Journal of proteome research

Article

Cite This: J. Proteome Res. 2019, 18, 2195−2205

pubs.acs.org/jpr

Machine-Learning-Based Predictor of Human−Bacteria Protein−Protein Interactions by Incorporating Comprehensive Host-Network Properties

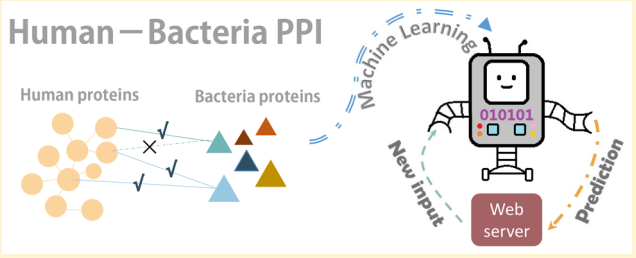
Xianyi Lian,[†,#] Shiping Yang,[†,#] Hong Li,[‡] Chen Fu,[†] and Ziding Zhang*[,†]

[†]State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China

[‡]Key Laboratory of Tropical Biological Resources of Ministry of Education, Hainan University, Haikou, 570228, China

Supporting Information

ABSTRACT: The large-scale identification of protein−protein interactions (PPIs) between humans and bacteria remains a crucial step in systematically understanding the underlying molecular mechanisms of bacterial infection. Computational prediction approaches are playing an increasingly important role in accelerating the identification of PPIs. Here, we developed a new machine-learning-based predictor of human−*Yersinia pestis* PPIs. First, three conventional sequence-based encoding schemes and two host network-property-related encoding schemes (i.e., NetTP and NetSS) were introduced. Motivated by previous human−pathogen PPI network analyses, we designed NetTP to systematically characterize the host proteins' network topology properties and designed NetSS to reflect the molecular mimicry strategy used by pathogen proteins. Subsequently, individual predictive models for each encoding scheme were inferred by Random Forest. Finally, through the noisy-OR algorithm, 5 individual models were integrated into a final powerful model with an AUC value of 0.922 in the 5-fold cross-validation. Stringent benchmark experiments further revealed that our model could achieve a better performance than two state-of-the-art human−bacteria PPI predictors. In addition to the selection of a suitable computational framework, the success of our proposed approach could be largely attributed to the introduction of two comprehensive host network-property-related feature sets. To facilitate the community, a web server implementing our proposed method has been made freely accessible at http://systbio.cau.edu.cn/intersppiv2/ or http://zzdlab.com/intersppiv2/.



KEYWORDS: *human−bacteria interaction, protein−protein interactions, machine learning, network properties, noisy-OR algorithm*

INTRODUCTION

Currently, infectious diseases (e.g., bacterial diseases) remain a major threat to human life and health, sickening and killing millions of people every year. On the one hand, the coinfection of bacterial and viral diseases has become a new trend.[1] On the other hand, the overuse or abuse of antibiotics has resulted in the rapid emergence of drug-resistant bacteria worldwide and has endangered the effectiveness of antibiotics.[2] These two factors make the treatment of bacterial diseases a long-term challenge, forcing us to accelerate the study of the molecular mechanisms of bacterial infection.

As the most important type of host−pathogen interaction, protein−protein interactions (PPIs) between host and pathogen play an important role in infection and disease progression.[3,4] Owing to the progress in high-throughput techniques, the identification of PPIs in individual organisms (intraspecies PPIs) that are verified by large-scale experiments has rapidly increased. In contrast, the identification of PPIs between different organisms (interspecies PPIs), such as host−pathogen PPIs (HP-PPIs), is only now emerging, and experimental data are generally limited. Moreover, experimental methods are often time-consuming and laborious, making it unfeasible to detect all possible HP-PPIs. Therefore, there is an urgent need to develop efficient and reliable computational prediction approaches to identify interaction candidates or to prioritize targets for high-throughput HP-PPI screening.

Traditional intraspecies PPI prediction approaches often use known PPIs, domain−domain interactions (DDIs) and domain−motif interactions (DMIs) as templates to infer the potential interaction relationships between query protein pairs, and these methods are generally referred to as interolog mapping,[5,6] the DDI-based method[7,8] and the DMI-based method,[9] respectively. In the meantime, machine learning (ML)-based prediction methods have also been booming in recent decades. ML-based methods typically convert the PPI prediction task into a binary classification framework. To train an ML predictive model, the encoding schemes converting protein pairs into feature vectors are required. Currently, a variety of protein-encoding schemes have been developed,

Received: January 29, 2019
Published: April 15, 2019

DOI: 10.1021/acs.jproteome.9b00074
J. Proteome Res. 2019, 18, 2195−2205

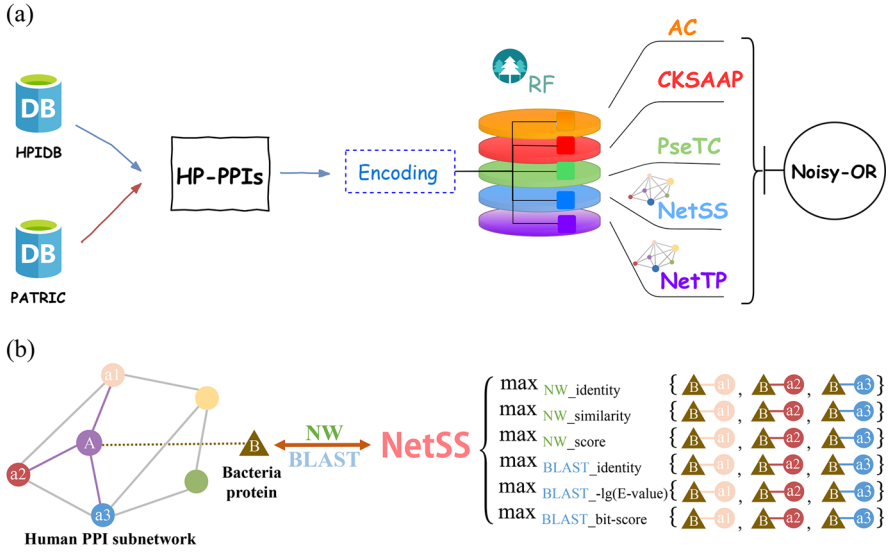


Figure 1. Flowchart and highlights of the proposed human–bacteria PPI predictor. (a) Computational framework of the prediction model. (b) A schematic example of the NetSS encoding scheme. Suppose that bacterial protein B interacts with human host protein A, which contains three partners: proteins a1, a2, and a3. We used NW and BLAST algorithms to obtain six similarity measures [NW_identity, NW_similarity, NW_score, BLAST_identity, BLAST_-lg (E-value), and BLAST_bit-score] between protein B with proteins a1, a2, and a3, individually. The corresponding maximum values were taken as the features of NetSS.

including sequence,[10−15] structure,[16] physiochemical properties,[10,17] and evolutionary profiles.[18,19]

Compared to intraspecies PPI prediction, the computational prediction of interspecies PPIs is a relatively young research topic. Traditional intraspecies PPI inference methods, such as interolog mapping,[20] the DDI-based method[20,21] and the DMI-based method,[22] have been directly adapted to predict interspecies PPIs, although their performance has not been intensively benchmarked. With the accumulation of experimentally verified HP-PPI data, ML-based HP-PPI prediction methods have also been proposed.[23] Most of these ML-based methods simply apply known intraspecies PPI prediction schemes to HP-PPI prediction without fully considering the biological characteristics of the interspecies PPIs. Overall, predicting interspecies PPIs is more challenging than intraspecies PPIs due to the finiteness of data resources and the complicated regulatory mechanisms of interspecies PPIs.

Regarding the host–pathogen system, most studies have focused on PPI prediction between human and pathogens. Indeed, a number of human–virus PPI prediction methods have been developed.[22,24−29] Some PPI prediction approaches related to human and bacteria have also been proposed,[21,30−35] but only a handful of them employed ML to build predictive models. Kshirsagar et al. proposed a multitask learning-based method for predicting human–bacteria PPIs, which was based on the biological hypothesis that proteins from different pathogens essentially target the same critical biological processes in human cells.[32] However, when put into practice, this method still has limitations in the acquisition of high-dimensional features such as the Gene Ontology (GO) annotations for query protein pairs. Very recently, Ahmed et al. used a multilayer neural network to predict human–*Bacillus anthracis* (human–*B. anthracis*) PPIs mainly by using a series of sequence features, including triplets and quadruplets of consecutive amino acids.[30] Although a promising performance has been reported,[30,32] the existing methods are not satisfying for practical use due to the lack of easy-to-use executive codes and web servers, suggesting that there is sufficient room for method improvement.

In the context of HP-PPI networks, some interesting network patterns have also been observed. According to several recent studies on the human–*Yersinia pestis* (human–*Y. pestis*) PPI network,[36,37] bacterial effector proteins have evolved to preferentially interact with important host proteins, which tend to be hubs (proteins with many interacting partners) and bottlenecks (proteins that lie in the shortest paths between many pairs of proteins) in the human PPI network. Moreover, it is well-established that the molecular mimicry of host proteins is a widely adopted strategy for pathogenic bacteria to exploit and subvert host processes during infection.[38−40] In brief, the bacterial effector proteins mimic the host-targeting proteins' partners and compete with them for the binding interface at the host proteins, making the host proteins unable to bind to their partners, thereby disrupting the normal host pathway. These previous network analyses not only allowed us to obtain a global landscape of host–pathogen PPIs but also provided important hints for the development of new interspecies PPI predictors.

In this work, we aimed to develop a new ML-based predictor of PPIs between human and *Y. pestis*. As a rod-shaped Gram-negative bacteria and plague pathogen, *Y. pestis* is classified as a potential agent of bioterrorism.[41] Historically, it has caused three massive pandemics that have killed tens of millions of people.[42] High-throughput experiments have been used to detect human–*Y. pestis* PPIs, providing sufficient data and an excellent opportunity for the development of ML-based prediction methods. To this end, efforts were made in two aspects. On the one hand, we dedicated ourselves to seeking new encoding schemes, which is an effective and important strategy to improve the performance of ML-based predictors. In addition to protein sequence information, we focused on the maximal utilization of host network property based features. On the other hand, we attempted to optimize an ML-based computational framework. We assessed different ML methods on the encoding schemes and selected the most suitable one,

DOI: 10.1021/acs.jproteome.9b00074
J. Proteome Res. 2019, 18, 2195−2205

as well as the corresponding integration strategy, to build the final predictive model.

MATERIALS AND METHODS

Data Sets

Regarding the ML-based HP-PPI prediction, experimentally verified HP-PPIs were treated as positive samples, while non-interacting protein pairs (non-PPIs) from the host and pathogen were treated as negative samples. The human−*Y. pestis* PPIs were downloaded from the Host−Pathogen Interaction Database (HPIDB)[43] and the Pathosystems Resource Integration Center (PATRIC).[44] The redundant PPIs were first removed, and then the PPIs containing proteins with fewer than 35 amino acids or with nonstandard amino acids were further filtered out; thus, 3892 PPIs between 1207 *Y. pestis* proteins and 2067 human proteins were retained and used as the positive samples. In total, 345 280 human PPIs were obtained from the BioGRID,[45] HPRD,[46] and I2D[47] databases to construct a human PPI network. We used a common method called random sampling to construct the negative samples, in which the human proteins were selected from our human PPI network, whereas the *Y. pestis* proteins were downloaded from UniProt.[48] Although highly imbalanced in the real world, the ratio of positive to negative samples was determined to be 1:1 to infer the ML-based models, which is commonly used in PPI prediction.[12,33] Finally, we obtained an initial data set containing 3892 PPIs and 3892 non-PPIs, which was further used to infer the predictive model.

Computational Framework of the Proposed Predictor

Based on the collected PPI data between human and *Y. pestis*, we designed a computational framework to develop a new HP-PPI predictor (Figure 1a). First, we introduced five different encoding schemes to construct feature vectors for protein pairs between human and *Y. pestis*. These five encoding schemes included three sequence-based encodings [i.e., auto covariance (AC), the composition of *k*-spaced amino acid pairs (CKSAAP), and pseudotripeptide composition (PseTC)], and two host network-property-related encodings [i.e., network topology parameters (NetTP) and sequence similarity measurements between the pathogen protein and the host protein's partners (NetSS)]. Subsequently, we built the individual predictive model of each encoding scheme by Random Forest (RF). Finally, we used the noisy-OR algorithm[49] to integrate the five individual models into a final predictive model. Additional details about the encoding schemes, the implementation of RF, and model integration are elaborated in the following sections.

Encoding Schemes

AC. AC accounts for the interactions between amino acid residues separated by a certain number of amino acids throughout the whole sequence.[10] Here, we employed seven physicochemical properties of amino acids (Table S1), including hydrophobicity, hydrophilicity, volumes of side chains, polarity, polarizability, solvent-accessible surface area, and net charge index of side chains, to infer the AC feature vector. To represent a protein sequence X with a sequence length of n, the AC variables are calculated according to eq 1:

$$\mathrm{AC}_{\mathrm{lag},j} = \frac{1}{n-\mathrm{lag}} \sum_{i=1}^{n-\mathrm{lag}} \left(P_{i,j} - \frac{1}{n} \sum_{i=1}^{n} P_{i,j} \right) \times \left(P_{(i+\mathrm{lag}),j} - \frac{1}{n} \sum_{i=1}^{n} P_{i,j} \right) \quad (1)$$

where lag is the sequence distance between residues and $P_{i,j}$ is the jth physicochemical property of the ith amino acid in the sequence X. A protein pair was characterized by concatenating the AC variables of two proteins. We set the maximal *lag* as 30; thus, a protein pair was converted into a 420-dimensional ($30 \times 7 \times 2$) vector.

CKSAAP. For a protein sequence, there are 20 common amino acids that make up a total of 400 amino acid pairs. These pairs can be extended to the *k*-spaced amino acid pairs (i.e., the pairs separated by *k* other amino acids).[50,51] Here, the CKSAAP encoding considered the *k*-spaced amino acid pairs, with $k = 0$, 1, 2, and 3. For instance, there is a protein sequence of "AAACC". When $k = 0$, the CKSAAP can be expressed as $(\mathrm{AA}, \mathrm{AC}, \mathrm{AD}, \ldots, \mathrm{CA}, \mathrm{CC}, \ldots, \mathrm{YY})_{400}$; this protein sequence contains 2 "AA" pairs, 1 "AC" pair and 1 "CC" pair, which can be expressed as a 400-dimensional vector $(2/4, 1/4, 0, \ldots, 0, 1/4, \ldots, 0)_{400}$. When $k = 1$, the CKSAAP can be expressed as $(\mathrm{A\_A}, \mathrm{A\_C}, \mathrm{A\_D}, \ldots, \mathrm{C\_A}, \mathrm{C\_C}, \ldots, \mathrm{Y\_Y})_{400}$; this protein sequence contains 1 "A_A" pair and 2 "A_C" pairs, which can be expressed as $(1/3, 2/3, 0, \ldots, 0, 0, \ldots, 0)_{400}$. Similarly, the CKSAAP for $k = 2$ and 3 can also be obtained. By simultaneously taking the above four vectors into account, the resulting CKSAAP of a protein sequence can be represented by a 400×4 dimensional feature vector. By further concatenating the vectors of two proteins, a protein pair was converted into a 3200-dimensional ($400 \times 4 \times 2$) vector.

PseTC. PseTC uses the tripeptide composition to represent a protein sequence. To avoid the dimensionality explosion problem, we divided the 20 amino acids into 13 groups (G, IV, FYW, A, L, M, E, QRK, P, ND, HS, T, and C)[14] and then calculated the group-based tripeptide composition (i.e., PseTC). Thus, a protein pair was converted into a 4394-dimensional ($13^3 \times 2$) vector.

NetTP. Considering that the host-targeting proteins contain specific network topology properties in the host PPI network, we used seven network topology parameters to infer the NetTP encoding, including degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, PageRank centrality, eccentricity, and clustering coefficient. The degree centrality of a node is the number of its neighboring nodes in the network. The betweenness centrality of a node is defined as the fraction of shortest paths between node pairs that pass through the node of interest. The closeness centrality of a node is defined by the inverse of the average length of the shortest paths to all the other nodes in the network. eigenvector centrality measures the centrality of a node by considering its neighboring nodes' centralities.[52] In particular, it can identify nodes with a low degree while bridging subnetworks with highly connected nodes. As a variant of eigenvector centrality, PageRank centrality was first proposed by Google to evaluate the importance of webpages.[53] It describes the importance of a node by considering both the number and importance of its parent nodes. The eccentricity of a node is calculated by measuring the shortest distance from the node to all nodes in the network and taking the maximum.

DOI: 10.1021/acs.jproteome.9b00074
J. Proteome Res. 2019, 18, 2195−2205

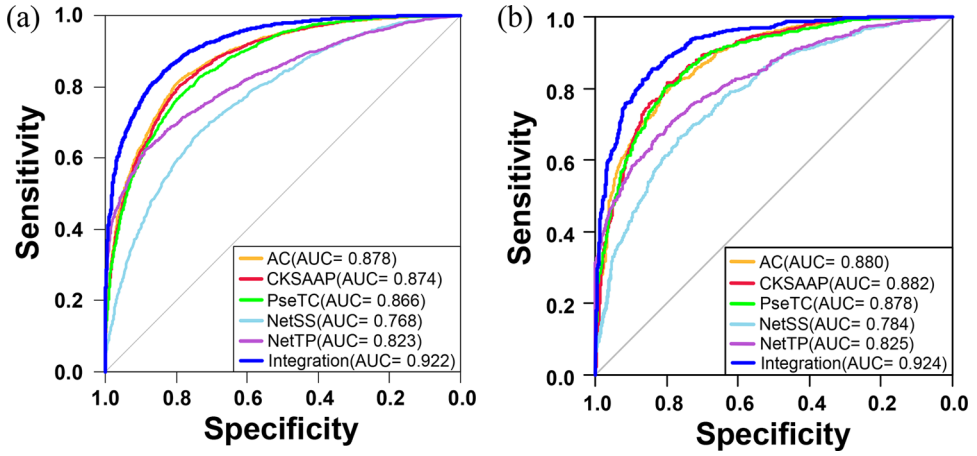


Figure 2. Performance of each individual model and the integrative model. (a) ROC curves of 5-fold cross-validation. (b) ROC curves of the independent test.

The clustering coefficient of a node is defined as the ratio of the edges present among its neighbors to all the possible edges that could be present among the neighbors. Here, we used the R package called igraph[54] to calculate these seven network topology parameters of proteins in our human PPI network. Note that this encoding scheme can only infer features for the human protein from the protein pair. Thus, a protein pair was converted into a seven-dimensional vector.

NetSS. Following the molecular mimicry strategy, the NetSS encoding scheme was designed to represent the sequence similarity between the pathogen protein and the host-targeting protein's partners. We used two sequence alignment algorithms [i.e., the Needleman-Wunsch (NW) algorithm[55] and BLAST][56] to infer the NetSS features. The identity, similarity, and alignment score between two protein sequences from the NW result, and the identity, E-value, and bit score from the BLAST result were used. When a host protein has more than one partner protein, the six sequence similarity measures between each partner and the pathogen protein were calculated, and the maximal value corresponding to each similarity measure was retained. A schematic example for the calculation of NetSS is shown in Figure 1b. Thus, a protein pair was converted into a six-dimensional vector.

Implementation of RF and Construction of the Integrative Model

RF is a flexible, popular, and easy-to-use ML method to build predictive models for both classification and regression problems.[57] Following the ensemble strategy, the RF model creates an entire forest of random uncorrelated decision trees to achieve the best possible result. In this work, we employed the Waikato Environment for Knowledge Analysis (WEKA)[58] to infer RF-based predictive models for different encoding schemes. For three sequence-based encoding schemes, the parameter called batch size was set as 2000. For the encoding schemes of NetTP and NetSS, the batch size was set as 100 and 300, respectively. Other parameters were set as defaults. For a query protein pair, five different prediction scores were obtained from the RF models corresponding to different encoding schemes. The noisy-OR algorithm[49] was further employed to integrate the five prediction scores into the final prediction score, which was carried out through the following three steps.

In the first step, for a protein pair (i,j), the interaction probability $I_{(i,j)}$ derived from the prediction score of each individual model was calculated, respectively, which was defined as:

$$P(I_{(i,j)}|s_{(i,j),k}) = \frac{|\text{positive training cases with scores} \geq s_{(i,j),k}|}{|\text{all training cases with scores} \geq s_{(i,j),k}|} \quad (2)$$

where $s_{(i,j),k}$ is the prediction score in the kth model (k is from 1 to 5). In the second step, the probabilities from all the prediction scores generated by individual models were integrated into a single probability using the noisy-OR model:

$$P_{(i,j),\text{noisy-OR}} = 1 - \prod_{k=1}^{n} (1 - P(I_{(i,j)}|s_{(i,j),k})) \quad (3)$$

where n is the total number of models. In the third step, a final probability of interaction was calculated as:

$$P(I_{(i,j)}|P_{(i,j),\text{noisy-OR}}) = \frac{|\text{positive training cases with } P_{\text{noisy-OR}} \geq P_{(i,j),\text{noisy-OR}}|}{|\text{all training cases with } P_{\text{noisy-OR}} \geq P_{(i,j),\text{noisy-OR}}|} \quad (4)$$

Performance Evaluation

To train and evaluate the predictive models, the initial human−*Y. pestis* data set was further partitioned into a training set (3135 PPIs and 3135 non-PPIs) and an independent test set (757 PPIs and 757 non-PPIs). To conduct a stringent performance assessment, both a 5-fold cross-validation and an independent test were carried out. We chose the receiver operating characteristic curve (ROC curve) and area under ROC curve (AUC) to assess the performance of our models. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. TPR is also known as sensitivity or recall, and FPR is equal to 1 − specificity. The formulas to calculate TPR and FPR are as follows:

$$\text{TPR} = \text{sensitivity} = \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} = 1 - \frac{\text{TN}}{\text{TN} + \text{FP}} = 1 - \text{specificity} \quad (6)$$

DOI: 10.1021/acs.jproteome.9b00074
J. Proteome Res. 2019, 18, 2195−2205

where the true positive (TP) is defined as the number of interacting samples classified correctly, the true negative (TN) stands for the number of non-interacting samples classified correctly, the false positive (FP) denotes the number of non-interacting samples classified mistakenly as interacting, and the false negative (FN) is defined as the number of interacting samples classified mistakenly as noninteracting. The larger the AUC, the better the predictive performance of the model.

A confusion matrix was also used to complement the performance assessment of the predictive models, which shows the TPR, false negative rate (FNR), FPR, and true negative rate (TNR) of a model at the maximum of MCC (Matthews correlation coefficient). The MCC is calculated according to the following formula:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \quad (7)$$

RESULTS AND DISCUSSION

Performance of Each Individual Model

For each encoding scheme, we used RF to construct the corresponding predictive model. The ROC curves of each model in the 5-fold cross-validation and independent test are illustrated in Figure 2. We found that the three sequence-based models achieved a relatively good and close performance. Regarding the 5-fold cross-validation, their AUC values ranged from 0.866 to 0.878. Moreover, the sequence-based models outperformed the other two network-based methods (NetTP and NetSS), whose AUC values were equal to 0.823 and 0.768, respectively. Notably, the ROC curves in the independent test set showed the same trend.

Performance of the Final Integrative Model

To improve the prediction performance, we used the noisy-OR algorithm to integrate these five individual models into a more powerful model. Generally, the final integrative model resulted in a significant performance improvement (Figure 2, DeLong's test, p value of $<2.2 \times 10^{-16}$). The AUC of the integrated model was 0.922, which was 0.044 higher than that of the best individual model (i.e., the AC encoding scheme) in the 5-fold cross-validation. The performance of the independent test also improved by 0.042 after integration compared to the best individual model (i.e., the CKSAAP encoding scheme). In practice, the performance at low-FPR control seems to be more important. Thus, we also calculated the corresponding AUC01 values (i.e., the partial AUCs, with the Specificity being 1 to 0.9). As listed in Table 1, the final integrative model also considerably outperformed any individual model. To further quantify the performance of the final integrative model, the confusion matrices of each individual model and the integrative model are also provided in Figure S1.

Considering that negative data are much more available than positive data, we also conducted computational experiments to test the influence of negative sampling. We randomly selected another nine groups of negative data sets and retrained the predictive models. The results showed that the model performance inferred from these 10 different groups of negative samples was quite stable. For instance, the average AUC value of the integrated model was 0.918 in the 5-fold cross-validation, which outperformed any individual model (Figure S2). Therefore, we concluded that the proposed prediction method is generally robust to the change of negative samples.

It is worth noting that existing HP-PPI prediction methods usually concatenate the heterogeneous features into a high-dimensional feature vector to facilitate the implementation of ML algorithms rather than take the computational framework used in this work. For comparison, we also tried to build the RF model based on the concatenation of the five encoding schemes into a high-dimensional vector. To avoid the potential curse of dimensionality, the feature selection approach called minimal-redundancy−maximal-relevance criterion (mRMR)[59] and the feature projection approach called principal component analysis (PCA)[60] were used to reduce the dimension. As shown in Table 2, the performance of the three models based on the concatenation of different encoding schemes was inferior to our integrative model, and these two feature selection methods did not result in performance improvement, which implied that our current computational framework was more suitable for predicting human−*Y. pestis* PPIs. We also used four other popular ML methods [Adaboost, Naive Bayes, Logistic Regression, and Support Vector Machine (SVM)] to verify our conclusions. Briefly, we used these four ML methods to obtain the predictive model of each encoding scheme, and then the integrative model corresponding to each ML method was obtained through the noisy-OR algorithm again. Note that Adaboost, Naive Bayes, and logistic regression were implemented through the WEKA[58] platform, while the training of the SVM model was carried out through the LIBSVM package.[61] Adaboost, short for adaptive boosting, is an ML meta-algorithm. In this work, the base classifier of Adaboost was set as RF, and the other parameters were set as defaults. The parameters in Naive Bayes

Table 1. AUC01 Values of Each Individual Model and the Final Integrative Model

method	5-fold cross-validation	independent test
AC	0.046	0.046
CKSAAP	0.044	0.043
PseTC	0.043	0.045
NetTP	0.048	0.046
NetSS	0.025	0.028
integrative model	0.060	0.059

Table 2. Performance (AUC) Comparison of Our Integrative Model and Other Models Based on the Concatenation of Different Encoding Schemes

method[a]	5-fold cross-validation	independent test
our model	0.922	0.924
concatenation[b]	0.887	0.896
concatenation plus PCA[c]	0.839	0.836
concatenation plus mRMR[d]	0.840	0.850

[a]We built the RF model based on the concatenation of the five encoding schemes into a high-dimensional vector. To avoid the potential curse of dimensionality, the feature selection approach called minimal-redundancy−maximal-relevance criterion (mRMR)[59] and the feature projection approach called principal component analysis (PCA)[60] were used to reduce the dimension. [b]Here, the dimensionality of concatenation was 8027. [c]After the feature selection of PCA, the dimension was reduced to 900. [d]After the feature selection of mRMR, the dimension of the retained features was 400.

DOI: 10.1021/acs.jproteome.9b00074
J. Proteome Res. 2019, 18, 2195−2205

and logistic regression were also set as defaults. For SVM, the parameters c and g of the five encoding schemes were optimized by grid search. The results showed that RF outperformed these 4 ML methods, either in the 5-fold cross-validation or in the independent test (Table 3), demonstrating that RF does exhibit a great tolerance for high-dimensional feature vectors and, therefore, is suitable for our classification task.

Table 3. Performance (AUC) Comparison of RF and the Other Four ML Methods in the Five-Fold Cross-Validation and Independent Test

ML method	5-fold cross-validation	independent test
Random Forest	0.922	0.924
Adaboost	0.889	0.895
SVM	0.893	0.871
naive bayes	0.734	0.740
logistic regression	0.753	0.696

Contribution of Network Property Based Encoding Schemes to the Integrative Model

To investigate the contributions of different encoding schemes, we also separately integrated three sequence-based encoding schemes and two network-based encoding schemes through the noisy-OR algorithm. The integration of the three sequence-based encoding schemes resulted in only minor improvements, with the AUC increasing from 0.878 (the best sequence encoding scheme, AC) to 0.889 (Table S2, DeLong's test, p value of 0.060), while the integrative model of the two network-based encoding schemes showed a more significant performance improvement, with the AUC increasing from 0.823 (NetTP) to 0.855 (DeLong's test, p value of 7.4×10^{-6}). Thus, these two network-based encoding schemes were complementary and they contributed considerably to the improved performance of the final integrative model, although their individual performances were weak.

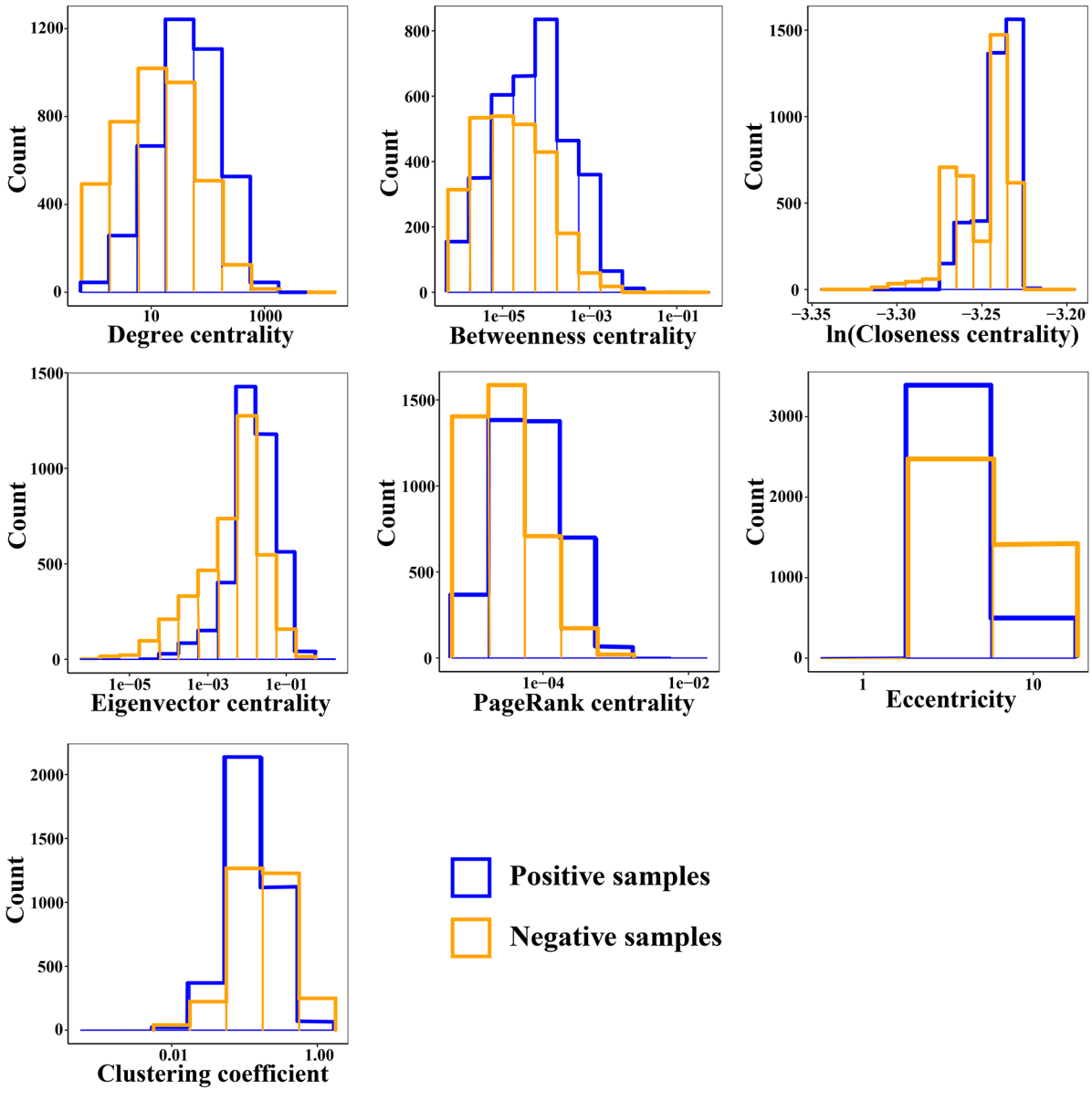


Figure 3. Distribution histograms of the NetTP features in positive and negative samples.

DOI: 10.1021/acs.jproteome.9b00074
J. Proteome Res. 2019, 18, 2195−2205

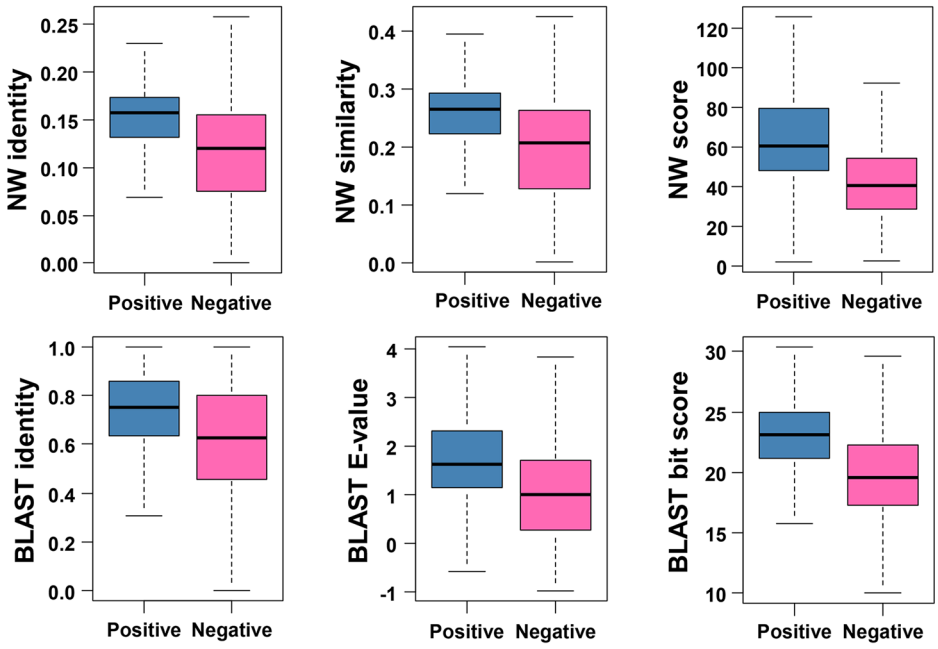


Figure 4. Box plots showing the distribution of the NetSS features in positive and negative samples. We took the "-lg" transformation of the original BLAST E-value.

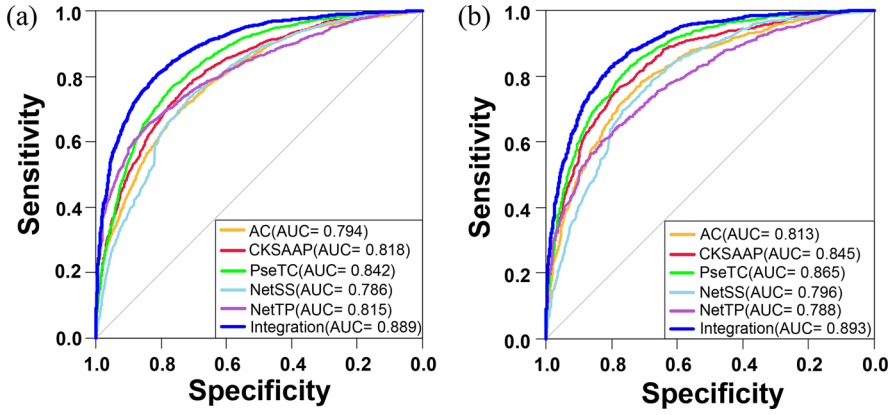


Figure 5. Performance of our model in predicting PPIs between human and two other bacterial species. (a) ROC curves in predicting human−*B. anthracis* PPIs. (b) ROC curves in predicting human−*F. tularensis* PPIs.

To quantitatively characterize the features of the two network-based encoding schemes and to understand the corresponding biological significance, we conducted statistical analyses of these features in positive samples and negative samples. Regarding NetTP, the distribution histograms of the seven network topology parameters were plotted to intuitively demonstrate their effectiveness in distinguishing positive samples from negative samples (Figure 3). In addition to the clustering coefficient (Wilcoxon test, two-tailed p value of 0.021), the distributions of the remaining six topology parameters were significantly different between positive samples and negative samples (Wilcoxon test, two-tailed p value of $<2.2 \times 10^{-16}$). For instance, we found that the average degree or betweenness centrality of human proteins in positive samples was significantly larger than that of human proteins in negative samples (Figure 3a,b). In line with previous analyses, these results showed that the effector proteins of *Y. pestis* were more likely to attack hubs and bottlenecks in the human PPI network to effectively infect the host. The host-targeting proteins had many interacting partners and lay in shortest paths between any two proteins, so they could control the flow of information between other nodes. Once they were attacked, the entire network would quickly collapse. It should be emphasized that PageRank and eigenvector centrality also have a high impact on the classification of PPIs and non-PPIs between human and *Y. pestis*. Indeed, PageRank has been proved to be more effective than the degree in the identification of crucial nodes in some biological networks.[62] To the best of our knowledge, these new network topology properties have not been used in the prediction of human−bacteria PPIs. In this work, these seven topology parameters were jointly used to comprise a comprehensive feature vector to maximally reflect the overall topology patterns of host-targeting proteins. Slightly different from NetTP, the distributions of the corresponding features in positive and negative samples are illustrated by box plots (Figure 4). Intuitively, the results showed that the six indicators were significantly higher for positive samples than for negative samples (Wilcoxon test, one-tailed p value of $<2.2 \times 10^{-16}$), indicating that the *Y. pestis* proteins tended to share sequence

DOI: 10.1021/acs.jproteome.9b00074
J. Proteome Res. 2019, 18, 2195−2205

similarity with their host-targeting proteins' partners. The results also provided indirect evidence to support that bacterial proteins follow the molecular mimicry strategy to interact with their host-targeting proteins. The NW algorithm identifies the global sequence similarity between two proteins, while BLAST reflects the local sequence similarity. In short, the NetSS encoding represented a comprehensive set of similarity indicators that we hoped could systematically characterize the sequence similarity between bacterial proteins and their host-targeting proteins' partners. Collectively, the aforementioned statistical analyses clearly demonstrated the effectiveness of these two comprehensive host network-property-related feature sets.

Ability to Predict PPIs between Human and Two Other Bacterial Species

To evaluate the extrapolation of our model, we also assessed its performance in predicting PPIs between human and other bacterial species. To this end, we downloaded 3039 human−*B. anthracis* PPIs and 1375 human−*Francisella tularensis* (human−*F. tularensis*) PPIs from HPIDB[43] and PATRIC.[44] The method of selecting the negative samples for human and the two bacterial species was the same as that used in the human−*Y. pestis* PPI data set. We found that our model showed a reasonably good performance for both the human−*B. anthracis* PPI data set and the human−*F. tularensis* PPI data set, and the corresponding AUC values were 0.889 and 0.893, respectively (Figure 5). The ROC curves of these two data sets both showed that the integrative model significantly outperformed the best individual model, further suggesting that our predictive model had strong robustness and generalization ability. Due to the overall lack of experimental HP-PPIs, it is impossible to develop specialized predictors for any pathogenic bacteria. In this context, our predictive model could be employed for predicting PPIs between human and other bacterial species, although the performance may be not optimal.

Comparison with Two Traditional PPI Prediction Methods

We compared our model with two traditional PPI inference methods: interolog mapping and the DDI-based method. To conduct the interolog mapping, we downloaded the intraspecies and interspecies PPIs from seven databases, including BioGRID,[45] DIP,[63] HPIDB,[43] HPRD,[46] IntAct,[64] MINT,[65] and PATRIC.[66] After removing redundant PPIs, self-interactions and human−*Y. pestis* PPIs, the remaining 907634 PPIs containing 108 031 proteins were used as interolog mapping templates. We compared human proteins and *Y. pestis* proteins with these 108031 proteins using BLAST to find homologous relationships with the following criteria: *E*-value of ≤0.00001, identity of ≥30%, and query coverage of ≥40%. The results showed that the corresponding Sensitivity and Specificity values of the interolog mapping method were 0.037 and 0.976, respectively (Table S3).

To conduct the DDI-based PPI inference, 11 200 DDIs were downloaded from the 3did database.[67] A basic assumption of this method is that proteins interact with each other through the domains they contain. If at least one DDI exists between two proteins, we can infer that these two proteins should interact with each other. We used the InterProScan[68] approach for domain scanning. The Sensitivity and Specificity values of the DDI-based prediction method were 0.004 and 0.999, respectively (Table S3). Collectively, the overall performance of these two traditional methods for the human−*Y. pestis* PPI data set was very poor (reflected as low sensitivity), meaning that the development of a specialized interspecies PPI predictor for human−*Y. pestis* is required.

To make a fair comparison, we compared the sensitivity of our model with the interolog mapping and the DDI-based method when the specificity was controlled at 0.999 and 0.976, respectively. As expected, the sensitivity of our model exceeded the two traditional prediction methods at the corresponding specificity control (Table S3).

Comparison with Existing Human−Bacteria PPI Prediction Methods

To comprehensively understand the pros and cons of a newly developed predictor, it is essential to benchmark the proposed method against existing prediction methods. Considering that the field of human−bacteria PPI predictions is far from mature, such a method comparison is challenging. To the best of our knowledge, none of the existing prediction methods provide a web server to the community. The source code provided by any existing method does not ensure that the program can be easily compiled and used properly because of the complexity of preparing the corresponding features. Here, we compared our model with a multitask learning-based method[32] and a multilayer neural network based method.[30]

In the multitask learning-based method, the performance of their model on a human−*Y. pestis* PPI data set containing 750 PPIs and 75 000 non-PPIs (the ratio of positive to negative samples was 1:100) was assessed. We also employed this data set to test our predictive model. To ensure a fair comparison, we removed the PPIs in this test set from our training data and retained 2490 PPIs. In view of the highly imbalanced test set, the ratio of positive samples to negative samples was set as 1:15 to retrain our model. In the multitask learning-based method, the performance was primarily measured by the F1 score, which is defined as the harmonic mean of precision and recall:

$$\mathrm{F1} = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{8}$$

where precision means the percentage of the true positive samples among the predicted positives. As shown in Table 4,

Table 4. Performance Comparison of Our Model and Two Existing ML-Based Human−Bacteria PPI Prediction Methods

method	AUC	AUPR	F1	positives/negatives (training)	positives/negatives (test)
Kshirsagar et al. (2013)[a]	–	–	0.288	1:100	1:100
our model[b]	0.867	0.218	0.291	1:15	1:100
Ahmed et al. (2018)[c]	0.930	0.926	–	1:1	1:1
our model[d]	0.950	0.948	0.896	1:1	1:1

[a]The corresponding F1 value was cited directly from Kshirsagar et al. (2013).[32] [b]The performance of our model was tested on a data set containing 750 PPIs and 75000 non-PPIs between human and *Y. pestis*. Here, the ratio of positive samples to negative samples used in the training model was 1:15. [c]The corresponding AUC and AUPR values were cited directly from Ahmed et al. (2018).[30] [d]Our model was retrained using a data set between human and *B. anthracis* (554 PPIs and 554 non-PPIs), and the performance was assessed through 10-fold cross-validation.

DOI: 10.1021/acs.jproteome.9b00074
J. Proteome Res. 2019, 18, 2195−2205

the optimal F1 score of our model on the test set was 0.291, which was slightly better than that of the multitask learning-based method (0.288). Indeed, the key idea of the multitask learning-based method was taking the commonality of interactions between different bacteria and human proteins into account by employing the pathway information. However, that method still has limitations regarding the acquisition of pathway information. Likewise, the GO annotations of the proteins the authors used to calculate the GO similarity features are difficult to obtain completely. The authors also used the host network topology properties as the feature, but we used more network topology properties than they did, and our integration approach highlighted the importance of this feature type.

Regarding the multilayer neural network-based method, the authors used a data set between human and *B. anthracis* (554 PPIs and 554 non-PPIs) to train their predictive model. To make a fair method comparison, we downloaded their data set from ftp://ftp.sanbi.ac.za/machine_learning/ and then rebuilt our integrative model using these data. Based on the 10-fold cross-validation, they quantified the performance of their model through the ROC curves and precision−recall (PR) curves. Therefore, we also evaluated our model in the same way (Figure S4 and Table 4). The mean AUC of our model for human−*B. anthracis* data was larger than that of their model (0.95 versus 0.93). In terms of the overall AUPR (area under PR curve) values, our model also revealed a better performance. Compared to the multilayer neural network-based method, our model used a different ML algorithm and computational framework. In addition to the sequence encodings, routine host network properties were also used in their method. Comparatively, we employed more comprehensive network property-based features in our model, especially the incorporation of the NetSS encoding scheme. In summary, the above benchmark experiments clearly showed that our method was fully competitive with these two state-of-the-art methods.

Web Server

To assist the community, we also built a web server for our model, which is freely accessible at http://systbio.cau.edu.cn/intersppiv2/ or http://zzdlab.com/intersppiv2/. The web server was implemented on a Linux operating system with CentOS-6.9 and Apache 2.2.15. Users need submit human and bacteria protein sequences in FASTA format; the web server will then calculate the prediction scores of all possible sequence pair combinations. In general, it takes approximately 1 min to complete the prediction for one protein pair. Considering that the ratio of positive to negative samples are highly imbalanced in the real world, it is important to set prediction thresholds at high-specificity controls so that we can ensure that the prediction results are generally reliable (i.e., the prediction yields a relatively high precision). Therefore, we provided three optional Specificity thresholds (0.95, 0.97, and 0.99) in the web server. A larger threshold generally corresponds to a higher prediction precision, but it also results in missing the detection of more true positives. The training data and independent test data used in this work are downloadable through the web server.

Current Limitations

Although the proposed method improves the prediction of human−bacteria PPIs, it has certain limitations in real applications. For instance, if the query human protein is not present in the human PPI network, our model would be invalid or the prediction result would be inaccurate. Indeed, this is a common phenomenon in the prediction of PPIs. We also noticed that an assessment based on balanced samples only may overestimate the performance in practical use due to the highly skewed ratio of positive to negative data in the real world. We conducted the following two computational experiments to elaborate this open issue. In the first experiment, we used the model trained on balanced positive and negative samples to assess its performance on two test data sets, with positive and negative sample ratios of 1:5 and 1:10. In terms of the ROC curves, the performance on these two data sets was quite similar (Figure S3a,b). When calculating the precision values at a 10% FPR control, the precision value decreased from 0.600 in the 1:5 independent test set to 0.426 in the 1:10 independent test set. Therefore, for practical use, one should choose high-specificity controls to ensure a reasonable precision control, which is suggested in our web server. Additionally, the ratio of positives to negatives in the test sets should be identical to ensure a fair comparison among different prediction methods, which was implemented when we compared the proposed method to the multitask learning-based method[32] and the multilayer neural network based method.[30] In the second computational experiment, we retrained the models based on two different ratios of positive to negative data in the training set (1:5 and 1:10), and tested their performance through 5-fold cross-validation (Figure S3c,d). The results showed that the ratio of positive to negative samples in the training data only slightly affect the ROC curves. We also observed minor precision changes based on different ratios of positives to negatives in training. Regarding the precision values at a 10% FPR control in the 1:5 test sets, the precision value increased from 0.600 (inferred from the balanced training set) to 0.613 (inferred from the 1:5 training set). This phenomenon suggests that a balanced positive-to-negative ratio may be not the optimal ratio, but it is reasonable enough to be commonly used in PPI prediction. As an open issue regarding the ML-based PPI prediction, the ratio of positives to negatives should be further taken into account in model training and performance assessment.

CONCLUSIONS

In this work, we have developed an RF-based predictor of *Y. pestis* PPIs. The highlights of the current work are summarized as follows. First, two comprehensive host network-property-related feature vectors reflecting the biological significance of HP-PPIs in network biology were adopted. Second, a suitable computational framework was selected to construct the predictive model. Third, rigorous benchmark experiments were conducted to quantify the performance of our proposed predictor. Finally, a web server that implements the proposed predictor has been made freely accessible to the community. Regarding future developments, the ML-based HP-PPI prediction will be more prosperous with the accumulation of experimental HP-PPI data. To provide reliable PPI prediction between human and any pathogenic bacteria, both generic and species-specific human−bacteria PPI predictors should be developed. Some advanced machine learning algorithms (e.g., deep learning)[69] have been applied in intraspecies PPI prediction, and this should also be rapidly employed to predict human−bacteria PPIs. Taken together, we hope this work will provide a useful tool to identify potential interactions or to prioritize targets for further experimental validation, which will

DOI: 10.1021/acs.jproteome.9b00074
J. Proteome Res. 2019, 18, 2195−2205

be helpful for achieving a more comprehensive understanding of the underlying mechanisms of bacterial infection and will provide new hints for drug development.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.9b00074.

Figures showing confusion matrices, the average performance of models constructed from ten different negative samples, and model performance, performance of our method on the data set used by the multilayer neural network-based method; tables showing original values of the seven physicochemical properties for each amino acid and performance comparisons (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: zidingzhang@cau.edu.cn; phone/fax: +86 10 6273 4376.

ORCID

Ziding Zhang: 0000-0002-9296-571X

Author Contributions

#X.L. and S.Y. contributed equally to this work. All the authors have read and approved the final manuscript.

Funding

This work was supported by the National Key Research and Development Program of China (grant no. 2017YFC1200205).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Yuan Zhou at Peking University for stimulating discussions on this topic.

■ REFERENCES

(1) Swaminathan, S.; Padmapriyadarsini, C.; Narendran, G. HIV-Associated Tuberculosis: Clinical Update. *Clin. Infect. Dis.* **2010**, *50* (10), 1377−1386.

(2) Rather, I. A.; Kim, B. C.; Bajpai, V. K.; Park, Y. H. Self-Medication and Antibiotic Resistance: Crisis, Current Challenges, and Prevention. *Saudi J. Biol. Sci.* **2017**, *24* (4), 808−812.

(3) Zoraghi, R.; Reiner, N. E. Protein Interaction Networks as Starting Points to Identify Novel Antimicrobial Drug Targets. *Curr. Opin. Microbiol.* **2013**, *16* (5), 566−572.

(4) Saha, S.; Sengupta, K.; Chatterjee, P.; Basu, S.; Nasipuri, M. Analysis of Protein Targets in Pathogen-Host Interaction in Infectious Diseases: A Case Study on *Plasmodium Falciparum* and *Homo Sapiens* Interaction Network. *Briefings Funct. Genomics* **2017**, *17* (6), 441−450.

(5) Yu, H.; Luscombe, N. M.; Lu, H. X.; Zhu, X.; Xia, Y.; Han, J. D. J.; Bertin, N.; Chung, S.; Vidal, M.; Gerstein, M. Annotation Transfer between Genomes: Protein-Protein Interologs and Protein-DNA Regulogs. *Genome Res.* **2004**, *14* (6), 1107−1118.

(6) He, F.; Zhang, Y.; Chen, H.; Zhang, Z.; Peng, Y. L. The Prediction of Protein-Protein Interaction Networks in Rice Blast Fungus. *BMC Genomics* **2008**, *9* (1), 519.

(7) Singhal, M.; Resat, H. A Domain-Based Approach to Predict Protein-Protein Interactions. *BMC Bioinf.* **2007**, *8* (1), 199.

(8) Dyer, M. D.; Murali, T. M.; Sobral, B. W. Computational Prediction of Host-Pathogen Protein-Protein Interactions. *Bioinformatics* **2007**, *23* (13), i159−i166.

(9) Pang, E.; Lin, K. Yeast Protein-Protein Interaction Binding Sites: Prediction from the Motif-Motif, Motif-Domain and Domain-Domain Levels. *Mol. BioSyst.* **2010**, *6* (11), 2164−2173.

(10) Guo, Y.; Yu, L.; Wen, Z.; Li, M. Using Support Vector Machine Combined with Auto Covariance to Predict Protein-Protein Interactions from Protein Sequences. *Nucleic Acids Res.* **2008**, *36* (9), 3025−3030.

(11) Zhou, Y.; Zhou, Y. S.; He, F.; Song, J.; Zhang, Z. Can Simple Codon Pair Usage Predict Protein-Protein Interaction? *Mol. BioSyst.* **2012**, *8* (5), 1396−1404.

(12) You, Z.-H.; Zhu, L.; Zheng, C.-H.; Yu, H.-J.; Deng, S.-P.; Ji, Z. Prediction of Protein-Protein Interactions from Amino Acid Sequences Using a Novel Multi-Scale Continuous and Discontinuous Feature Set. *BMC Bioinf.* **2014**, *15* (15), S9.

(13) You, Z.-H.; Lei, Y.-K.; Zhu, L.; Xia, J.; Wang, B. Prediction of Protein-Protein Interactions from Amino Acid Sequences with Ensemble Extreme Learning Machines and Principal Component Analysis. *BMC Bioinf.* **2013**, *14* (8), S10.

(14) Huang, Q.; You, Z.; Zhang, X.; Zhou, Y. Prediction of Protein-Protein Interactions with Clustered Amino Acids and Weighted Sparse Representation. *Int. J. Mol. Sci.* **2015**, *16* (5), 10855−10869.

(15) Yang, S.; Li, H.; He, H.; Zhou, Y.; Zhang, Z. Critical Assessment and Performance Improvement of Plant-Pathogen Protein-Protein Interaction Prediction Methods. *Briefings Bioinf.* **2019**, *20* (1), 274−287.

(16) Zhang, Q. C.; Petrey, D.; Deng, L.; Qiang, L.; Shi, Y.; Thu, C. A.; Bisikirska, B.; Lefebvre, C.; Accili, D.; Hunter, T.; et al. Structure-Based Prediction of Protein-Protein Interactions on a Genome-Wide Scale. *Nature* **2012**, *490* (7421), 556−560.

(17) Zahiri, J.; Mohammad-Noori, M.; Ebrahimpour, R.; Saadat, S.; Bozorgmehr, J. H.; Goldberg, T.; Masoudi-Nejad, A. LocFuse: Human Protein-Protein Interaction Prediction via Classifier Fusion Using Protein Localization Information. *Genomics* **2014**, *104* (6), 496−503.

(18) Zahiri, J.; Yaghoubi, O.; Mohammad-Noori, M.; Ebrahimpour, R.; Masoudi-Nejad, A. PPIevo: Protein-Protein Interaction Prediction from PSSM Based Evolutionary Information. *Genomics* **2013**, *102* (4), 237−242.

(19) Hamp, T.; Rost, B. Evolutionary Profiles Improve Protein-Protein Interaction Prediction from Sequence. *Bioinformatics* **2015**, *31* (12), 1945−1950.

(20) Li, Z.-G.; He, F.; Zhang, Z.; Peng, Y.-L. Prediction of Protein-Protein Interactions between *Ralstonia Solanacearum* and *Arabidopsis Thaliana*. *Amino Acids* **2012**, *42* (6), 2363−2371.

(21) Schleker, S.; Garcia-Garcia, J.; Klein-Seetharaman, J.; Oliva, B. Prediction and Comparison of *Salmonella*-Human and *Salmonella*-*Arabidopsis* Interactomes. *Chem. Biodiversity* **2012**, *9* (5), 991−1018.

(22) Evans, P.; Dampier, W.; Ungar, L.; Tozeren, A. Prediction of HIV-1 Virus-Host Protein Interactions Using Virus and Host Sequence Motifs. *BMC Med. Genomics* **2009**, *2* (1), 27.

(23) Nourani, E.; Khunjush, F.; Durmuåÿ, S. Computational Approaches for Prediction of Pathogen-Host Protein-Protein Interactions. *Front. Microbiol.* **2015**, *6* (FEB), 94.

(24) Mei, S. Probability Weighted Ensemble Transfer Learning for Predicting Interactions between HIV-1 and Human Proteins. *PLoS One* **2013**, *8* (11), No. e79606.

(25) Qi, Y.; Tastan, O.; Carbonell, J. G.; Klein-Seetharaman, J.; Weston, J. Semi-Supervised Multi-Task Learning for Predicting Interactions between HIV-1 and Human Proteins. *Bioinformatics* **2010**, *26* (18), i645−i652.

(26) Emamjomeh, A.; Goliaei, B.; Zahiri, J.; Ebrahimpour, R. Predicting Protein-Protein Interactions between Human and Hepatitis C Virus via an Ensemble Learning Method. *Mol. BioSyst.* **2014**, *10* (12), 3147−3154.

(27) Nourani, E.; Khunjush, F.; Durmuş, S. Computational Prediction of Virus−human Protein−protein Interactions Using

DOI: 10.1021/acs.jproteome.9b00074
J. Proteome Res. 2019, 18, 2195−2205

Embedding Kernelized Heterogeneous Data. *Mol. BioSyst.* **2016**, *12* (6), 1976−1986.
(28) Dyer, M. D.; Murali, T. M.; Sobral, B. W. Supervised Learning and Prediction of Physical Interactions between Human and HIV Proteins. *Infect., Genet. Evol.* **2011**, *11* (5), 917−923.
(29) Barman, R. K.; Saha, S.; Das, S. Prediction of Interactions between Viral and Host Proteins Using Supervised Machine Learning Methods. *PLoS One* **2014**, *9* (11), No. e112034.
(30) Ahmed, I.; Witbooi, P.; Christoffels, A. Prediction of Human-*Bacillus Anthracis* Protein-Protein Interactions Using Multi-Layer Neural Network. *Bioinformatics* **2018**, *34* (24), 4159−4164.
(31) Huo, T.; Liu, W.; Guo, Y.; Yang, C.; Lin, J.; Rao, Z. Prediction of Host-Pathogen Protein Interactions between *Mycobacterium Tuberculosis* and *Homo Sapiens* Using Sequence Motifs. *BMC Bioinf.* **2015**, *16* (1), 100.
(32) Kshirsagar, M.; Carbonell, J.; Klein-Seetharaman, J. Multitask Learning for Host-Pathogen Protein Interactions. *Bioinformatics* **2013**, *29* (13), i217−i226.
(33) Mei, S.; Zhu, H. AdaBoost Based Multi-Instance Transfer Learning for Predicting Proteome-Wide Interactions between *Salmonella* and Human Proteins. *PLoS One* **2014**, *9* (10), No. e110488.
(34) Rapanoel, H. A.; Mazandu, G. K.; Mulder, N. J. Predicting and Analyzing Interactions between *Mycobacterium Tuberculosis* and Its Human Host. *PLoS One* **2013**, *8* (7), No. e67472.
(35) Schleker, S.; Kshirsagar, M.; Klein-Seetharaman, J. Comparing Human-*Salmonella* with Plant-*Salmonella* Protein-Protein Interaction Predictions. *Front. Microbiol.* **2015**, *6* (JAN), 45.
(36) Dyer, M. D.; Neff, C.; Dufford, M.; Rivera, C. G.; Shattuck, D.; Bassaganya-Riera, J.; Murali, T. M.; Sobral, B. W. The Human−Bacterial Pathogen Protein Interaction Networks of *Bacillus Anthracis*, *Francisella Tularensis*, and *Yersinia Pestis*. *PLoS One* **2010**, *5* (8), No. e12089.
(37) Yang, H.; Ke, Y.; Wang, J.; Tan, Y.; Myeni, S. K.; Li, D.; Shi, Q.; Yan, Y.; Chen, H.; Guo, Z.; et al. Insight into Bacterial Virulence Mechanisms against Host Immune Response via the *Yersinia Pestis*-Human Protein-Protein Interaction Network. *Infect. Immun.* **2011**, *79* (11), 4413−4424.
(38) Doxey, A. C.; McConkey, B. J. Prediction of Molecular Mimicry Candidates in Human Pathogenic Bacteria. *Virulence* **2013**, *4* (6), 453−466.
(39) Gowthaman, U.; Eswarakumar, V. P. Molecular Mimicry: Good Artists Copy, Great Artists Steal. *Virulence* **2013**, *4* (6), 433−434.
(40) Guven-Maiorov, E.; Tsai, C.-J.; Nussinov, R. Pathogen Mimicry of Host Protein-Protein Interfaces Modulates Immunity. *Semin. Cell Dev. Biol.* **2016**, *58*, 136−145.
(41) Riedel, S. Plague: From Natural Disease to Bioterrorism. *Baylor Univ. Med. Cent. Proc.* **2005**, *18* (2), 116−124.
(42) Gilbert, M. T. P. *Yersinia Pestis:* One Pandemic, Two Pandemics, Three Pandemics, More? *Lancet Infect. Dis.* **2014**, *14* (4), 264−265.
(43) Ammari, M. G.; Gresham, C. R.; McCarthy, F. M.; Nanduri, B. HPIDB 2.0: A Curated Database for Host−pathogen Interactions. *Database* **2016**, *2016*, baw103.
(44) Wattam, A. R.; Davis, J. J.; Assaf, R.; Boisvert, S.; Brettin, T.; Bun, C.; Conrad, N.; Dietrich, E. M.; Disz, T.; Gabbard, J. L.; et al. Improvements to PATRIC, the All-Bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res.* **2017**, *45*, D535−D542.
(45) Stark, C. BioGRID: A General Repository for Interaction Datasets. *Nucleic Acids Res.* **2006**, *34*, D535−D539.
(46) Peri, S. Human Protein Reference Database as a Discovery Resource for Proteomics. *Nucleic Acids Res.* **2004**, *32*, 497−501.
(47) Brown, K. R.; Jurisica, I. Unequal Evolutionary Conservation of Human Protein Interactions in Interologous Networks. *Genome Biol.* **2007**, *8* (5), R95.
(48) Apweiler, R. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* **2009**, *38*, D190−D195.
(49) Kotlyar, M.; Pastrello, C.; Pivetta, F.; Lo Sardo, A.; Cumbaa, C.; Li, H.; Naranian, T.; Niu, Y.; Ding, Z.; Vafaee, F.; et al. In Silico Prediction of Physical Protein Interactions and Characterization of Interactome Orphans. *Nat. Methods* **2015**, *12* (1), 79−84.
(50) Chen, K.; Kurgan, L. A.; Ruan, J. Prediction of Flexible/Rigid Regions from Protein Sequences Using k-Spaced Amino Acid Pairs. *BMC Struct. Biol.* **2007**, *7* (1), 25.
(51) Chen, Y.-Z.; Tang, Y.-R.; Sheng, Z.-Y.; Zhang, Z. Prediction of Mucin-Type O-Glycosylation Sites in Mammalian Proteins Using the Composition of k-Spaced Amino Acid Pairs. *BMC Bioinf.* **2008**, *9* (1), 101.
(52) Bonacich, P. Some Unique Properties of Eigenvector Centrality. *Soc. Networks* **2007**, *29* (4), 555−564.
(53) Brin, S.; Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Comput. Networks ISDN Syst.* **1998**, *30* (1−7), 107−117.
(54) Csárdi, G.; Nepusz, T. The Igraph Software Package for Complex Network Research. *InterJournal Complex Syst.* **2006**, *1695*, 1−9.
(55) Needleman, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48* (3), 443−453.
(56) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215* (3), 403−410.
(57) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1947−1958.
(58) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explor. Newsl.* **2009**, *11* (1), 10.
(59) Peng, H.; Long, F.; Ding, C. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27* (8), 1226−1238.
(60) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2* (1−3), 37−52.
(61) Chang, C.-C.; Lin, C.-J. Libsvm: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2* (3), 1−27.
(62) Bánky, D.; Iván, G.; Grolmusz, V. Equal Opportunity for Low-Degree Network Nodes: A PageRank-Based Method for Protein Target Identification in Metabolic Graphs. *PLoS One* **2013**, *8* (1), No. e54204.
(63) Salwinski, L. The Database of Interacting Proteins: 2004 Update. *Nucleic Acids Res.* **2004**, *32* (32), 449−451.
(64) Hermjakob, H. IntAct: An Open Source Molecular Interaction Database. *Nucleic Acids Res.* **2004**, *32*, 452−455.
(65) Licata, L.; Briganti, L.; Peluso, D.; Perfetto, L.; Iannuccelli, M.; Galeota, E.; Sacco, F.; Palma, A.; Nardozza, A. P.; Santonico, E.; et al. MINT, the Molecular Interaction Database: 2012 Update. *Nucleic Acids Res.* **2012**, *40* (D1), D857−D861.
(66) Wattam, A. R.; Abraham, D.; Dalay, O.; Disz, T. L.; Driscoll, T.; Gabbard, J. L.; Gillespie, J. J.; Gough, R.; Hix, D.; Kenyon, R.; et al. PATRIC, the Bacterial Bioinformatics Database and Analysis Resource. *Nucleic Acids Res.* **2014**, *42*, D581−D591.
(67) Stein, A.; Céol, A.; Aloy, P. 3did: Identification and Classification of Domain-Based Interactions of Known Three-Dimensional Structure. *Nucleic Acids Res.* **2011**, *39*, D718−D723.
(68) Quevillon, E.; Silventoinen, V.; Pillai, S.; Harte, N.; Mulder, N.; Apweiler, R.; Lopez, R. InterProScan: Protein Domains Identifier. *Nucleic Acids Res.* **2005**, *33*, W116−W120.
(69) Sun, T.; Zhou, B.; Lai, L.; Pei, J. Sequence-Based Prediction of Protein Protein Interaction Using a Deep-Learning Algorithm. *BMC Bioinf.* **2017**, *18* (1), 277.

DOI: 10.1021/acs.jproteome.9b00074
J. Proteome Res. 2019, 18, 2195−2205