BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 03/10 pp21–31
Volume 6, Number 1, March 2023
DOI: 10.26599/BDMA.2022.9020031

RF-PSSM: A Combination of Rotation Forest Algorithm and Position-Specific Scoring Matrix for Improved Prediction of Protein-Protein Interactions Between Hepatitis C Virus and Human

Xin Liu*, Yaping Lu, Liang Wang*, Wei Geng, Xinyi Shi, and Xiao Zhang*

Abstract: The identification of hepatitis C virus (HCV) virus-human protein interactions will not only help us understand the molecular mechanisms of related diseases but also be conductive to discovering new drug targets. An increasing number of clinically and experimentally validated interactions between HCV and human proteins have been documented in public databases, facilitating studies based on computational methods. In this study, we proposed a new computational approach, rotation forest position-specific scoring matrix (RF-PSSM), to predict the interactions among HCV and human proteins. In particular, PSSM was used to characterize each protein, two-dimensional principal component analysis (2DPCA) was then adopted for feature extraction of PSSM. Finally, rotation forest (RF) was used to implement classification. The results of various ablation experiments show that on independent datasets, the accuracy and area under curve (AUC) value of RF-PSSM can reach 93.74% and 94.29%, respectively, outperforming almost all cutting-edge research. In addition, we used RF-PSSM to predict 9 human proteins that may interact with HCV protein E1, which can provide theoretical guidance for future experimental studies.

Key words: protein-protein interactions; hepatitis C virus; position specific scoring matrix; two-dimensional principal component analysis; rotation forest

1 Introduction

Viral diseases, which are caused by various viruses, kill millions of people every year. For example, nearly 71 million people died each year from the hepatitis C virus (HCV) complications such as cirrhosis and hepatocellular carcinoma[1,2]. The notorious Ebola virus affected 28 000 cases and over 11 000 deaths were reported during the 2014 epidemic[3]. For many viral diseases, there is currently no effective vaccine or treatment, due to the unclear pathogenic mechanisms and fast mutation rates of virus genomes[4].

Therefore, the identification of interactions among viral and host proteins is significant for comprehending the molecular mechanisms of viral infection and identifying antiviral drugs[5]. In this paper, we mainly studied interactions among HCV and humans. HCV genome can be translated into 11 proteins, including four structural proteins, six non-structural proteins, and an F protein[6–8]. Chronically infected HCV patients usually present liver injuries associated with hepatic cirrhosis,

• Xin Liu, Wei Geng, and Xiao Zhang are with the School of Medical Informatics and Engineering, Xuzhou Medical University, Xuzhou 221000, China. E-mail: liuxin@xzhmu.edu.cn; gw@xzhmu.edu.cn; changshui@hotmail.com.
• Yaping Lu is with the College of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China. E-mail: yplu@cumt.edu.cn.
• Liang Wang is with the Laboratory Medicine, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou 510080, China. E-mail: healthscience@foxmail.com.
• Xinyi Shi is with the Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310005, China. E-mail: 1613679323@qq.com.
* To whom correspondence should be addressed.
Manuscript received: 2022-07-04; revised: 2022-08-22; accepted: 2022-08-30

hepatic steatosis, and hepatocarcinoma if not properly treated[9]. However, the treatment of HCV is very expensive, and also has severe adverse effects. Thus, if we can obtain a clear map of HCV-human protein-protein interactions (PPIs), it would make us better understand the mechanisms of HCV infection. However, due to the shortcomings of traditional biological experimental methods, such as high cost and long cycles, large-scale screening is not possible. By contrast, computational methods solve this issue with higher efficiency and more accuracy, which have been gaining more and more attention[10, 11].

In proteomic studies, a growing number of machine learning algorithms are proposed to predict PPIs in the same species[12, 13], however the PPIs between different species deserve more attention. Moreover, PPIs are more conservative within species than between species[14], thus traditional computational method for PPIs prediction may not be suitable for PPIs between different species. Recently, along with the advancement of biological experimental technology, more and more virus host PPI data have been accumulated, which provides a good data foundation for the research based on machine learning prediction methods. The earliest research on virus-host interaction based on machine learning was reported in 2012[15], which focused on the model construction on human papilloma virus (HPV) and HCV interactions with human proteins based on transforming protein sequences into amino acid triplets, respectively. Next, the authors investigated the performance of HCV-host interaction prediction models based on more feature representation methods and ensemble learning[16]. In particular, a total of 6 features such as amino acid composition (ACC) were used to encode proteins, and ensemble learning was developed based on four different base classifiers. In addition, some people took into account information such as network structure information[17]. Certainly, more studies prefer to construct features directly from amino acid sequence, such as the frequency difference between amino acid triplets (FDAT)[18], repeat patterns and compositions of amino acids[19], and so on. Although machine learning has made many achievements in the field of virus-host PPIs prediction, it is still challenging to develop effective models to improve the predicting performance of viral-host PPIs.

In general, feature representation and selection of classification models are two key factors for successful constructions of PPI predicting models. In order to achieve good predicting performance, many models first combine multiple features and then build classification models through appropriate feature selection or feature dimensionality reduction, which increases the complexity of the model. According to Occam's Razor theory, the simpler, the better[20]. Therefore, in this paper, we aim to build a valid classifier with fewer feature representations. Considering that position-specific scoring matrix (PSSM) has been widely used in various proteomics studies with good capability[21–23], such as subcellular locations[24], protein secondary structure prediction[25], protein folding patterns[26], and di-sulfide connectivity[27]. However, most of them just converted the PSSM from matrix $L \times 20$ to vector of length $L \times 20$, which may lead to information loss. Therefore, for the purpose of fully utilizing the information contained in PSSM, we proposed a new computational model that could obtain more information from PSSM by adopting an effective feature extraction method.

The model present in this study is based on rotation forest (RF) and PSSM, termed RF-PSSM, which is a predictor of HCV-human PPIs based on feature extraction from PSSM. Specifically, position specific-iterated basic local alignment search tool (PSI-BLAST) was used to generate PSSM for each protein, and then 2DPCA was used to further extract features from PSSM. Finally, RF was utilized as classifier. We have conducted several ablation experiments, and the results show that RF-PSSM is superior to almost all the cutting-edge methods. Furthermore, RF-PSSM was used to predict proteins that may interact with E1, which may provide theoretical guidance for subsequent experimental verification.

2 Experiment

2.1 Datasets

The HCV-human PPIs dataset was sourced from VirHostNet[28], which contains 477 PPIs among HCV and human proteins, and is treated as a positive dataset. Then we randomly select 477 datasets that do not overlap with positive samples from the Human Protein Reference Database (HPRD) as negative dataset. Therefore, the HCV-human PPIs dataset is a balanced dataset with 954 sets of data (see Table S1, which is in the Electronic Supplmentary Material (ESM) of the online version of this article). Then, the HCV-human PPIs dataset was divided into two parts: 20% was utilized as independent dataset for test ($n = 191$) and the remaining as training dataset ($n = 763$).

The PPI network was shown in Fig. 1, which was generated by CytoScape (version 3.8.0). The network includes 432 nodes in which 11 hub nodes (large nodes) are HCV proteins while 421 nodes are human proteins (small nodes). A total of 477 edges were drawn to depict the PPIs between HCV and human. The specific quantity of human proteins interacting with each HCV protein is shown in Table 1. The third column represent the length of HCV protein, where the bracket AA represents the length unit of amino acid.

2.2 Algorithms

2.2.1 PSSM

For an element $P_{i,j}$ in PSSM, its value indicates the possibility that the i-th amino acid is mutated into the j-th amino acid during evolution. If the value is positive, it indicates the greater the possibility; Otherwise, it means that the probability is smaller. PSSM has been used in various bioinformatics fields because of its high quality to preserve the evolutionary information of each protein[22,29]. The PSSM is obtained by PSI-BLAST which hunting for the NCBI non-redundant database[30,31]. The PSSM matrix of protein with length L is $L \times 20$[32].

2.2.2 2DPCA

2DPCA is usually applied to two-dimensional matrices such as digital images and board games, etc., and has become an effective method by reducing the computational complexity and singularity during feature extraction[33–36]. Therefore, we try to use 2DPCA for feature extraction of the PSSMs that we constructed in this study. Suppose there are L samples, then the i-th sample can be expressed as a matrix P_i of $m \times n$ ($i = 1, 2, \ldots, L$), and $\bar{P}$ represents the average of P_i. Project each P onto the best projection matrix, and the formula is as follows:

$$H = PX \tag{1}$$

Therefore, H is projection vector. X is an n-dimensional column vector. The best projection axis X is defined by the divergence distribution of H:

$$J(X) = trace(S_x) \tag{2}$$

where S_x represents the covariance matrix of H, $trace(S_x)$ denotes the trace of S_x.

$$trace(S_x) = trace(X^{\mathrm{T}}\left[E(P - E(P))^{\mathrm{T}}(P - E(P))\right]X) \tag{3}$$

where E represents the expectation.

M_t represents total scatter matrix, as follows:

$$M_t = E\left[(P - E(P))^{\mathrm{T}}(P - E(P))\right] = \frac{1}{L}\sum_{i=1}^{L}\left(P_j - \bar{P}\right)^{\mathrm{T}}(P_i - \bar{P}) \tag{4}$$

Consequently, the criterion function is shown as below:

$$J(X) = trace(X^{\mathrm{T}} M_t X) \tag{5}$$

The first d eigenvalues constitute the best orthogonal projection axis $X_1, X_2, \ldots, X_d$, the matrix P is projected into the projection axis, and the formula is as follows:

$$H_k = PX_k, k = 1, 2, \ldots, d \tag{6}$$

Finally, a new set of eigenvectors of matrix P, $H_1, H_2, \ldots, H_d$, can be obtained by above calculation. Therefore, 2DPCA can retain as much useful information as possible.

2.2.3 Rotation forest

Rotation forest uses linear analysis theory and decision tree ensemble classification algorithm, which can still achieve good results even with few classifiers, and can ensure the performance of ensemble classification[37]. Since RF was proposed, it has been used in protein interaction prediction and cancer classification[38,39], and so on. The detailed process is as follows.

First, suppose Z is training sample set, Y is corresponding label, and F is feature set. Then the training set Z containing N samples and n features can be expressed as $N \times n$. In this study, we adopt k-nearest

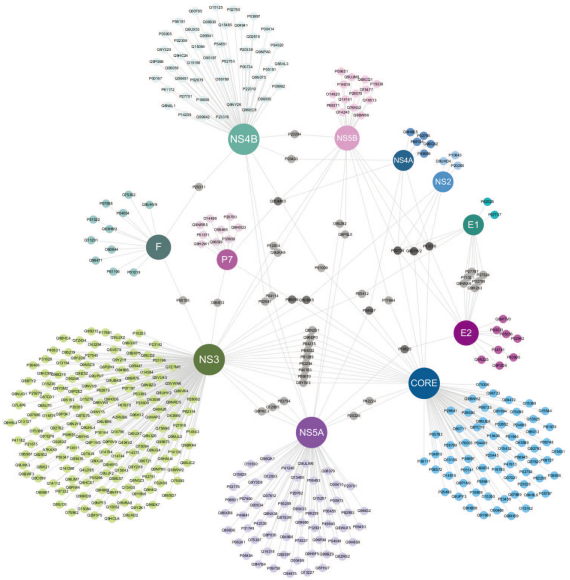
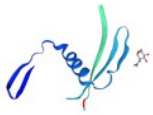
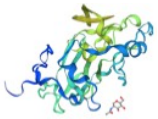
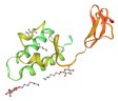
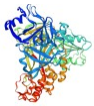
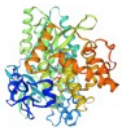
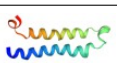


Fig. 1 Illustration of PPIs network among HCV and human. The network was constructed and visualized by CytoScape (version 3.8.0).

Table 1 Basic description of HCV proteins and the quantity of human proteins interacting with each HCV protein.

HCV protein	UniProt ID	Length (AA)	PDB ID	3D structure	Number of human proteins
Core	NP_751919.1	191	1xcq.2.C		84
E1	NP_751920.1	192	4uoi.3.B		11
E2	NP_751921.1	363	6mej.1.C		18
F	NP_803170.1	161	2q6u.1.A		13
NS2	NP_751923.1	217	2hd0.1.A		7
NS3	NP_803144.1	631	3o8d.1.A		177
NS4A	NP_751925.1	54	6uju.1.A		10
NS4B	NP_751926.1	261	2kdr.1.A		52
NS5A	NP_751927.1	448	4cl1.1.A		69
NS5B	NP_751928.1	591	3hkw.1.A		24
P7	NP_751922.1	63	3zd0.1.A		12

Note: 3D structures of HCV proteins were constructed via Swiss-Model for demonstration purposes.

neighbor (KNN) as basic model, for which there are two parameters that need to be predefined. The first is Q, which is the number of basic classifiers in a rotation forest; the second is K, which is the number of feature subset. The implementation of the RF can be divided into training phase and classification phase. The training process of the RF is shown as follows. Divide the feature set F into K subsets $F_{i,j}$, $j = 1, 2, \ldots, k$. For each subset $F_{i,j}$, first choose the homologous feature column in subset $F_{i,j}$ in training sample set Z, forming a novel matrix $Z_{i,j}$; 75% sample of $Z_{i,j}$ are then collected by bootstrapping approach, forming a matrix $Z'_{i,j}$; feature transformation is finally made in $Z'_{i,j}$ to obtain the matrix $D_{i,j}$. The j-th column in $D_{i,j}$ is the coefficient of j-th feature component. Construct a block diagonal matrix R_i by matrix $D_{i,j}$; Adjust the rows of matrix R_i to be accordance with the feature order of feature set F, and obtain the rotation matrix R_i^a.

$$R_i^a = \begin{bmatrix} a_{i,1}^{(1)}, \ldots, a_{i,1}^{(M_1)} & 0 & \cdots & 0 \\ 0 & a_{i,2}^{(1)}, \ldots, a_{i,2}^{(M_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{i,K}^{(1)}, \ldots, a_{i,K}^{(M_K)} \end{bmatrix} \tag{7}$$

The classification procedure is as follows. $ZR_i^a \ Y$ is the input of basic classifier D_i; for the test sample z, D_i could generate the labels y_i; the confidence level can be obtained by averaging D_i results:

$$\omega_j(z) = \frac{1}{Q} \sum_{i=1}^{Q} D_{i,j}\left(ZR_i^a\right) \tag{8}$$

Then assign the category with the largest $\omega_j(z)$ value to z.

2.2.4 Support vector machine

The basic principle of support vector machine (SVM) is to realize the linear classification function by finding the separation hyperplane that maximizes the interval in the feature space[40,41]. Specifically, if samples are linearly separable, linear classifier could be learned; if samples are approximately linearly separable, a slack variable is introduced, and a soft margin is maximized to learn a linear classifier. When samples are linearly inseparable, the kernel technique and soft interval maximization can be used to learn the nonlinear SVM. SVM has good generalization ability and has excellent performance in various fields, including antifungal peptides prediction[42], cancer prediction[43,44], protein secondary structure prediction[45], and so on.

2.2.5 *K*-fold cross-validation

Cross-validation can not only solve the problem that the amount of data in the dataset is not large enough, but also solve the problem of parameter tuning. *K*-fold cross-validation means separate the dataset into K mutually exclusive subsets of the same size, and keep the distribution consistent by sampling K subsets in a stratified manner. $K-1$ of these sets are utilized for training, and the remaining set for evaluating, and the average value obtained after K repetitions is used as a measure of model performance.

2.3 Overall procedure

First of all, the positive dataset is obtained by searching in the VirHostNet database, which has 477 HCV-human interaction datasets. Then, the negative sample containing 477 non-interacting pairs was constructed by searching for human proteins from the HPRD that did not interact with HCV. After that, 80% of the data are used as training sets and the rest are independent sets. In this paper, PSSM is used for protein characterization, and 2DPCA was utilized to extract latent feature from PSSM. Finally, rotating forests are used as classifiers and the models are evaluated using cross-validation and other methods.

2.4 Performance measurement

As usual, we adopted the following six metrics to evaluate RF-PSSM, the first five of them are calculated as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \times 100\% \tag{9}$$

$$\mathrm{Sen} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100\% \tag{10}$$

$$\mathrm{Spec} = \frac{\mathrm{TN}}{\mathrm{FN}+\mathrm{TN}} \times 100\% \tag{11}$$

$$\mathrm{Pre} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \times 100\% \tag{12}$$

$$\mathrm{MCC} = \frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}} \times 100\% \tag{13}$$

ACC represents the correct proportion predicted in all samples. Sen represents the proportion of all instances that are actually positive and are predicted to be positive, which is equivalent to the recall. Spec describes the proportion of predicted negative samples to actual negative samples. Pre represents the proportion of the predicted positive samples to the actual positive samples. MCC describes the correlation coefficient between actual classification and prediction classification, and its value range is $[-1, 1]$.

Among them, P in TP, TN, FP, FN means positive, N means negative, and T means correctly predicted, N means wrongly predicted. For example, TP represents the number of positive samples that are correctly predicted. Area under curve (AUC) and receiver operating characteristic (ROC) curve are also used to evaluate performance of RF-PSSM.

3 Result and Discussion

3.1 Performance of RF-PSSM on the training dataset

In this paper, the value of two important parameters K and L in RF are got by grid search ($K = 6, L = 5$), respectively, where K means the quantity of feature subsets, L means the quantity of base classifiers. The

performance on training dataset is depict in Table 2. As seen from Table 2 that the variance of each indicator changes gently which means the model is relatively robust. For example, the variances of ACC and Sen are 2.63% and 2.00%, respectively.

Meanwhile, the ROC curves of five-fold cross-validations are shown in Fig. 2. The mean AUC value of RF-PSSM on five-fold cross-validation is 0.9286, and with gentle change in each validation, which show stability of RF-PSSM.

3.2 Comparison with other methods

To evaluate the effectiveness of the rotation forest algorithm in the HCV-host interaction model, we compared it with the classical SVM and some other ensemble learning methods, respectively.

3.2.1 Comparison with SVM-based methods

Considering the ubiquitous application of SVM in the field of bioinformatics, we made comparison with it to verify the effectiveness of RF-PSSM[42–45]. In specificity, we obtained the best kernel function, c and g by grid search, respectively[41,46].

In the SVM-based model, the feature representation and feature selection methods are the same as those of RF-PSSM, we first analyze its performance on the training set. The five-fold cross-validations results of SVM on training dataset was depicted in Table 3. It can be seen from Table 3 that the SVM-based model is not very robust because many indicators fluctuate greatly. For example, the ACC value varies between 68.42% and 82.24%. To be more intuitive, the ROC curves of the five-fold cross-validations are shown in Fig. 3.

Table 2 Performance of RF-PSSM on training dataset.

Training set	ACC (%)	Sen (%)	Pre (%)	MCC (%)	AUC
1	94.08	100.00	89.16	88.81	0.9553
2	92.76	98.77	88.89	86.32	0.9106
3	91.45	98.61	85.54	84.27	0.9160
4	88.16	100.00	0.7978	78.77	0.9128
5	94.84	95.12	95.12	90.18	0.9484
Average	92.26± 2.63	98.50± 2.00	87.70± 5.61	85.67± 4.48	0.9286± 0.0214

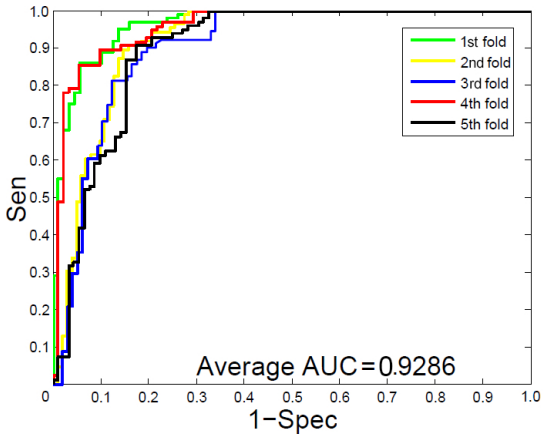


Fig. 2 ROC curves of RF-PSSM on training dataset.

As can be seen from Tables 2 and 3, RF-PSSM not only outperforms SVM on the training dataset in all metrics, but is also more stable, as the variance of each metric is also lower than that of SVM.

Furthermore, we also made comparison of the two algorithms on independent datasets that are virtually unaffected by the training dataset. The result is shown in Table 4. It can be seen that RF-PSSM is better than SVM in all aspects, especially in the Sen value RF is 19.59% higher than SVM. This may be due to the fact that RF integrates multiple base classifiers to improve model performance.

3.2.2 Comparison with ensemble learning-based methods

Since the rotation forest adopted in this paper is an ensemble learning algorithm, in order to further verify its performance, we compared it with some other excellent

Table 3 Performance of SVM-based model on training dataset.

Training set	ACC (%)	Sen (%)	Pre (%)	MCC (%)	AUC
1	82.24	90.54	77.01	70.47	0.9414
2	81.58	92.59	77.32	68.73	0.9231
3	68.42	69.44	65.79	56.73	0.8720
4	76.97	95.77	68.00	62.69	0.8943
5	74.19	79.27	73.86	61.29	0.9001
Average	76.68± 5.69	85.52± 10.93	72.40± 5.26	63.89± 5.62	0.9062± 0.0268

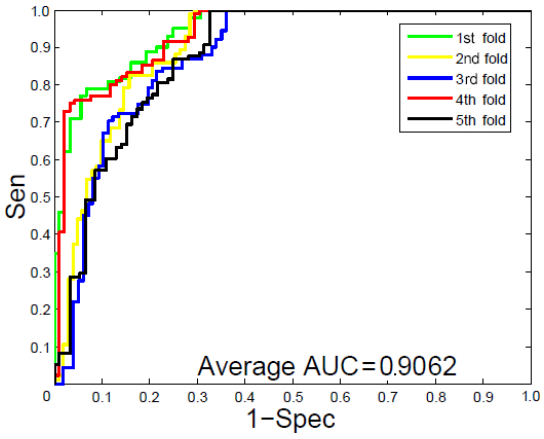


Fig. 3 ROC curves of SVM-based method.

Table 4 Comparison performance of RF-PSSM and SVM on independent dataset.

Model	ACC (%)	Sen (%)	Pre (%)	MCC (%)	AUC
RF-PSSM	93.74	98.97	89.72	88.14	0.9429
SVM	89.01	79.38	87.62	80.10	0.9061

ensemble learning algorithms, such as random forest, Xgboost, and Adaboost[47]. We first adopted the feature representation method proposed in this paper, and then use random forest and Adaboost as classifiers to build models, and compared them with RF-PSSM on the training set and independent dataset. The results are shown in Fig. 4.

As can be seen from Fig. 4 the rotation forest slightly outperforms random forest and Adaboost on both the training and independent datasets. This indicates that rotation forest is more suitable for the characteristics of HCV-host interaction data presented in this paper than other integrated learning algorithms.

3.3 Comparison to other protein representations

Feature representation is critical to the construction of predictive models, it determines the upper bound of model performance. Even though many excellent feature representation methods have been proposed in previous studies for HCV-host PPIs prediction, such as amino acid triplet, physical and chemical properties, PSSM, network structure information, PTM, and so on[15–19]. However, they lack in-depth mining of features, and the performance of the model can be further improved. Considering that PSSM contains both the position information of amino acid sequence and chemical information, we adopted 2DPCA to extract features of PSSM, which is called 2DPCA-PSSM. To verify the effectiveness of it, we compared the performance of 2DPCA-PSSM with PSSM and four other feature representation methods on independent datasets, respectively.

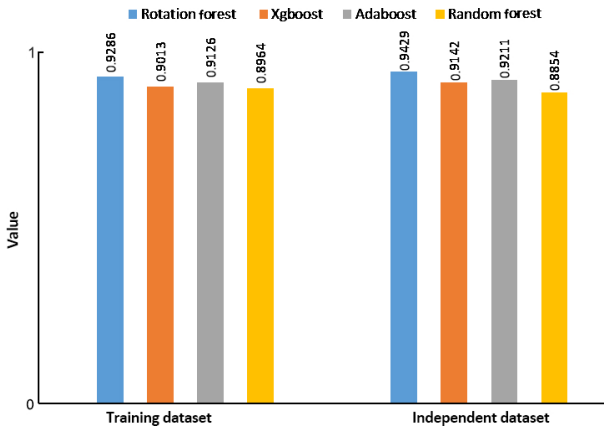


Fig. 4 Comparison of AUC values between rotation forest, Xgboost, Adaboost, and random forest, on different dataset.

For the comparison of 2DPCA-PSSM and PSSM, we multiplied the PSSM matrix of each protein with its transpose matrix to obtain a 20×20 matrix, which is then vectorized to 400 dimensions. Thus, for an HCV-host interaction pair, it can be represented as 800-dimensional vector. The ROC curve obtained on the independent dataset are shown in Fig. 5. It can be seen that the AUC value based on PSSM model only achieve 0.6669, which is 27.6% lower than that of using 2DPCA-PSSM. This indicates that 2DPCA effectively extracts latent information of PSSM, while direct vectorization of the PSSM may lose a lot of location information.

Here we selected four commonly used protein feature representation methods and compare them with 2DPCA-PSSM, including amino acid composition, autocorrelation, pseudo amino acid composition, and profile-based features[48]. Since many features contain high-dimensional redundant information, we first adopted extra-tree for dimension reduction. The comparison results on independent datasets are shown

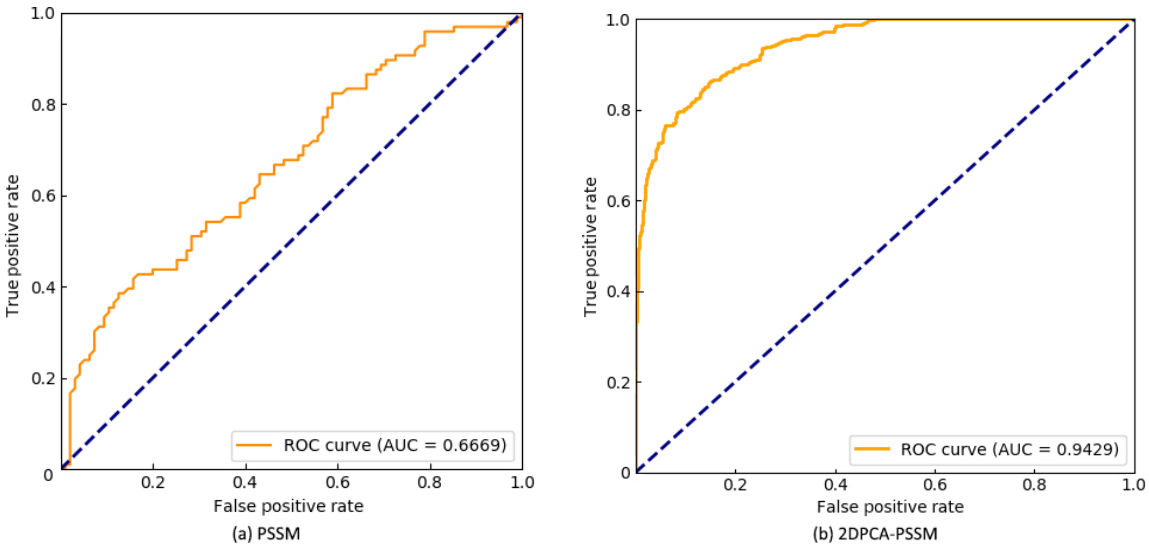


Fig. 5 ROC curve of PSSM and 2DPCA-PSSM on independent dataset.

below. As shown in Fig. 6, 2DPCA-PSSM outperforms the other four features in all metrics. In particular, the MCC value is nearly double that of autocorrelation.

In conclusion, the feature representation method of 2DPCA-PSSM can better represent the key features of HCV-host interaction.

3.4 Comparison with cutting-edge methods

Finally, we also made a comparison between RF-PSSM and several cutting-edge methods in Table 5. As shown in Table 5, we can see that RF-PSSM outperforms other methods in basically all metrics, except for the Pre value obtained using the naive Bayes method in Ref. [17]. The reason may be mainly due to the comprehensive use of 2DPCA for effective feature extraction of PSSM and the rotation forest model. At the same time, we analyze that the reason why some studies integrate multiple features but their performance is not very good may be that they have not conducted in-depth research on feature extraction. To sum up, effective feature engineering and models are indispensable.

3.5 Case study

Furthermore, we adopted RF-PSSM to find proteins that have potential interaction with E1. The steps are as follows: (1) NCBI protein BLAST was adopted to look for proteins similar to the proteins that interact with E1; (2) Screening for all proteins with similarity of 60% yielded in the first step; (3) PSSM features of these proteins were obtained by PSI-BLAST, and further features were extracted by 2DPCA; (4) The RF-PSSM is used to predict the potential proteins which are likely to interact with HCV E1. Eventually, 9 proteins were predicted to potentially interact with E1 (Table S2 in the ESM).

Table 5 Comparison of cutting-edge methods and RF-PSSM on independent dataset.

Model	ACC (%)	Sen (%)	Pre (%)	MCC (%)	AUC
SVM[15]	81.60	77.80	–	–	–
MLP[16]	83.00	84.00	–	–	–
SVM[17]	74.00	67.00	72.00	44.00	0.7300
Naive Bayes[17]	68.50	37.49	98.80	47.00	0.7100
Random forest[17]	72.41	55.66	82.26	48.00	0.7600
SVM[18]	88.80	89.40	88.60	77.40	–
SVM[19]	73.20	94.37	66.30	51.20	0.9250
RF-PSSM	93.74	98.97	89.72	88.14	0.9429

4 Conclusion

PSSM is often used in various proteomic studies along with other features after simple processing (vectorization from 2 to 1). However, few studies have been devoted to the effective feature extraction of PSSM. In this study, we proposed the RF-PSSM method, which first extracted the effective features from PSSM through 2DPCA, and then rotation forest was used to establish the prediction model. The experiment results showed the satisfactory prediction performance of the RF-PSSM. We also compared 2DPCA-PSSM with PSSM and four other feature representation methods to further verify the effectiveness of 2DPCA-PSSM. Furthermore, we also made comparisons with SVM and other cutting-edge approaches, and the results indicated the excellent of RF-PSSM. Finally, we adopted RF-PSSM to find some potential proteins that may interact with E1, which may provide guidance for future wet experiments.

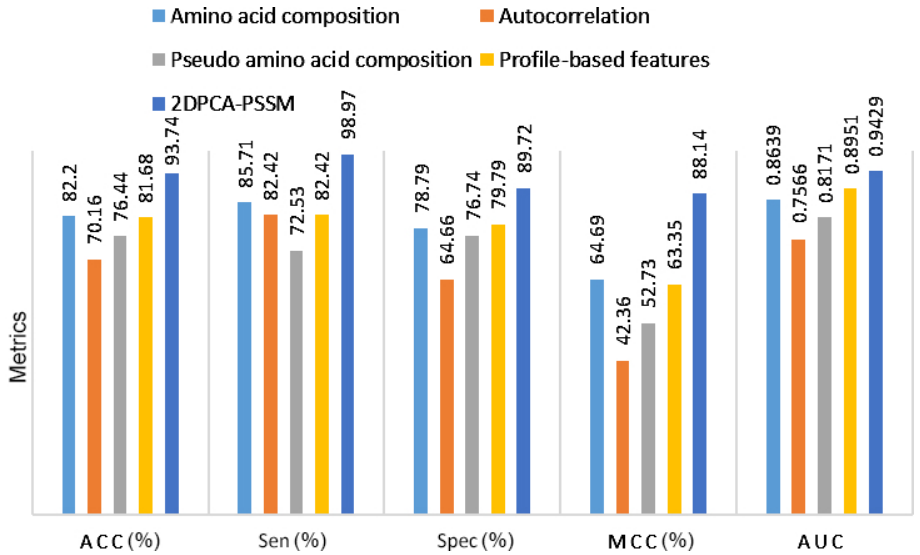


Fig. 6 Performance of four features compared with 2DPCA-PSSM on independent test set.

The dataset and code are available at https://github.com/flyinsky6/RF-PSSM.

Of course, the lack of developing special software or website for the algorithm proposed in this paper is the shortcoming of this paper. Furthermore, we will focus on improve the performance of the prediction of HCV-host interaction field in the following areas, such as adding the network structure features of HCV-host interaction through graph neural network technology[49], adding protein structure features, physicochemical properties and so on.

Electronic Supplementary Material

Supplementary materials including

- HCV-human PPIs dataset, and
- potential human proteins that may interact with E1

are available in the online version of this article at https://doi.org/10.26599/BDMA.2022.9020031.

References

[1] A. E. Jordan, D. C. Perlman, J. Reed, D. J. Smith, and H. Hagan, Patterns and gaps identified in a systematic review of the hepatitis C virus care continuum in studies among people who use drugs, *Front. Public Health*, vol. 5, p. 348, 2017.

[2] R. Rashti, S. M. Alavian, Y. Moradi, H. Sharafi, A. Mohamadi Bolbanabad, D. Roshani, and G. Moradi, Global prevalence of HCV and/or HBV coinfections among people who inject drugs and female sex workers who live with HIV/AIDS: A systematic review and meta-analysis, *Arch. Virol.*, vol. 165, no. 9, pp. 1947–1958, 2020.

[3] R. Ansumana, S. Keitell, G. M. T. Roberts, F. Ntoumi, E. Petersen, G. Ippolito, and A. Zumla, Impact of infectious disease epidemics on tuberculosis diagnostic, management, and prevention services: Experiences and lessons from the 2014–2015 Ebola virus disease outbreak in West Africa, *Int. J. Infect. Dis.*, vol. 56, pp. 101–104, 2017.

[4] R. V. Thurber, J. P. Payet, A. R. Thurber, and A. M. S. Correa, Virus-host interactions and their roles in coral reef health and disease, *Nat. Rev. Microbiol.*, vol. 15, no. 4, pp. 205–216, 2017.

[5] A. F. Brito and J. W. Pinney, Protein-protein interactions in virus-host systems, *Front. Microbiol.*, vol. 8, p. 1557, 2017.

[6] C. K. Lai, K. S. Jeng, K. Machida, and M. M. C. Lai, Association of hepatitis C virus replication complexes with microtubules and actin filaments is dependent on the interaction of NS3 and NS5A, *J. Virol.*, vol. 82, no. 17, pp. 8838–8848, 2008.

[7] M. A. Ansari, V. Pedergnana, C. L. C. Ip, A. Magri, A. Von Delft, D. Bonsall, N. Chaturvedi, I. Bartha, D. Smith, G. Nicholson, et al., Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus, *Nat. Genet.*, vol. 49, no. 5, pp. 666–673, 2017.

[8] M. Dimitrova, I. Imbert, M. P. Kieny, and C. Schuster, Protein-protein interactions between hepatitis C virus nonstructural proteins, *J. Virol.*, vol. 77, no. 9, pp. 5401–5414, 2003.

[9] M. Irshad, P. Gupta, and K. Irshad, Molecular basis of hepatocellular carcinoma induced by hepatitis C virus infection, *World J. Hepatol.*, vol. 9, no. 36, pp. 1305–1314, 2017.

[10] F. E. Eid, M. ElHefnawi, and L. S. Heath, DeNovo: Virus-host sequence-based protein-protein interaction prediction, *Bioinformatics*, vol. 32, no. 8, pp. 1144–1150, 2016.

[11] A. Zhang, L. He, and Y. Wang, Prediction of GCRV virus-host protein interactome based on structural motif-domain interactions, *BMC Bioinformatics*, vol. 18, no. 1, p. 145, 2017.

[12] Z. H. You, K. C. C. Chan, and P. Hu, Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest, *PLoS One*, vol. 10, no. 5, p. e0125811, 2015.

[13] J. Zahiri, O. Yaghoubi, M. Mohammad-Noori, R. Ebrahimpour, and A. Masoudi-Nejad, PPIevo: Protein-protein interaction prediction from PSSM based evolutionary information, *Genomics*, vol. 102, no. 4, pp. 237–242, 2013.

[14] S. Mika and B. Rost, Protein-protein interactions more conserved within species than across species, *PLoS Comput. Biol.*, vol. 2, no. 7, p. e79, 2006.

[15] G. Cui, C. Fang, and K. Han, Prediction of protein-protein interactions between viruses and human by an SVM model, *BMC Bioinformatics*, vol. 13, no. 7S, p. S5, 2012.

[16] A. Emamjomeh, B. Goliaei, J. Zahiri, and R. Ebrahimpour, Predicting protein-protein interactions between human and hepatitis C virus via an ensemble learning method, *Mol. Biosyst.*, vol. 10, no. 12, pp. 3147–3154, 2014.

[17] R. K. Barman, S. Saha, and S. Das, Prediction of interactions between viral and host proteins using supervised machine learning methods, *PLoS One*, vol. 9, no. 11, p. e112034, 2014.

[18] B. Kim, S. Alguwaizani, X. Zhou, D. S. Huang, B. Park, and K. Han, An improved method for predicting interactions between virus and human proteins, *J. Bioinform. Comput. Biol.*, vol. 15, no. 1, p. 1650024, 2017.

[19] S. Alguwaizani, B. Park, X. Zhou, D. S. Huang, and K. Han, Predicting interactions between virus and host proteins using repeat patterns and composition of amino acids, *J. Healthc. Eng.*, vol. 2018, p. 1391265, 2018.

[20] P. Domingos, The role of Occam's razor in knowledge discovery, *Data Min. Knowl. Discov.*, vol. 3, no. 4, pp. 409–425, 1999.

[21] J. Wang, B. Yang, J. Revote, A. Leier, T. T. Marquez-Lago, G. Webb, J. Song, K. C. Chou, and T. Lithgow, POSSUM: A bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles, *Bioinformatics*, vol. 33, no. 17, pp. 2756–2758, 2017.

[22] Y. Wang, Y. Ding, F. Guo, L. Wei, and J. Tang, Improved detection of DNA-binding proteins via compression technology on PSSM information, *PLoS One*, vol. 12, no. 9, p. e0185587, 2017.

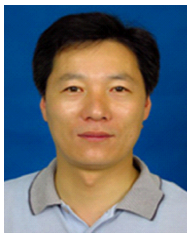
[23] Z. Li, P. Han, Z. H. You, X. Li, Y. Zhang, H. Yu, R. Nie, and X. Chen, In silico prediction of drug-target interaction networks based on drug chemical structure and protein sequences, *Sci. Rep.*, vol. 7, no. 1, p. 11174, 2017.

[24] C. Huang and J. Yuan, Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites, *Biosystems*, vol. 113, no. 1, pp. 50–57, 2013.
[25] S. Ding, Y. Li, Z. Shi, and S. Yan, A protein structural classes prediction method based on predicted secondary structure and PSI-BLAST profile, *Biochimie*, vol. 97, pp. 60–65, 2014.
[26] Y. O. J. Hong, S. V. Chintapalli, K. D. Ko, G. Bhardwaj, Z. Zhang, D. Van Rossum, and R. L. Patterson, Predicting protein folds with fold-specific PSSM libraries, *PLoS One*, vol. 6, no. 6, p. e20557, 2011.
[27] P. R. Wills, D. J. Scott, and D. J. Winzor, The osmotic second virial coefficient for protein self-interaction: Use and misuse to describe thermodynamic nonideality, *Anal. Biochem.*, vol. 490, pp. 55–65, 2015.
[28] T. Guirimand, S. Delmotte, and V. Navratil, VirHostNet 2.0: Surfing on the web of virus/host molecular interactions data, *Nucleic Acids Res*., vol. 43, pp. D583–D587, 2015.
[29] N. Q. Khanh Le, Q. H. Nguyen, X. Chen, S. Rahardja, and B. P. Nguyen, Classification of adaptor proteins using recurrent neural networks and PSSM profiles, *BMC Genomics*, vol. 20, p. 966, 2019.
[30] S. F. Altschul and E. V. Koonin, Iterated profile searches with PSI-BLAST–A tool for discovery in protein databases, *Trends Biochem. Sci.*, vol. 23, no. 11, pp. 444–447, 1998.
[31] P. J. A. Cock, J. M. Chilton, B. Grüning, J. E. Johnson, and N. Soranzo, NCBI BLAST+ integrated into Galaxy, *GigaScience*, vol. 4, p. 39, 2015.
[32] Y. A. Huang, Z. H. You, X. Gao, L. Wong, and L. Wang, Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence, *Biomed Res. Int.*, vol. 2015, p. 902198, 2015.
[33] J. Yang, D. Zhang, A. F. Frangi, and J. Y. Yang, Two-dimensional PCA: A new approach to appearance-based face representation and recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131–137, 2004.
[34] J. J. Hu, G. Z. Tan, F. G. Luan, and A. S. M. Libda, 2DPCA versus PCA for face recognition, *J. Cent. South Univ.*, vol. 22, no. 5, pp. 1809–1816, 2015.
[35] J. Yang and J. Y. Yang, From image vector to matrix: A straightforward image projection technique-IMPCA vs. PCA, *Pattern Recogn.*, vol. 35, no. 9, pp. 1997–1999, 2002.
[36] Z. Li, R. Nie, Z. You, C. Cao, and J. Li, Using discriminative vector machine model with 2DPCA to predict interactions among proteins, *BMC Bioinformatics*, vol. 20, p. 694, 2019.
[37] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, Rotation forest: A new classifier ensemble method, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1619–1630, 2006.
[38] K. H. Liu and D. S. Huang, Cancer classification using Rotation Forest, *Comput. Biol. Med.*, vol. 38, no. 5, pp. 601–610, 2008.
[39] L. Wong, Z. H. You, Z. Ming, J. Li, X. Chen, and Y. A. Huang, Detection of interactions between proteins through rotation forest and local phase quantization descriptors, *Int. J. Mol. Sci.*, vol. 17, no. 1, p. 1, 2016.
[40] W. S. Noble, What is a support vector machine? *Nat. Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, 2006.
[41] K. Duan, S. S. Keerthi, and A. N. Poo, Evaluation of simple performance measures for tuning SVM hyperparameters, *Neurocomputing*, vol. 51, pp. 41–59, 2003.
[42] M. Mousavizadegan and H. Mohabatkar, Computational prediction of antifungal peptides via Chou's PseAAC and SVM, *J. Bioinform. Comput. Biol.*, vol. 16, no. 4, p. 1850016, 2018.
[43] J. Zhou, L. Li, L. Wang, X. Li, H. Xing, and L. Cheng, Establishment of a SVM classifier to predict recurrence of ovarian cancer, *Mol. Med. Rep.*, vol. 18, no. 4, pp. 3589–3598, 2018.
[44] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, Applications of support vector machine (SVM) learning in cancer genomics, *Cancer Genomics Proteomics*, vol. 15, no. 1, pp. 41–51, 2018.
[45] Y. Ge, S. Zhao, and X. Zhao, A step-by-step classification algorithm of protein secondary structures based on double-layer SVM model, *Genomics*, vol. 112, no. 2, pp. 1941–1946, 2020.
[46] C. C. Chang and C. J. Lin, LIBSVM: A library for support vector machines, *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011.
[47] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, A survey on ensemble learning, *Front. Comput. Sci.*, vol. 14, no. 2, pp. 241–258, 2020.
[48] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K. C. Chou, Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences, *Nucleic Acids Res.*, vol. 43, no. W1, pp. W65–W71, 2015.
[49] T. Zhong, Z. Li, Z. H. You, R. Nie, and H. Zhao, Predicting miRNA-disease associations based on graph random propagation network and attention network, *Brief Bioinform.*, vol. 23, no. 2, p. bbab589, 2022.

Xin Liu received the PhD and MS degrees from China University of Mining and Technology in 2016 and 2006, respectively. She is an associate professor and works at the Department of Intelligent Medical Engineering, School of Medical Informatics and Engineering, Xuzhou Medical University. Her research interests include bioinformatics data mining, machine learning, and protein representation learning.

Yaping Lu obtained the BS and MS degrees in computer science and technology from China University of Mining and Technology in 2003 and 2006, respectively. At present, she is working in the laboratory centre of the School of Humanities and Arts of China University of Mining and Technology. Her research interests mainly include virtual reality, mobile agents, big data analysis, and intelligent processing.

Liang Wang received the PhD degree from University of Western Australia and worked as a postdoctoral fellow in both Canada and Australia for several years. He is currently a full-time principal investigator (PI) at Laboratory Medicine in Guangdong Provincial People's Hospital and Guangdong Academy of Medical Sciences, Guangzhou, China. He has published more than 70 peer-reviewed journal articles and his research interests focus on bioinformatics, infectious diseases, microbiology, and biomacromolecules.



Wei Geng received the MS degree from China University of Mining and Technology in 2009. He is currently a senior experimentalist at the School of Modical Informatics and Engineering, Xuzhou Medical University. He has published 10 papers and 2 books (medical information technology, overview of hospital information system). His current research interest is medical informatics.

Xinyi Shi is persuing the MS degree in bioinformatics at Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences. Her research interests include biological information and data mining.

Xiao Zhang received the PhD degree from University of Wisconsin-Milwaukee, WI, USA. He is a professor and the dean at the School of Medical Informatics and Engineering, Xuzhou Medical University, China. His research interests include precision medicine, artificial intelligence, and machine learning in healthcare and medical field.