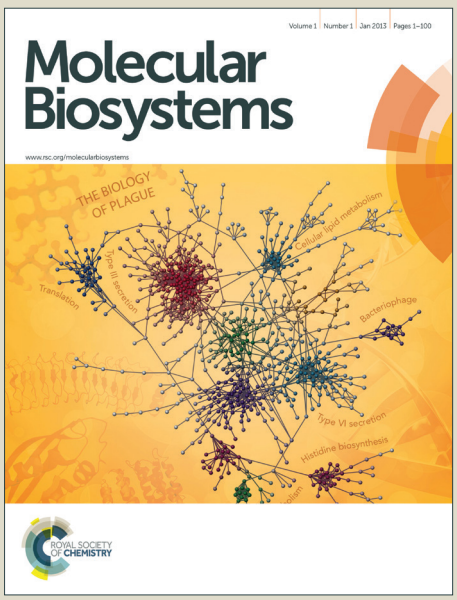
Molecular BioSystems

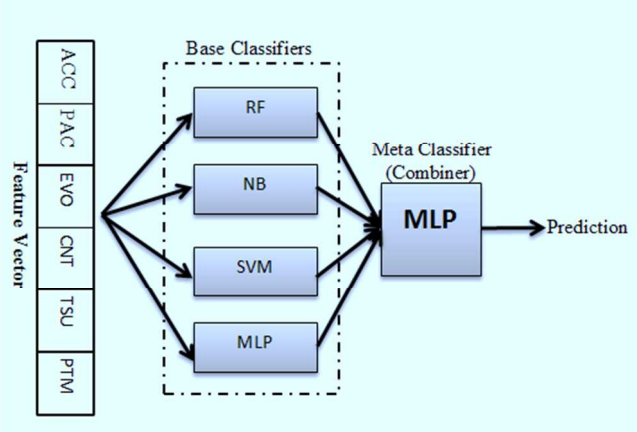
Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Graphical abstract



The highlighting the novelty of the work:

We developed novel method to predict human-HCV protein-protein interaction that is the most comprehensive study of this type.

Molecular BioSystems

RSCPublishing

PAPER

Predicting of protein–protein interactions between human and hepatitis C virus via an ensemble learning method

Cite this: DOI: 10.1039/x0xx00000x

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

Abbasali Emamjomeh,[a] Bahram Goliaei,*[a] Javad Zahiri[ba] and Reza Ebrahim-pour[c],

An estimated 170 million people, approximately 3% of the world population, are chronically infected with the Hepatitis C virus (HCV). More than 350,000 deaths are reported annually, which are caused by HCV. HCV, similar to a variety of viruses, causes disease in humans by altering protein–protein interactions within the host cells. Experimental approaches for the detection of host–virus PPIs have many inherent limitations. Computational approaches to predict these interactions are therefore of significant importance. While many studies have been developed to predict intra-species PPIs in the last decade, predictions of inter-species PPIs such as human-HCV PPIs are rare. In this study, we developed an ensemble learning method to predict PPIs between human and HCV. Our model utilises four well-established diverse learners as base classifiers including random forest (RF), Naïve Bayes (NB), support vector machine (SVM) and multilayer perceptron (MLP). In addition, an MLP was used as a meta-learner to combine base learners' predictions to produce the final prediction. To encode human and HCV proteins as feature vectors, we used six different descriptors as follows: amino acid composition (ACC), pseudo amino acid composition (PAC), evolutionary information feature, network centrality measures, tissue information and post-translational modification information. To assess the prediction power of the proposed method, we assembled a benchmark dataset composed of confident positive and negative PPIs. In a 10-fold cross-validation experiment, our prediction method achieved an accuracy and a specificity as high as 83% and 94%, respectively. Further, in an independent test set the proposed method achieved an accuracy of 84% and a specificity of 92%. When compared with the existing method, our method showed a better performance. These results revealed that our method is suitable for performing PPI prediction in a host–pathogen context.

Introduction

Hepatitis C virus (HCV), an enveloped positive stranded RNA virus, is a major cause of liver disease in chronically infected individuals.[1] An estimated 170 million people are chronically infected with the HCV and more than one million new infection cases and about 350,000 deaths are reported annually.[2-5] Virus proteins manipulate host cell machinery by competing with host proteins in the host PPIs.[6-8] Considering the limitations of experimental approaches for detection of PPIs in the host–pathogen context,[9,10] computational methods for predicting these interactions are of significant importance. Computational PPI prediction approaches provide opportunities for identifying specific targets for further experimental work,[1] especially for drug design and developing more effective treatments.[11-13]

While many studies have been performed on intra-species PPI prediction in the last decade,[14] predictions of inter-species PPI are rare,[1] and are especially so for host–pathogen PPIs like human-HCV. Existing methods for host–pathogen PPI prediction can be roughly categorised into four different divisions:[13] homology-based methods;[15,16] methods based on structural information;[9,17] using sequence information[6,18] and machine learning-based methods.[19-21]

In this study, we propose an ensemble learning method, which exploits six different features to predict PPIs between humans and HCV, including: 1, amino acid composition (ACC); 2, pseudo amino acid composition (PAC); 3, evolutionary information feature; 4, network centrality measures; 5, tissue information feature and 6, post-translational modification

information. Our method achieved an accuracy of 0.83 in a cross validation analysis and 0.84 in an independent test set.

Material and methods

Benchmark dataset

To assess the proposed method, two datasets were constructed: Positive dataset and negative dataset.

Positive interactions

For constructing positive human-HCV PPIs, we extracted all hepatitis C virus interactions from the IntAct database.[22] Then, interactions between HCV and other organisms except human were removed. Finally, those interactions that annotated as 'physical association' or 'direct interaction' were considered as positive interaction set (PS). PS contained 657 interactions.

Negative interactions

Selecting appropriate negative PPIs is very challenging in PPI prediction problem.[23,24] To select confident negative PPIs, we paired all human proteins with all HCV proteins that had been examined in a specific experiment (according to PubMed IDs of publications), and then all protein pairs that were not reported as an interaction in the PS were considered as negative interactions (Figure 1).

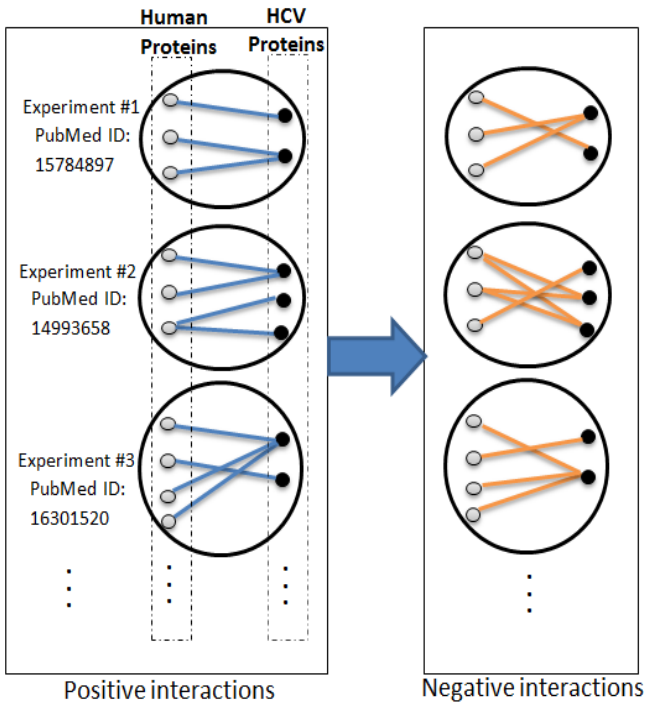


Figure 1: Schematic view of constructing positive and negative interactions.

Obtained negative interactions of all the experiments constructed the negative interaction set (NS); the total number of negative interactions was 2910. Last, PS and NS constructed our benchmark dataset.

Encoding proteins as feature vectors

We used different features to encode human and HCV proteins as feature vectors. These features included: amino acid composition (ACC), pseudo amino acid composition (PAC), evolutionary information feature, network centrality feature, tissue information feature and post-translational modification feature. These features have been described below.

Amino acid composition

For computing amino acid composition (ACC), we clustered twenty naïve amino acids into eight clusters, C_1–C_8, regarding 544 physicochemical and biochemical properties of amino acids that exist in the AAindex database.[25] So, each protein was coded as an 8-dimentional feature vector. The details of the clustering procedure described below.

Firstly, to prevent outweighing indices with larger magnitude over indices with smaller magnitude, the values of each index were standardised according to the mean and variance of the index. Then, k-means method was used as the clustering algorithm by varying the number of clusters from three to nine. To find an appropriate clustering, the k-means was run 1000 times for each cluster number, and the Dunn index[26] was used to measure the validity of the obtained clustering.

Suppose $C = \{c_1, c_2, \ldots c_N\}$ shows a clustering that consists of $c_1 \ldots c_N$ as clusters, then the Dunn index for this clustering is computed as follows:

$$1.\quad DunnIndex(C) = \min_{1\le i\le N}\left\{\min_{i+1\le j\le N}\left(\frac{dist(c_i,c_j)}{\max_{1\le l\le N}(diam(c_l))}\right)\right\}.$$

Where, $dist(c_i, c_j)$ is the distance between clusters c_i and c_j, and $diam(c_l)$ is diameter of the cluster c_l:

$$2.\quad dist(c_i,c_j) = \min_{x\in c_i, y\in c_j}\{dist(x,y)\}$$

And

$$3.\quad diam(c_i) = \max_{x,y\in c_i}\{dist(x,y)\}$$

According to the values of the Dunn index, the best clustering is as follow:

C_1= {A, E}; C_2={R, Q, K, H}; C_3= {N, D, S, T}; C_4= {G}; C_5= {P}; C_6= {I, L, M, F, V}; C_7= {W, Y}; C_8= {C}

Pseudo amino acid composition

The concept of pseudo amino acid composition (PAC) was originally introduced in order to predict protein subcellular localisation and membrane protein type.[27] Despite the ACC, PAC uses the sequence-order information (for precise definition of what PAC refers to).[28]

The PAC has been used in various protein annotation predictions such as post-translational modification,[29] secondary structure prediction,[30] protein folding rates prediction,[31] and identifying bacterial secreted proteins.[32] We used the PseAAC webserver[33] to compute two types of PAC features for each protein.

Evolutionary information

We used recently proposed evolutionary features by Zahiri *et al.*[34] that have been extracted from Position-Specific Scoring Matrices (PSSM) and have been successfully exploited for human PPI prediction. The PSSM matrix is composed of L × 20 entries, where L is the length of the protein of interest. The rows and columns of the matrix are indexed by the protein residues and the 20 naive amino acids, respectively, and entries of the matrix represent the log odds scores of the occurrences of different amino acids at the different positions of the protein sequence.

To compute PSSMs, we used the Position Specific Iterated BLAST (PSI-BLAST) tool,[35] with three iterations and the e-value of 0.0001, against the NCBI non-redundant dataset on a local machine for human proteins.

NETWORK CENTRALITY FEATURE

It has been shown that centrality of host proteins are important for being targeted by pathogen proteins.[1] In this study, different centrality measures for human proteins have been computed using NetworkAnalyzer,[36] which is a cytoscape[37] plugin. We used the BioGrid[38] database to construct a human protein–protein interaction network (PPIN) for computing centrality measures. Topological measures that were computed for each protein, denoted by P, in human PPIN are briefly described below.

Degree: the degree (or connectivity) of a protein is the number of its interacting partners. *Neighborhood connectivity:* the average degree (discarding self-interactions) of all neighbours of P. *Average shortest path length*: the average length of all shortest paths from P to any other protein. *Stress*: the number of shortest paths between all protein pairs in the human PPIN that pass through a particular protein. This centrality represents the workload the protein carries in a graph.[39] *Eccentricity*: the greatest distance between P and any other protein in the human PPIN. *Closeness*: the reciprocal of the total distance from P to all the other proteins in the human PPIN. *Betweenness*: this centrality is a semi-normalised version of the stress centrality[39] and addresses the limitations of some classical centralities.[40] Betweenness for a protein P is the ratio of the number of shortest paths passing through P to the number of all shortest paths between all protein pairs in the human PPIN. *Radiality*: Computed by subtracting the average shortest path length of P from the length of the longest path of the connected component, plus 1. Finally, for normalisation the radiality of each protein is divided by the length of the longest path of the connected component. *Clustering coefficient*: the ratio of the number of edges between the neighbours of P to the number of edges that could possibly exist among them. The clustering coefficient measures the density in the local region of a protein.

In addition to the above-mentioned topological features, all computed for human proteins, we also considered the number of interactions for each HCV protein as a new feature. The numbers of interactions for HCV proteins were extracted from the IntAct database using uniprot IDs of proteins.

TISSUE INFORMATION FEATURE

Considering the importance of tissue information in human-pathogens PPI,[41] we used 582 different tissue terms of human proteins that were mentioned in the HPRD database.[42] Each human protein was coded as a 582-dimentional binary feature vector; each element of this vector shows whether the protein expressed is in a specific tissue or not.

POST-TRANSLATIONAL MODIFICATION

There is increasing evidence that shows post-translational modifications (PTM) are crucial for the control of protein functions and specially affect the PPIs.[43,44] Considering the importance of PTMs, we used 31 PTM types (e.g. deacetylation, phosphorylation, glycosylation and others) that were mentioned in the HPRD database to represent human proteins. Each 20 amino acids may undergo 31 PTM types, so each human protein was represented as a 620-length binary feature vector; each element of this vector, for a protein of interest, shows whether a specific amino acid is modified with a specific PTM type or not.

PREDICTION ALGORITHM

It is shown that in challenging prediction problems, combining diverse classifiers (ensemble learning) can lead to better performance.[45,46] In this study, we used an ensemble learning method to predict human-HCV PPIs. Our base classifiers are random forest (RF), Naïve Bayes (NB), support vector machine (SVM) and multilayer perceptron (MLP); these classifiers had successful applications in similar problems.[13,14] Here, we used stacking,[47] also called stacked generalisation, to combine base classifiers in order to predict human-HCV PPIs (Figure 2).

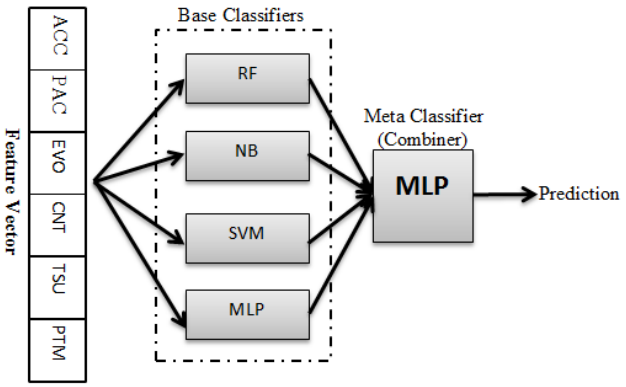


Figure 2: The ensemble learning method that was used for predicting protein–protein interactions between human and Hepatitis C virus. A stacking method was used to combine four diverse base classifiers including: random forest (RF), Naïve Bayes (NB), support vector machine (SVM) and multilayer perceptron (MLP). Input feature vector contains six different feature types for each protein pair: amino acid composition (ACC), pseudo amino acid composition (PAC), evolutionary information (EVO), network centrality measures (CNT), tissue information feature (TSU) and post-translational modification information (PTM).

In stacking, the output of the base classifiers was considered as input of a meta-learner. The meta-learner tries to learn how best to combine the base classifiers' outputs to produce the final output. Herein, an MLP was used as meta-learner.

PERFORMANCE ASSESSMENT MEASURES

After construction of a benchmark dataset, the prediction performance of human-HCV PPI can be evaluated in terms of

different measures based on four basic parameters: TP (True Positive) that denotes the number of interacting proteins correctly predicted, TN (True Negative) that is the number of non-interacting proteins correctly predicted, FP (False Positive) that denotes the number of non-interacting proteins incorrectly predicted as interacting proteins and FN (False Negative) that is the number of interacting proteins incorrectly predicted as non-interacting proteins. Having these basic parameters, we used precision, recall, accuracy and area under the ROC curve (AUC) to assess the performance of the proposed model. Recall is a fraction of real interactions correctly identified by the model; precision is a fraction of interaction predictions that are correct and the accuracy is the proportion of correct predictions. The area under the ROC curve, which plots sensitivity vs. one minus the specificity, is an important measure of prediction performance.

4. $Sensitivity = \frac{TP}{TP + FN}$

5. $Specificity = \frac{TN}{TN + FP}$

6. $Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$

Results and discussion

To estimate the performance of the proposed prediction method, we used a 10-fold cross validation and an independent test. For cross-validation analysis, the PPIs that reported after 2012 have been excluded from the benchmark dataset to be used as an independent test set.

Cross-validation analysis

In the 10-fold cross validation, the dataset is randomly partitioned into 10 equal sets, out of which nine sets are used for training and the remaining one for testing. The subsets were rotated 10 times, such that each subset was used for both training and testing, and each protein pair was used for testing exactly once. The final prediction result is the average of the 10 testing sets. Table 1 shows the performance of our method in predicting human-HCV PPIs: sensitivity, specificity and accuracy are 0.79, 0.94 and 0.83, respectively.

Independent test set

We used those PPIs from the benchmark dataset that reported after 2012 as an independent test set. This dataset contained 139 interacting and 555 non-interacting protein pairs.
The performance of our model in the 10-fold cross-validation procedure is shown in Table 1.
As we can see, the proposed method achieved a sensitivity of 0.73, a specificity of 0.94 and an accuracy of 0.84 in the independent test set.

Table 1: Prediction performance of the proposed method in 10-fold cross validation and independent test set and details about used data sets.

	Sensitivity	Specificity	Accuracy	Interacting protein pairs	Non-interacting protein pairs
Cross-validation	0.79	0.94	0.83	518	2355
Independent test	0.73	0.92	0.84	139	555

Comparison with the current state-of-the-art methods

Computational prediction of PPIs between human and HCV is a new problem in bioinformatics and computational biology. Recently, Cui *et al.*[21] proposed a prediction method using an SVM model to address the problem. Cui *et al.* used three different pairs of train-test set to evaluate the performance of the method.
To compare with Cui's method, we also run our method using their train-test sets. As shown in Table 2, our method outperformed Cui's method considering all performance measures.

Table 2: Comparison of our method with the method previously proposed[21] for predicting protein–protein interaction between human Hepatitis C virus.

	Sensitivity	Specificity	Accuracy
Our method	0.84	0.87	0.83
Cui's method	0.78	0.85	0.81

Table 3: Feature importance: to compute the importance of the different features, we removed each feature type in turn and then computed the accuracy of the proposed prediction model. The highest loss of accuracy shows the most important feature.

Feature	Accuracy of our method when removing the feature	Loss of accuracy
ACC	0.69	0.14
PAC	0.76	0.07
EVO	0.74	0.09
CNT	0.75	0.08
TSU	0.78	0.05
PTM	0.72	0.11

Feature importance

In this study, we used heterogeneous descriptors for proteins to predict PPIs between human and HCV. To compute the contribution of the different descriptors in prediction, we removed each feature type in turn and then computed the accuracy of the proposed prediction model (Table 3); the higher the loss of accuracy, the more important the feature.

Accordingly, we can say amino acid composition and post-translational modification are the most important features in predicting PPIs between human and HCV.

Table 4: Enriched pathways and GO (Gene Ontology) terms in the set of interacting human proteins with HCV (a term was considered significantly enriched if the Benjamini corrected P-value was less than 0.005).

	Type of data	Enriched feature	Benjamini corrected P-value
GO [48]	Biological process	multi-organism process (GO:0051704)	1.3E-12
		cellular component organization (GO:0016043)	2.5E-08
		cellular process (GO:0009987)	7.2E-08
		biological regulation (GO:0065007)	3.5E-06
		death (GO:0016265)	1.7E-05
		biological adhesion (GO:0022610)	7.9E-05
	Molecular function	binding (GO:0005488)	1.2E-10
		structural molecule activity (GO:0005198)	1.3E-03
		enzyme regulator activity (GO:0030234)	1.2E-03
	Cellular component	organelle part (GO:0044422)	3.6E-07
		organelle (GO:0043226)	2.6E-07
		extracellular region part (GO:0044421)	1.9E-06
		macromolecular complex (GO:0032991)	3.7E-06
		membrane-enclosed lumen (GO:0031974)	2.6E-04
Pathway	KEGG pathway [49]	Focal adhesion (hsa04510)	4.5E-05
		Pathways in cancer (hsa05200)	3.7E-03
	REACTOME pathway [50]	Hemostasis (REACT_604)	2.3E-04
		Integrin cell surface interactions (REACT_13552)	3.1E-04
		Apoptosis (REACT_578)	3.3E-03

Individual classifier's contribution to the whole ensemble

The most accurate individual classifiers are not necessarily the most important contributors to the ensemble performance.[51] Therefore, to evaluate each individual classifier's contribution to the whole ensemble, we removed each classifier in turn and then computed the accuracy of the ensemble model. The loss of accuracy was 0.09, 0.07, 0.04 and 0.03 for MLP, RF, SVM and NB, respectively. According to the results, the individual classifiers which made most contribution to the ensemble, in order, are: MLP, RF, SVM and NB.

Enrichment analysis

We used "Database for Annotation, Visualization and Integrated Discovery (DAVID)"[52] for enrichment analysis in the set of interacting human proteins with HCV. Table 4 shows Gene Ontology (GO) terms and pathways that significantly enriched (a term was considered significantly enriched if the Benjamini corrected P-value[53] was less than 0.005). The analysis highlights GO terms and pathways, such "as binding (GO:0005488)"; "extracellular region part (GO:0044421)" and "Pathways in cancer (hsa05200)", which have been reported previously as important features of the proteins that interact with the HCV.[8,54] Furthermore, table 5 shows the enriched domains in the set of predicted interacting human proteins with HCV. Some of the enriched domains, such as "Spectrin repeat", have been reported in experimental studies.[8] In addition, the results highlight some new features of human proteins that significantly enriched and can be used in the future experimental and computational studies [Supplementary file.xlsx].

Table 5: Enriched domains in the set of interacting human proteins with HCV (a term was considered significantly enriched if the Benjamini corrected P-value was less than 0.005).

Database name	Enriched feature	Benjamini corrected P-value
SMART [55]	EGF_CA (SM00179)	3.1E-11
	EGF (SM00181)	1.6E-08
	SH3 (SM00326)	9.1E-04
	SPEC (SM00150)	2.1E-03
BLOCKS [56]	EGF-like calcium-binding (IPB001881)	1.4E-04
	Spectrin repeat (IPB002017)	9.0E-04
	Aspartic acid and asparagine hydroxylation site (IPB000152)	1.9E-03
PFAM [57]	EGF_CA (PF07645)	1.1E-08
	EGF (PF00008)	1.9E-07
	SH3_1 (PF00018)	8.6E-04
	Intermediate filament head (PF04732)	2.4E-03
	Spectrin (PF00435)	3.9E-03
	Filament_head (PF04732)	3.6E-03
PROSITE [58]	EGF_CA (PS01187)	2.6E-11
	Calcium-binding EGF-like domain signature (PS01187)	1.5E-11
	ASX_HYDROXYL (PS00010)	1.1E-11
	Aspartic acid and asparagine hydroxylation site (PS00010)	2.1E-11
	EGF_1 (PS00022)	3.3E-11
	EGF_2 (PS01186)	3.3E-11
	EGF-like domain signature 2 (PS01186)	9.7E-10
	EGF-like domain profile (PS50026)	1.4E-09
	EGF-like domain signature 1 (PS00022)	4.5E-08
	Src homology 3 (SH3) domain profile (PS50002)	5.6E-04

Conclusions

In this study, we have proposed an ensemble learning method in order to predict PPI between human and hepatitis C virus

(HCV). Six different descriptors include: amino acid composition, pseudo amino acid composition, evolutionary information, network centrality measures, tissue information feature and post-translational modification information have been used to encode protein pairs. Four diverse classifiers: random forest, Naïve Bayes, support vector machine (SVM), and multilayer perceptron (MLP) are used as base classifiers. An MLP was used as meta-learner to combine the base classifiers' predictions in order to produce final output. The results show the satisfactory prediction performance of our method, the method achieved an accuracy of 0.83, a specificity of 0.94, and a sensitivity of 0.79 in a 10-fold cross validation analysis on our benchmark dataset. In addition, in an independent test set achieved an accuracy of 0.84, a specificity of 0.92 and a sensitivity of 0.73. The results revealed the better performance of the proposed method in comparison with the current state-of-the-art method. According to our analysis, amino acid composition and post-translational modification contribute most in predicting PPIs between human and HCV. The proposed method can be extended to predict other host–pathogen PPIs.

Notes and references

[a] Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran.
[b] Department of Mathematics, K.N. Toosi University of Technology, Tehran, Iran.
[c] Brain and Intelligent Systems Research Lab, Department of Electrical and Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.
*Corresponding author: Enghelab Square, University of Tehran, Institute of Biochemistry and Biophysics (IBB), Postal code:1417614411, Tehran, Iran. Tel /fax:+ 98 21 66498672. E-mail address: goliaei@ibb.ut.ac.ir.
See DOI: 10.1039/b000000x/

1 McDermott JE, Diamond DL, Corley C, Rasmussen AL, Katze MG, Waters KM. Topological analysis of protein co-abundance networks identifies novel host targets important for HCV infection and pathogenesis. BMC systems biology. 2012;6:28.
2 Wilkins T, Malcolm JK, Raina D, Schade RR. Hepatitis C: diagnosis and treatment. American Family Physician. 2010;81:1351-7.
3 De Chassey B, Navratil V, Tafforeau L, Hiet M, Aublin-Gex A, Agaugue S, *et al*. Hepatitis C virus infection protein network. Molecular Systems Biology. 2008;4.
4 Nishiya AS, de Almeida-Neto C, Ferreira SC, Alencar CS, Di-Lorenzo-Oliveira C, Levi JE, *et al*. HCV Genotypes, Characterization of Mutations Conferring Drug Resistance to Protease Inhibitors, and Risk Factors among Blood Donors in São Paulo, Brazil. PloS one. 2014;9:e86413.
5 Fan X, Xue B, Dolan PT, LaCount DJ, Kurgan L, Uversky VN. The intrinsic disorder status of the human hepatitis C virus proteome. Molecular BioSystems. 2014;10:1345-63.
6 Evans P, Dampier W, Ungar L, Tozeren A. Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. BMC medical genomics. 2009;2:27.
7 Tournier J-N, Quesnel-Hellmann A. Host–Pathogen Interactions: A Biological Rendez-Vous of the Infectious Nonself and Danger Models? PLoS Pathogens. 2006;2:e44.
8 Dolan PT, Zhang C, Khadka S, Arumugaswami V, Vangeloff AD, Heaton NS, *et al*. Identification and comparative analysis of hepatitis C virus–host cell protein interactions. Molecular BioSystems. 2013;9:3199-209.
9 Doolittle JM, Gomez SM. Mapping protein interactions between Dengue virus and its human and insect hosts. PLoS neglected tropical diseases. 2011;5:e954.
10 Zhao C, Sacan A. Prediction of HIV-1 and human protein interactions based on a novel evolution-aware structure alignment method.
11 Arkin MR, Wells JA. Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. Nature reviews Drug discovery. 2004;3:301-17.
12 Mukhopadhyay A, Ray S, Maulik U. Incorporating the type and direction information in predicting novel regulatory interactions between HIV-1 and human proteins using a biclustering approach. BMC bioinformatics. 2014;15:26.
13 Zhou H, Jin J, Wong L. Progress in computational studies of host-pathogen interactions. Journal of bioinformatics and computational biology. 2013;11.
14 Zahiri J, Hannon Bozorgmehr J, Masoudi-Nejad A. Computational Prediction of Protein–Protein Interaction Networks: Algorithms and Resources. Current genomics. 2013;14:397-414.
15 Krishnadev O, Srinivasan N. Prediction of protein–protein interactions between human host and a pathogen and its application to three pathogenic bacteria. International journal of biological macromolecules. 2011;48:613-9.
16 Wuchty S. Computational prediction of host-parasite protein interactions between P. falciparum and H. sapiens. PloS one. 2011;6:e26960.
17 Doolittle JM, Gomez SM. Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens. Virology journal. 2010;7:82.
18 Dyer MD, Murali T, Sobral BW. Computational prediction of host-pathogen protein–protein interactions. Bioinformatics. 2007;23:i159-i66.
19 Dyer MD, Murali T, Sobral BW. Supervised learning and prediction of physical interactions between human and HIV proteins. Infection, Genetics and Evolution. 2011;11:917-23.
20 Qi Y, Tastan O, Carbonell JG, Klein-Seetharaman J, Weston J. Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. Bioinformatics. 2010;26:i645-i52.
21 Cui G, Fang C, Han K. Prediction of protein-protein interactions between viruses and human by an SVM model. BMC bioinformatics. 2012;13:S5.
22 Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, *et al*. The IntAct molecular interaction database in 2012. Nucleic acids research. 2012;40:D841-6.
23 Yu J, Guo M, Needham C, Huang Y, Cai L, Westhead D. Simple sequence-based kernels do not predict protein-protein interactions. Bioinformatics. 2010;26:2610 - 4.
24 Park Y, Marcotte EM. Revisiting the negative example sampling problem for predicting protein–protein interactions. Bioinformatics. 2011;27:3024-8.

25 Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic acids research. 2008;36:D202-D5.

26 Dunn JC. Well-separated clusters and optimal fuzzy partitions. Journal of cybernetics. 1974;4:95-104.

27 Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics. 2001;43:246-55.

28 Chou K-C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Current Proteomics. 2009;6:262-74.

29 Xu Y, Wen X, Shao X-J, Deng N-Y, Chou K-C. iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. International Journal of Molecular Sciences. 2014;15:7594-610.

30 Chen C, Chen L, Zou X, Cai P. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein and peptide letters. 2009;16:27-31.

31 Guo J, Rao N, Liu G, Yang Y, Wang G. Predicting protein folding rates using the concept of Chou's pseudo amino acid composition. Journal of computational chemistry. 2011;32:1612-7.

32 Yu L, Guo Y, Li Y, Li G, Li M, Luo J, *et al*. SecretP: Identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition. Journal of Theoretical Biology. 2010;267:1-6.

33 Shen H-B, Chou K-C. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Analytical biochemistry. 2008;373:386-8.

34 Zahiri J, Yaghoubi O, Mohammad-Noori M, Ebrahimpour R, Masoudi-Nejad A. PPIevo: Protein-Protein Interaction Prediction from PSSM Based Evolutionary Information. Genomics. 2013.

35 Altschul S, Madden T, Schffer A, J Zhang Z, Miller W, Lipman D. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic acids research. 1997;25:3389 - 340.

36 Assenov Y, Ramírez F, Schelhorn S-E, Lengauer T, Albrecht M. Computing topological parameters of biological networks. Bioinformatics. 2008;24:282-4.

37 Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011;27:431-2.

38 Chatr-aryamontri A, Breitkreutz B-J, Heinicke S, Boucher L, Winter A, Stark C, *et al*. The BioGRID interaction database: 2013 update. Nucleic acids research. 2013;41:D816-D23.

39 Zhang A. Protein interaction networks: computational analysis: Cambridge University Press; 2009.

40 Freeman LC. A set of measures of centrality based on betweenness. Sociometry. 1977:35-41.

41 Tastan O, Qi Y, Carbonell JG, Klein-Seetharaman J. Prediction of interactions between HIV-1 and human proteins by information integration. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing: NIH Public Access; 2009. p. 516.

42 Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, *et al*. Human Protein Reference Database--2009 update. Nucleic acids research. 2009;37:D767-72.

43 Kontaki H, Talianidis I. Cross-talk between post-translational modifications regulates life or death decisions by E2F1. Cell Cycle. 2010;9:3836-7.

44 Woodsmith J, Kamburov A, Stelzl U. Dual Coordination of Post Translational Modifications in Human Protein Networks. PLoS computational biology. 2013;9:e1002933.

45 Kuncheva L. Combining Pattern Classifiers: Methods and Algorithms. Combining Pattern Classifiers: Methods and Algorithms. 2004.

46 Saha I, Zubek J, Klingström T, Forsberg S, Wikander J, Kierczak M, *et al*. Ensemble learning prediction of protein–protein interactions using proteins functional annotations. Molecular BioSystems. 2014;10:820-30.

47 Witten IH, Frank E, Mark A. Hall. 2011. Data Mining: Practical machine learning tools and techniques. San Francisco: Morgan Kaufmann Publishers Inc; 2011.

48 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature genetics. 2000;25(1):25-9.

49 Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic acids research. 2014;42(D1):D199-D205.

50 Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic acids research. 2014;42(D1):D472-D7.

51 Lu Z, Wu X, Zhu X, Bongard J, editors. Ensemble pruning via individual contribution ordering. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining; 2010: ACM.

52 Da Wei Huang BTS, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols. 2008;4(1):44-57.

53 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995:289-300.

54 Memmott RM, Dennis PA. Akt-dependent and-independent mechanisms of mTOR regulation in cancer. Cellular signalling. 2009;21(5):656-64.

55 Letunic I, Doerks T, Bork P. SMART 7: recent updates to the protein domain annotation resource. Nucleic acids research. 2012;40(D1):D302-D5.

56 Pietrokovski S, Henikoff JG, Henikoff S. The Blocks database - a system for protein classification. Nucleic acids research. 1996;24(1):197-200.

57 Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic acids research. 2013:gkt1223.

58 Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, et al. The PROSITE database. Nucleic acids research. 2006;34(suppl 1):D227-D30.