional data collection and cleansing methods from the literature [7-9].

3.2 Data Curation

In this section, we briefly describe the major statistics for ‘golden dataset’ curation, which will be thoroughly surveyed in the following sections.

** Positive Interactions*

Six different types of bacteria were selected, and the related data were pre-processed from the available databases. We identified the bacteria by mapping the taxonomy IDs according to the NCBI Taxonomy database. In **Table 4**, the corresponding information, including taxonomy ID, organism name, total pair number from the database and the number after cleansing, are presented. These 11 databases were accessed and downloaded in September 2018.

[Table 4]

In **Table 4**, the statistics refer to the results of the representative proteins. Meanwhile, any proteins with fewer than 50 amino acids were removed since these proteins may be non-functional fragments. The protein sequence information was primarily from the SwissProt/UniProtKB database [77].

** Negative Interactions*

How to select feasible negative PPIs remains an active topic for the prediction of PPIs. Currently, there is not a standard protocol defining both the negative pairing strategy and the ratio to positive interactions. In most cases, building a negative interaction dataset by randomly selecting protein pairs from a set of unknown interacting relationships between protein pairs is utilized. This heuristic approach works well in practice as the interaction ratio (i.e. the number of positive interactions in a large, random set of protein pairs) is expected to be very low, which in the work of [7] was defined as 25, 50, and 100 times as many negative examples as positive examples. In the study by [9], the ratio was set to 1/100. The assumption in this approach is that the probability that the selected negatives contain true positives is negligible.

Thus, we followed the traditional approaches from the literature [7, 9, 10, 21]. A random pairing for a negative PPI was first undertaken between different proteins sets, which in this study was between the chosen bacterium pathogens (listed in **Table 4**) and *Homo sapiens* proteins (taxonomy ID: 9606). Then, we randomly selected a subset from this random pairing set to be the negative dataset. The negative interactions were selected with different ratios: 1:1, 1:25, 1:50 and 1:100.

** Protein Information*

When building machine-learning models to predict PPIs, HB-PPIs are needed to utilise the diverse protein information, which can be divided into three groups: structure-based, domain-based and sequence-based protein information.

Numerous studies have utilized and examined different information when predicting specific HP-PPIs [7, 78, 79]. Particularly, domain-domain and structure-structure interaction methods are the two main approaches used to complement existing high-confidence interactions [7, 79]. Also, structural similarity, which refers to a result of homology-based modelling, is an important alternative for detecting proteins with a homogeneous structure based on experimentally verified HP-PPIs [78].

Although structure-based and domain-based information have some benefits for exploring HPIs [80, 81], they can limit the scope of studied HP-PPIs to specific genres and species, such as HIV-1, HCV, Ebola viruses and so on [7, 10, 58, 82-84]. One dominant reason is the limited amount of available experimentally determined structures and domain information, particularly for bacteria. Imputation remains a core technology to compensate for the dearth of protein information and helps to address the challenge of interaction predictions [79]. Imputation for missing data also impacts the prediction performance since it brings putative information, which might not be accurate. Thus, utilizing structure-based and domain-based information limits the availability and scalability to a wide range of studies of HB-PPIs.

Alternatively, there has been a research trend of predicting PPIs from sequence-based protein information [35, 85]. Sequence-based protein information is one of the most abundant sources of protein data, which has stimulated ongoing research to improve the prediction performance of novel feature-representation and machine-learning models [8, 14, 21, 86, 87]. Sequence-based methods enable the models to be applied to large datasets and various species and genres.

** Independent Datasets*

To help our readers understand each dataset's information, in **Table 5**, all the protein numbers related to the different subsets were included. This information, which was related to the reviewed sequence information from the UniProtKB database [26], was last updated on 30th Oct, 2018. In all, we collected 18,181 *Homo sapiens* protein sequence information, and the corresponding protein numbers for each taxonomy ID are reported in **Table 5**.

[Table 5]

Evaluation of the models requires careful preparation of the independent datasets. Generally, cross-validation shows better performance than the independent-testing model for an unseen dataset. To give a general performance evaluation, we followed [8] when we built the independent datasets. The difference was that we further built five-fold independent datasets, which helped us to better measure the means and variations of the machine-learning models. The independent datasets were not used during the training, and various measurements were included to evaluate the performance of different models based on the independent datasets. Thus, we first randomly selected one-fifth of the PPIs from both positive and negative interactions to be the independent dataset. The remaining PPIs of positive and negative interactions were then combined as the training set. We assembled the negative interactions with a random sampling method, where random sampling of the negative interactions was conducted five times, which allowed us to evaluate the different models with statistic means and variations to reduce the bias caused by negative interactions. The involved protein numbers for *Homo sapiens* and the corresponding bacterium taxonomy IDs are reported in detail in the appendix. We have reported the number of utilized proteins for each species for different ratio settings. We anticipate that this experimental setting and details will help to provide more information to build novel machine learning methods in future work.

[Figure 2]

The framework of our evaluation study is presented in **Figure 2**. In **Figure 2**, a clear process procedure from databases to training and independent datasets, followed by the feature representation algorithms and machine-learning model evaluations, are mapped in a coherent line. The best model selection and prediction are given as the main outcome of this framework.

4 EVALUATION RESULTS

4.1 Evaluation Metrics

A set of six popular performance evaluation metrics, including precision (Pre), accuracy (Acc), sensitivity (Sn), specificity (Sp), F1-score and Matthew's correlation coefficient (MCC) were applied to evaluate the overall prediction performance of the models [46, 88-94]. The measurements are defined as follows:

$$
\begin{aligned}
&Pre = \frac{TP}{TP+FP} \\
&Acc = \frac{TP+TN}{TP+FP+TN+FN} \\
&Sn = \frac{TP}{TP+FN} \\
&Sp = \frac{TN}{TN+FP} \\
&F1 = \frac{2 * Pre}{Pre+Rec} \\
&MCC = \frac{(TP * TN) - (FN * FP)}{\sqrt{(TP+FN) * (TN+FP) * (TP+FP) * (TN+FN)}}
\end{aligned}
\tag{22}
$$

where *TP*, *FP*, *TN* and *FN* represent the numbers of true positives, false positives, true negatives and false negatives, respectively. Also, the receiver operating characteristic (ROC) curve and the area under the curve (AUC) were included to quantify the model performances.

4.2 Performance Evaluation Based on Different Class Ratios

One primary evaluation of this study was the ratio impact of different predictors, which was the ratio between positive and negative protein interactions. We herein present the F1 score and Acc value from our measurements for feature 'ACC' for the evaluation discussion. Since the curated

HP-PPI datasets involve different ratios between positive and negative interactions data, Acc is able to more precisely measure the performance of the model in a more accurate way for the ratio of 1:1. However, when the ratios become skewed, such as 1:25 to 1:100, F1 score will be a more suitable performance measurement. The mean value and deviation of each of the five independent tests were calculated in terms of different bacterial species and building ratio settings between the positive and negative pairs. In general, the ability to predict positive interactions as negative pairs decreases both the F1 and Acc results. Here, we found that the Acc was as high as 0.990099 when all the test data were predicted as negative interactions for a ratio of 1:100 between the positive and negative interactions. For ratios of 1:25, 1:50 and 1:100 between the positive and negative interactions, the datasets were considered as imbalanced datasets. Therefore, the F1 score was more suitable for measuring the performance of imbalanced datasets.

From **Figure 3**, it is easy to see that the F1 scores present a trend of getting worse as the dataset becomes larger and more complex, which means more protein nodes and edges are involved in the dataset. For example, when the positive to negative ratio was 1:1, a 1.0±0.0 F1 score was found for the RF algorithm and the taxonomy ID is “1491”. However, the F1 score became 0.96±0.0 with RF for ID “644”,0.817555±0.029558 with LR for ID “623”, 0.730386±0.005192 with RF for ID “177416”, 0.770171±0.007703 with RF for ID “1392” and 0.752226±0.006632 with RF for ID “632”.

[Figure 3]

[Figure 4]

In **Figure 4** and **Figure 5**, feature “PseAAC” from the PseAAC method and feature “BlockPSSM” from the evolutionary information method are also included for different ratios. The performance comparison between these two different sequence-based features also indicate the impact of the ratio upon the F1 and Acc results.

[Figure 5]

From **Figure 3**, we can see that all the predictors have worse performance for all datasets when the ratio increases from 1:1 to 1:25, 1:50 and 1:100, especially when the dataset is with more than one hundred thousand samples. For example, for taxonomy ID “632”, the F1 score was 0.752226±0.006632 for a 1:1 ratio, however, the F1 scores dropped to 0.312530±0.010944 for a 1:25 ratio, 0.243679±0.012883 for a ratio of 1:50 and 0.154535±0.012569 for the 1:100 ratio. These results were all achieved with the RF algorithm.

In **Figure 6** and **Figure 7**, the results of the existing available methods are included. **Figure 6** contains the Acc, F1 and MCC scores for IDs “1491”, “644” and “623”, and **Figure 7** contains the results for IDs “177419”, “1392” and “632”. Both **Figure 6** and **Figure 7** indicate the performance variation when the dataset changes from taxonomy ID “1491” to “644” and “623”, which becomes worse for taxonomy IDs “177419”, “1392” and “632”. Even though the existing methods in **Figure 6** and **Figure 7** have incorporated several novel sequential feature-representation algorithms, their performance has not improved.

[Figure 6]

[Figure 7]

4.3 Overall Performance

Figure 8 and **Figure 9** show the ROC curves for taxonomy IDs “177416” and “644”, respectively.

[Figure 8]

[Figure 9]

We have listed the six evaluated machine-learning models as two groups. One group contains tree-based models, which includes DT, RF and GBM. The other group consists of kernel-based models including SVM, LR and the Naïve Bayes model. The performances are presented as mean ROC curves from five-fold independent test results for different ratios.

Because there are 1207 positive interaction pairs for taxonomy ID “177416”, the dataset size is 121907 for a ratio of 1:100, which is larger than that of taxonomy ID “644”. Somehow, the predictors performance became worse for the larger dataset. Although the tree-based models still outperformed the kernel-based models for each dataset, the overall performance was not stable across the different host-bacterium systems.

In **Table 6**, the best results of all the predictors are listed accordingly for taxonomy ID “632”. For example, for the AC feature representation algorithm dataset, the best results of for ratios of 1:1, 1:25, 1:50 and 1:100 were all achieved by RF model with accuracies of 0.757082±0.008000, 0.967350±0.000365, 0.982521±0.000128, and 0.990674±0.000043, respectively. The tree-based models, including DT, RF, and GBM, have demonstrated a strong generalization ability in terms of providing effective and efficient performance. The other models, such as kernel-based model, including SVM, Gaussian Naïve Bayes (GNB) model and the LR model, however, are less robust compared with the tree-based models. Meanwhile, the training time was in higher demand than for the tree-based models. Taking CTM as the feature representation algorithm, the time spent training GBM for the dataset of ratio 1:100 on taxonomy ID “632” was over 1,500 seconds. However, the time spent training the SVM model was more than 23,000 seconds.

[Table 6]

4.4 Further Discussion

Given different PPI networks, such as the HB-PPI between *Homo sapiens* and *Clostridium botulinum* (ID: 1491), and the interaction between *Homo sapiens* and *Yersinia pseudotuberculosis* *subsp. pestis* (ID: 632), the positive interactions networks have presented different complexities. As we can see, it still requires huge amounts of work towards the completeness of human-bacterium protein-protein interactions network. They have indicated different pathways between the different species. **Figures 10** and **11** show diagrams of two different interaction networks for taxonomy IDs 1491 and 632, respectively.

[Figure 10]

[Figure 11]

To accomplish a robust performance of predicting HB-PPIs, the relationship between positive and negative protein interactions requires further consideration. There have been several methods dedicated to one-class classification tasks, such as semi-supervised learning [95-97], to leverage the power of singularly labelled data and unlabelled data. This may help to improve the performance of protein interaction prediction regardless of the ratio between the positive and negative protein interactions. Meanwhile, since sequential feature-representation algorithms have been an active and challenging area, a better feature-representation algorithm is needed to help build a sequence-based end-to-end machine-learning model [98-100] for predicting HB-PPIs. Regarding the potential information of host-pathogen protein-protein interactions networks, cutting-edge machine learning algorithms are expected to more effectively decipher the code of protein information, in particular deep learning algorithms such as graph neural networks [101], long short-term memory and convolutional neural network model [102]. How to efficiently distil the useful information and features from the HP-PPIs networks by leveraging these advanced deep learning techniques to further enhance the predictive performance remains a challenge. By benefitting from the advanced machine learning models, the study on protein-protein interactions networks will eventually shed the lights on our understanding of infectious-disease mechanisms. Since publishing accessible portals for computational analysis and prediction has become a practice, it is essential to construct webservers [103-108] to support and publish standard alone tools to enhance the research communication and facilitate future discoveries of HP-PPIs. Given the increasing number of developed models for high-throughput prediction of HP-PPIs, the future work will involve the development of user-friendly tools and web servers for HP-PPIs evaluation and prediction.

5 CONCLUSIONS

In this study, we evaluated HB-PPIs in a systematic manner, where the focus was on leveraging machine-learning-based models as the primary computational method. We first presented a wide and deep review on currently available data sources and tools. As noted in the literature review (Section 2) of computational tools developed for prediction tasks of HP-PPIs, a careful data curation phase was implemented and a pipeline for HB-PPI studies was summarized, which included numerous sequential feature-representation algorithms and machine-learning models. Several other computational methods concerning HB-PPIs were also evaluated.

Given the study of HP-PPIs, we have tried to determine the impacts caused by different ratios of benchmark datasets, different feature-representation algorithms and different machine-learning models. The experimental results indicated that to better utilise machine-learning models and harness the power of accumulated protein interaction data, a more robust and more powerful computational model is required to achieve better performance across different HB-PPI prediction tasks. To facilitate the usage and study of HB-PPIs, a complete evaluation report and databases analysis have released along with this review to the wider biomedical research community.

Key Points

- A comprehensive review on currently available data sources and computational tools is presented.
- A comprehensive framework for HBI studies was summarized whilst both the datasets and computational methods were substantially reviewed and collected.
- A systematic evaluation of machine-learning-based computational prediction was delivered. Although numerous existing studies have reported the performance of traditional machine-learning methods separately, in this study, we evaluated a larger scope of machine-learning models as well as feature-representation algorithms. The evaluation was conducted by reporting multiple metrics and comparing between the different models.
- By composing the comprehensive pipeline for HBI studies, we have tried to answer the following questions: (a) How do machine-learning-based models perform on the prediction task of HB-PPIs? (b) How do feature-representation algorithms based on sequence information affect the model performance? (c) Do the ratios between positive and negative interactions have an impact on the model performance?

References

1. Prashanthi K, Chandra N. Host-Pathogen Interactions. In: Dubitzky W., Wolkenhauer O., Cho K.-H. et al. eds). Encyclopedia of Systems Biology. New York, NY: Springer New York, 2013, 904-908.
2. Mock M, Fouet A. Anthrax, Annual Reviews in Microbiology 2001;55:647-671.
3. Maresso AW, Garufi G, Schneewind O. Bacillus anthracis secretes proteins that mediate heme acquisition from hemoglobin, PLoS Pathogens 2008;4:e1000132.
4. Dyer MD, Nef C, Dufford M et al. The Human-Bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis, PLoS ONE 2010;5:e12089.
5. Nourani E, Khunjush F, Durmuş S. Computational approaches for prediction of pathogen-host protein-protein interactions, Frontiers in Microbiology 2015;6:1-10.
6. Durmus S, Çakir T, Özgür A et al. A review on computational systems biology of pathogen-host interactions, Frontiers in Microbiology 2015;6:1-19.
7. Dyer MD, Murali TM, Sobral BW. Supervised learning and prediction of physical interactions between human and HIV proteins, Infection, Genetics and Evolution 2011;11:917-923.
8. Cui G, Fang C, Han K. Prediction of protein-protein interactions between viruses and human by an SVM model, BMC Bioinformatics 2012;13:S5-S5.
9. Kshirsagar M, Carbonell J, Klein-Seetharaman J. Multitask learning for host-pathogen protein interactions, Bioinformatics 2013;29:217-226.
10. Emamjomeh A, Goliaei B, Zahiri J et al. Predicting protein–protein interactions between human and hepatitis C virus via an ensemble learning method, Molecular Biosystems 2014;10:3147-3154.
11. Eid FE, Elhefnawi M, Heath LS. DeNovo: Virus-host sequence-based protein-protein interaction prediction, Bioinformatics 2016;32:1144-1150.
12. Sen R, Nayak L, De RK. A review on host–pathogen interactions: classification and prediction, European Journal of Clinical Microbiology and Infectious Diseases 2016;35:1581-1599.
13. Zhou H, Jin J, Wong L. Progress in Computational Studies of Host–Pathogen Interactions, Journal of Bioinformatics and Computational Biology 2013;11:1230001-1230001 -- 1230001-1230026.
14. Zhang J, Kurgan L. Review and comparative assessment of sequence-based predictors of protein-binding residues, Briefings In Bioinformatics 2017:1-17.
15. Krishnadev O, Srinivasan N. A data integration approach to predict host-pathogen protein-protein interactions: application to recognize protein interactions between human and a malarial parasite, In Silico Biol 2008;8:235-250.
16. Huo T, Liu W, Guo Y et al. Prediction of host-pathogen protein interactions between Mycobacterium tuberculosis and Homo sapiens using sequence motifs, BMC Bioinformatics 2015;16:1-9.
17. Hwang H, Dey F, Petrey D et al. Structure-based prediction of ligand–protein interactions on a genome-wide scale, Proceedings of the National Academy of Sciences 2017;114:13685-13690.
18. Dyer MD, Murali TM, Sobral BW. Computational prediction of host-pathogen protein-protein interactions, Bioinformatics 2007;23:i159-i166.
19. Wuchty S. Computational prediction of Host-Parasite protein interactions between P. falciparum and H. sapiens, PLoS ONE 2011;6:e26960: 26961-26968.
20. Mei S. Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins, PLoS ONE 2013;8:1-13.
21. Ahmed I, Witbooi P, Christoffels A. Prediction of human-Bacillus anthracis protein–protein interactions using multi-layer neural network, Bioinformatics 2018;34:4159-4164.
22. Driscoll T, Dyer MD, Murali TM et al. PIG - The pathogen interaction gateway, Nucleic Acids Research 2009;37:647-650.
23. Squires B, Macken C, Garcia-Sastre A et al. BioHealthBase: Informatics support in the elucidation of influenza virus host-pathogen interactions and virulence, Nucleic Acids Research 2008;36:497-503.

24. Wattam AR, Abraham D, Dalay O et al. PATRIC, the bacterial bioinformatics database and analysis resource, Nucleic Acids Research 2014;42:581-591.

25. Braxton SM, Onstad DW, Dockter DE et al. Description and analysis of two internet-based databases of insect pathogens: EDWIP and VIDIL, Journal of Invertebrate Pathology 2003;83:185-195.

26. Consortium U. UniProt: the universal protein knowledgebase, Nucleic Acids Research 2017;45:D158-D169.

27. Durmuş Tekir S, Çakir T, Ardiç E et al. PHISTO: Pathogen-host interaction search tool, Bioinformatics 2013;29:1357-1358.

28. Chautard E, Dana JM, Rivas JDL et al. PSICQUIC and PSISCORE: accessing and scoring molecular interactions, Nature Methods 2012;8:528-529.

29. Xenarios I, Salwínski L, Duan XJ et al. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions, Nucleic Acids Research 2002;30:303-305.

30. Joshi-Tope G, Gillespie M, Vastrik I et al. Reactome: A knowledgebase of biological pathways, Nucleic Acids Research 2005;33:428-432.

31. Xiang Z, Tian Y, He Y. PHIDIAS: A pathogen-host interaction data integration and analysis system, Genome Biology 2007;8:R150.

32. Yue J, Zhang D, Ban R et al. PCPPI: A comprehensive database for the prediction of Penicillium-crop protein-protein interactions, Database 2017;2017:1-9.

33. Kerrien S, Aranda B, Breuza L et al. The IntAct molecular interaction database in 2012, Nucleic Acids Research 2012;40:841-846.

34. Calderone A, Castagnoli L, Cesareni G. Mentha: A resource for browsing integrated protein-interaction networks, Nature Methods 2013;10:690-691.

35. Shen J, Zhang J, Luo X et al. Predicting protein-protein interactions based only on sequences information, Proceedings of the National Academy of Sciences 2007;104:4337-4341.

36. Guo Y, Yu L, Wen Z et al. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences, Nucleic Acids Research 2008;36:3025-3030.

37. Davies MN, Secker A, Freitas AA et al. Optimizing amino acid groupings for GPCR classification, Bioinformatics 2008;24:1980-1986.

38. Chou K-C. Prediction of Protein Cellular Attributes Using Pseudo-Amino Acid Composition, Proteins: Struct., Funct., Genet. 2001;43:246-255.

39. Shen HB, Chou KC. PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition, Analytical Biochemistry 2008;373:386-388.

40. Chou K-C. Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology, Current Proteomics 2009;6:262-274.

41. Altschul SF, Madden TL, Schäffer AA et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Research 1997;25:3389-3402.

42. Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins, BMC Bioinformatics 2005;6:1-6.

43. Xia J-F, Han K, Huang D-S. Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor, Protein and Peptide Letters 2010;17:137-145.

44. Chou KC, Shen HB. MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM, Biochemical and Biophysical Research Communications 2007;360:339-345.

45. Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition, Journal of Theoretical Biology 2011;273:236-247.

46. Chen Z, Zhao P, Li F et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data, Briefings In Bioinformatics 2019;10.
47. Chen Z, Zhao P, Li F et al. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 2018;1:1-4.
48. Zahiri J, Yaghoubi O, Mohammad-Noori M et al. PPIevo: Protein-protein interaction prediction from PSSM based evolutionary information, Genomics 2013;102:237-242.
49. Wang J, Yang B, Song J. Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors, Bioinformatics 2018;34:2546-2555.
50. Uddin MR, Sharma A, Farid DM et al. EvoStruct-Sub: An accurate Gram-positive protein subcellular localization predictor using evolutionary and structural features, Journal of Theoretical Biology 2018;443:138-146.
51. Göktepe YE, Kodaz H. Prediction of Protein-Protein Interactions Using An Effective Sequence Based Combined Method, Neurocomputing 2018;303:68-74.
52. Zhang B, Li J, Lü Q. Prediction of 8-state protein secondary structures by a novel deep learning architecture, BMC Bioinformatics 2018;19:1-13.
53. Wang Y-B, You Z-H, Li L-P et al. Improving Prediction of Self-interacting Proteins Using Stacked Sparse Auto-Encoder with PSSM profiles, International Journal of Biological Sciences 2018;14:983-991.
54. Dayhoff MO. A model of evolutionary change in proteins, Atlas of protein sequence and structure 1972;5:89-99.
55. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks, Proceedings of the National Academy of Sciences 1992;89:10915-10919.
56. Jeong JC, Lin X, Chen X-w. On Position-Specific Scoring Matrix for Protein Function Prediction, IEEE/ACM Transactions on Computational Biology and Bioinformatics 2011;8:308-315.
57. Liu T, Zheng X, Wang J. Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile, Biochimie 2010;92:1330-1334.
58. Halder AK, Dutta P, Kundu M et al. Review of computational methods for virus–host protein interaction prediction: a case study on novel Ebola–human interactions, Briefings in Functional Genomics 2018;17:381-391.
59. Arnold R, Boonen K, Sun MGF et al. Computational analysis of interactomes: Current and future perspectives for bioinformatics approaches to model the host-pathogen interaction space, Methods 2012;57:508-518.
60. Cortes C, Vapnik V. Support-Vector Networks, Machine Learning 1995;20:273-297.
61. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology, IEEE transactions on systems, man, and cybernetics 1991;21:660-674.
62. Breiman L. Random forests, Machine Learning 2001;45:5-32.
63. Li F, Li C, Revote J et al. GlycoMine struct: a new bioinformatics tool for highly accurate mapping of the human N-linked and O-linked glycoproteomes by incorporating structural features, Scientific Reports 2016;6:1-16.
64. Li F, Li C, Wang M et al. GlycoMine: a machine learning-based approach for predicting N-, C-and O-linked glycosylation in the human proteome, Bioinformatics 2015;31:1411-1419.
65. Pedregosa F, Varoquaux G, Gramfort A et al. Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 2011;12:2825-2830.
66. Song J, Li F, Leier A et al. PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy, Bioinformatics 2018;34:684-687.
67. Li F, Li C, Marquez-Lago TT et al. Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome, Bioinformatics 2018;34:4223-4231.

68. Lewis DD. Naive (Bayes) at forty: The independence assumption in information retrieval. In: European conference on machine learning. 1998, p. 4-15. Springer.
69. Zhang H. The optimality of naive Bayes. In: The 17th International FLAIRS Conference. Miami Beach, Florida, USA, 2004, p. 562-567.
70. Friedman JH. Greedy Function Approximation : A Gradient Boosting Machine, The Annals of Statistics 2001;29:1189-1232.
71. Prieto C, De Las Rivas J. APID: Agile protein interaction DataAnalyzer, Nucleic Acids Research 2006;34:298-302.
72. Licata L, Briganti L, Peluso D et al. MINT, the molecular interaction database: 2012 Update, Nucleic Acids Research 2012;40:857-861.
73. Breuer K, Foroushani AK, Laird MR et al. InnateDB: Systems biology of innate immunity and beyond - Recent updates and continuing curation, Nucleic Acids Research 2013;41:1228-1233.
74. Kumar R, Nanduri B. HPIDB - a unified resource for host-pathogen interactions, BMC Bioinformatics 2010;11:S16.
75. Ammari MG, Gresham CR, McCarthy FM et al. HPIDB 2.0: a curated database for host-pathogen interactions, Database : the journal of biological databases and curation 2016;2016:1-9.
76. Chatr-Aryamontri A, Oughtred R, Boucher L et al. The BioGRID interaction database: 2017 update, Nucleic Acids Research 2017;45:D369-D379.
77. Boutet E, Lieberherr D, Tognolli M et al. Uniprotkb/swiss-prot. Plant bioinformatics. Springer, 2007, 89-112.
78. Davis FP, Barkan DT, Eswar N et al. Host pathogen protein interactions predicted by comparative modeling, Protein science : a publication of the Protein Society 2007;16:2585-2596.
79. Mariano R, Wuchty S. Structure-based prediction of host–pathogen protein interactions, Current Opinion in Structural Biology 2017;44:119-124.
80. Franzosa EA, Xia Y. Structural principles within the human-virus protein-protein interaction network, Proceedings of the National Academy of Sciences 2011;108:10538-10543.
81. Franzosa EA, Garamszegi S, Xia Y. Toward a three-dimensional view of protein networks between species, Frontiers in Microbiology 2012;3:1-6.
82. Qi Y, Tastan O, Carbonell JG et al. Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins, Bioinformatics 2010;26:i645-i652.
83. Tastan O, Qi Y, Carbonell JG et al. Prediction of interactions between HIV-1 and human proteins by information integration. Biocomputing 2009. World Scientific, 2009, 516-527.
84. Tyagi N, Krishnadev O, Srinivasan N. Prediction of protein–protein interactions between Helicobacter pylori and a human host, Molecular Biosystems 2009;5:1630-1635.
85. Gomez SM, Noble WS, Rzhetsky A. Learning to predict protein-protein interactions from protein sequences, Bioinformatics 2003;19:1875-1881.
86. Zhang L. Sequence-Based Prediction of Protein-Protein Interactions Using Random Tree and Genetic Algorithm, Intelligent Computing Technology 2012:334-341.
87. Yang S, Li H, He H et al. Critical assessment and performance improvement of plant–pathogen protein–protein interaction prediction methods, Briefings In Bioinformatics 2017:1-11.
88. Mei S, Li F, Leier A et al. A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction, Briefings In Bioinformatics 2019:bbz051:051-017.
89. Li F, Zhang Y, Purcell AW et al. Positive-unlabelled learning of glycosylation sites in the human proteome, BMC Bioinformatics 2019:1-17.
90. Zhang M, Li F, Marquez-Lago TT et al. MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters, Bioinformatics 2019;35:2957-2965.
91. Chen Z, Liu X, Li F et al. Large-scale comparative assessment of computational predictors for lysine post-translational modification sites, Briefings In Bioinformatics 2019;20:2267-2290.

92. Li F, Wang Y, Li C et al. Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: a comprehensive revisit and benchmarking of existing methods, Briefings In Bioinformatics 2019;20:2150-2166.
93. Song J, Wang Y, Li F et al. iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites, Briefings In Bioinformatics 2019;20:638-658.
94. Li F, Chen J, Leier A et al. DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites, Bioinformatics 2020;36:1057-1065.
95. Manevitz LM, Yousef M. One-class SVMs for document classification, Journal of Machine Learning Research 2001;2:139-154.
96. Chidlovskii B, Hovelynck M. Multi-modality classification for one-class classification in social networks. Google Patents, 2013.
97. Ruff L, Vandermeulen R, Goernitz N et al. Deep one-class classification. In: International Conference on Machine Learning. 2018, p. 4393-4402.
98. Min S, Lee B, Yoon S. Deep learning in bioinformatics, Briefings In Bioinformatics 2017;18:851-869.
99. Perera P, Patel VM. Learning deep features for one-class classification, IEEE Transactions on Image Processing 2019;28:5450-5463.
100. Greener JG, Kandathil SM, Jones DT. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints, Nature Communications 2019;10:1-13.
101. Zhang C, Song D, Huang C et al. Heterogeneous graph neural network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019, p. 793-803.
102. Hanson J, Paliwal K, Litfin T et al. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks, Bioinformatics 2019;35:2403-2410.
103. Li F, Fan C, Marquez-Lago TT et al. PRISMOID: a comprehensive 3D structure database for post-translational modifications and mutations with functional im-pact, Briefings In Bioinformatics 2019:bbz050.
104. Hong J, Luo Y, Mou M et al. Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery, Briefings In Bioinformatics 2019:bbz120.
105. Hong J, Luo Y, Zhang Y et al. Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning, Briefings In Bioinformatics 2019:bbz081.
106. Tang J, Wang Y, Fu J et al. A critical assessment of the feature selection methods used for biomarker discovery in current metaproteomics studies, Briefings In Bioinformatics:bbz061.
107. Lian X, Yang S, Li H et al. Machine-Learning-Based Predictor of Human-Bacteria Protein-Protein Interactions by Incorporating Comprehensive Host-Network Properties, Journal of Proteome Research 2019;18:2195-2205.
108. Yang S, Fu C, Lian X et al. Understanding Human-Virus Protein-Protein Interactions Using a Human Protein Complex-Based Analysis Framework, MSystems 2019;4:e00303-00318.