Original Article

Using Machine Learning Approaches for Computational Prediction of Human and Hepatitis C Virus Proteins

Masoud Akbari[1], Mohamad Reza Ramezani[1], Hadi Esmaili GouvarchinGhaleh[1], Ruhollah Dorostkar[1], Mahdieh Farzanehpour[1*]

[1]Applied Virology Research Center, Baqiyatallah University of Medical Sciences, Tehran, Iran

ABSTRACT

Background:
Hepatitis C virus (HCV) infection affects around 170 million individuals worldwide, with an estimated 3% of the world's population presently afflicted. More than 350,000 people are killed each year by HCV throughout Asia and the rest of the globe due to liver disorders such as cirrhosis, chronic hepatitis, and hepatocellular carcinoma (HCC). Understanding viral-host protein interactions are essential for understanding viral infection, disease etiology, and the development of innovative therapeutics. This is due to the inherent limits of laboratory techniques for finding host-virus protein-protein interactions (PPIs). There seems to be a strong computational effect on the research of cellular infection.

Materials and Methods:
In this study, we predicted the interaction between human and HCV proteins using an ensemble learning technique.Support vector machines (SVMs) nuclear liner and radial are the cornerstones of our model, as are K-Nearest Neighbors (KNN) and Random Forest (RF). Four different feature vectors were used to encode human and HCV proteins: the tripeptide composition (TPC), The composition of k-spaced acid pairs (CKSAAP), the amino acid autocorrelation-autocovariance (AAutoCor), and the conjoint triad (CT).

Results:
The predictive power of the suggested technique is evaluated using a benchmark dataset that contains both consistently positive and negative PPIs. A support vector machine (Radial-SVM) model was used to predict which human proteins interact with HCV. To achieve accuracy and specificity of 84.9 and 88.3 percent, we employed tenfold cross-validation and principal component analysis (PCA).

Conclusion:
Our technique correctly predicts PPIs based on human and HCV proteins. The discovery of HCV-human protein interaction networks, enriched pathways, gene ontology, and functional categories has improved our knowledge of HCV infection.

Keywords: HCV, PPIs, Computational, SVM, Prediction

Please cite this paper as:
Akbari M, Ramezani MR, Esmaili GouvarchinGhaleh H, Dorostkar R, Farzanehpour M. Using machine learning approaches for computational prediction of human and hepatitis C virus proteins. *Govaresh* 2023;28: 66-77.

Corresponding author:
Mahdieh Farzanehpour, MD,
Applied Virology Research Center,
Baqiyatallah University of Medical Sciences, Tehran, Iran
Email: Mah_farzanehpour@yahoo.com

Received: 21 Aug. 2022
Revised: 7 Jan. 2023
Accepted: 8 Jan. 2023

INTRODUCTION

As obligate intracellular parasites, the genome replication and propagation of viruses are entirely reliant on host cellular machinery. The chronic hepatitis C virus (HCV) was the first positive-strand RNA virus family to be identified in Decades after its discovery, infection remains a severe public health issue that costs the medical system billions of dollars annually. According to the World Health Organization (WHO), around 3-4 million, new instances of HCV infection occur each year, with approximately 350,000 deaths (1-6). The Hepaciviruses are a group of viruses that belong to the Flaviviridae family. HCV is a hepacivirus. The HCV genome contains six non-structural (NS) and four structural proteins. Nucleoside analogs 2, 3, 4, and 5 represent structural HCV proteins, whereas nucleoside analogs 1, 2, and 7 represent non-structural HCV proteins (7-9).

To undertake disease studies and develop new therapeutics, we need to understand how proteins interact with one another. Experiments and computer simulations may gather data on a wide range of protein-protein interactions. Among the downsides of experimental techniques include high false positive rates, expensive costs, and complex and time-consuming processes, researchers have turned to computational approaches (10-12). Using computational PPI prediction methods, more studies may be done on particular targets. Computational approaches to forecasting PPIs may be more efficient and cost-effective than experimental methods (13,14). Even more advantageous than experimental approaches is the ability to examine proteins by mapping binary links in a large network according to their different performance using computational techniques (15-17). Four major groups of current host-pathogen PPI prediction approaches exist: those based on homology, those based on structural information, those based on sequence, and those that employ machine learning (18,19).

A combination of practical and computational approaches was used to create the HCV protein interactome map. In contrast to experimental investigations, computational studies have generally focused on large-scale research that either validated experimental results or predicted PPIs computationally (20–22).

HCV interactome was developed jointly, however, it seems that the growing number of released data will help us comprehend the molecular components of HCV and the process by which its proteins are infected. Based on the workflow shown in Figure 1, in this research, PPIs were predicted for human-HCV infection. The approach for predicting PPIs between human and HCV proteins was developed using ensemble learning. Random forest, support vector machine (radial and liner nuclear), and KNN are some of the most often used machine learning algorithms that the SVM (radial nuclear SVM with the principle component analysis (PCA)) showed a reasonable level of performance (an average sensitivity of 81.6%, specificity of 88.3%, and accuracy of 84.9%). Also, with the help of exploring and analyzing the HCV-human network, it was determined that human genes had some critical interaction with most HCV proteins. Finally, in other to diminish the complexity and highlight biological processes gene ontology (GO) enrichment analysis was done.

MATERIALS AND METHODS

Benchmark dataset

Two datasets were constructed to assess the approach: both positive and negative datasets.

Positive interactions

The IntAct database (23) was used to collect all human-hepatitis C virus PPIs. Then a strategy was created to eliminate HCV infection in non-human species. An excellent meeting that stands out involves direct physical contact or touching (Pairwise Similarity (PS)). There were 1115 interactions in PS.

Negative interactions

Finding correct negative PPIs is one of the most difficult PPI prediction tasks (24,25). During an experiment, all human protein combinations that did not show any interaction in the PS were considered negative data. Due to the mismatch between the positive and negative data sets, we produced a negative dataset with the same amount of proteins. The study found over 1100 negative interactions in the negative interaction set (NS). This completed our benchmark dataset development process.

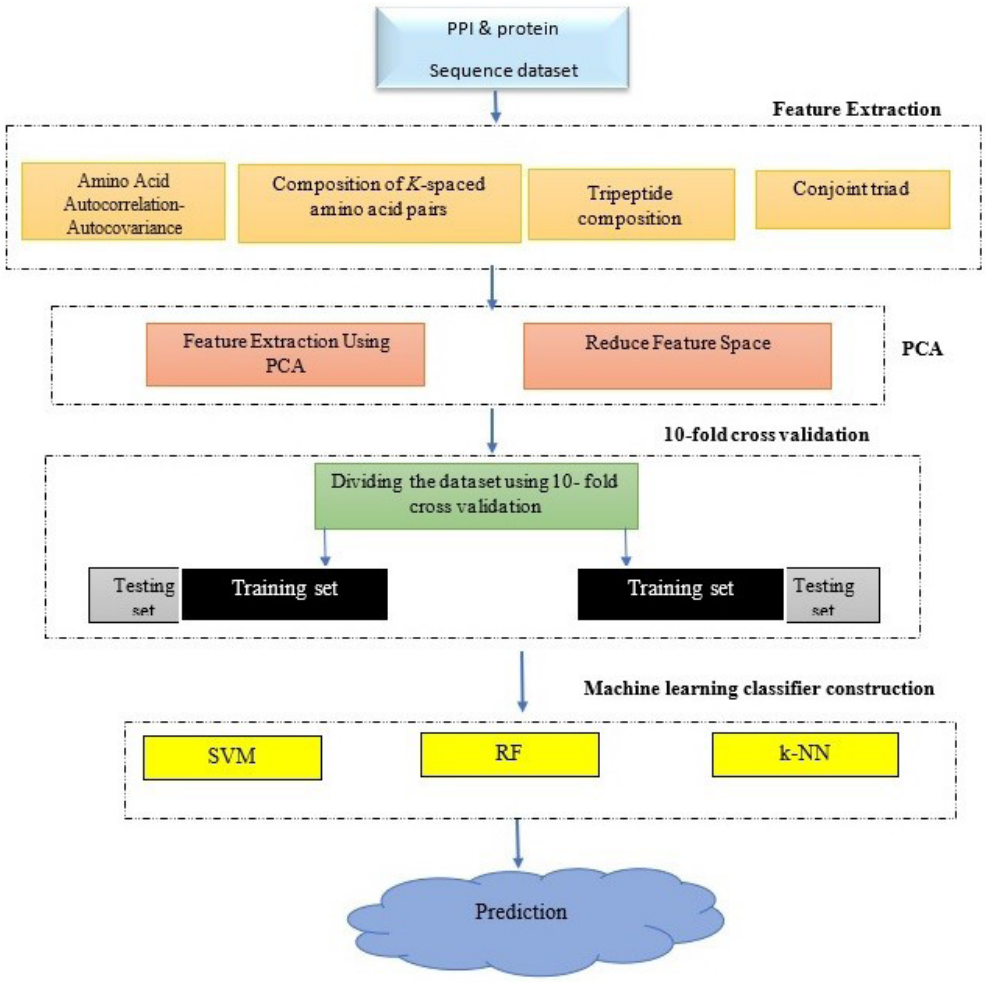

Figure 1. The workflow of our pipeline to predict human-virus PPIs

Protein sequence representation

To predict PPIs using machine learning methods, computational challenges must be addressed before protein sequences are processed.

The human and HCV proteins were encoded in the feature vectors using a total of four attributes. This list includes autocorrelation-autocovariance and conjoint triads (CT). Each of these points will be examined in further depth in the paragraphs that follow.

Tripeptide composition

The following formula was used to identify each of the 8000 possible dipeptides created from 20 amino acids. Additionally, it may be used to categorize samples based on their composition (26).

$$TPC(i) = \frac{\text{Total number of tripeptides}(i)}{\text{Total number of all possible tripeptides}} \times 100$$

where TPC(n) is a tripeptide n out of 8000 tripeptides.

Composition of K-spaced amino acid pairs

Chen and his teammates' CKSAAP were initially utilized in bioinformatics research back in the 1980s (27). CKSAAP's approach is outlined in the following paragraphs. Since there are 21 types of amino acids (including the gap), a sequence fragment with a window size of 2r+1 and (21*21)=441 different amino acid pairings may be constructed for each of the 4 k values (O). A distance of k amino acids separates two amino acids. Three k-spaces make up the name "AXXXA". According

to the findings of this experiment, the kmax value of three was picked to collect 1764 distinct amino acid pairings for each sequence. To create the feature vector with k=2, the following formula was applied (FV) (28-30).

$$FV=\left(\frac{NAxxA}{Ntotal},\frac{NAxxC}{Ntotal},\frac{NAxxG}{Ntotal}\ldots,\frac{NXxxX}{Ntotal}\right)_{441}$$

A fragment residue of 36 amino acids with k=0, 1, 2, 3, 4, 5, N entire=L-k-1 will be 35, 34, 33, 32, 31, and 30 for the overall length of the composition, where «x» signifies any of the composition›s 21 amino acids and Ntotal denotes the total composition residues› length.

Conjoint triad

Shen and colleagues reported that the conjoined triad (CT) was shown (31). These seven groups were constructed based on their dipoles and volumes, as stated in Table 1, to facilitate the code representation of the 20 standard amino acids and to account for synonymous mutations. Three nearby amino acids may be regarded as a single unit, and the characteristics of particular amino acids and their adjacent amino acids are well understood (31). The following procedure is used to produce descriptor vectors.

The protein sequence begins by replacing each amino acid with an index depending on the amino acid's categorization. As an example, "RLASCTELRTLNLARN" has been replaced with 5213736253242154. The next step is to represent an amino acid sequence in binary space (V, F). The vector spaces (V and F) and frequency vectors (V and F) of sequence characteristics are shown here. As a result, V should include seven times seven amino acids, resulting in I=1, 2, 3, 443. Each protein is covalently linked to a unique F vector. In other words, the length of the amino acid sequence influences the value of fi (the frequency of conjoint triad) (directly. As a result, a longer amino acid sequence has a larger value of fi, making comparisons between two distinct proteins more difficult. As a consequence, they resorted to normalization to address the problem: Between 0 and 1, there is a normalized di? value equal to (fi–min/max)/max (where the normalized range of di is between 0 and 1). Additionally, they coupled the vector spaces of two proteins to bring everything together. When two proteins are coupled, they form a total of 686-dimensional vectors (343 for each protein).

Table 1. Division of amino acids based on the dipoles and volumes of the side chains

No.	Group
1	A, G, V
2	C
3	D, E
4	F, I, L, P
5	H, N, Q, W
6	K, R
7	M,S,T,Y

Principal component analysis (PCA)

PCA (may be used to analyze large multidimensional data sets by simplifying them. With this widely used data analysis method, we may reduce the system's dimensionality while still preserving information about the relationships between its many constituent parts (32,33). The "principal components (PC)" are a collection of linear combinations formed by PCA. By eliminating low-variability characteristic bits, we may dramatically decrease the dimensionality of data. The original M-dimensional data patterns can be correctly converted into a two-dimensional feature space. Due to its simplicity both theoretically and computationally, PCA is a great tool for understanding both.

Evaluation measures

For the sake of determining the accuracy of human-HCV PPI predictions, four essential features may be used: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The total system's reliability, sensitivity, and precision were all evaluated using the tenfold cross-validation method (Prec.). The next lines describe each of them:

RESULTS AND DISCUSSION

Cross-validation analysis

To reduce the effect of training and testing data, the proposed prediction technique was assessed using a tenfold cross-validation process.

Each dataset is divided into ten equal training and testing sets. Each subgroup was utilized independently for training and testing, and only one protein combination was

investigated. The ultimate prediction result is calculated by averaging the results of all ten testing sets.

Text mining of HCV interaction protein

We identified 1325 human proteins that are particularly targeted by HCV using the IntAct Molecular Interaction Database. After removing duplicate contacts, 11 HCV proteins were identified as distinct HCV-human interactors, accounting for 1115 interactions (Supplementary data). Numerous studies 1-4 have shown that the HCV proteins distribute their interactions at various speeds. To get a better knowledge of HCV-human protein interactions, we examined the whole interaction between several HCV isolation proteins and human proteins. In this investigation, several isolates (genome polyprotein) were revealed to have the highest number of contacts with other isolates, including genotype 1b (isolate con1) and genotype 2a (isolate JFH-1).

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

Bioinformatics and computational biology have a challenge in accurately predicting human-HCV PPIs computationally. Multiple studies have recently proposed an approach for dealing with problems (34–36) by employing different models to forecast them. According to table 2, for multidimensional data sets, SVM (radial nuclear) and PCA may help minimize the number of dimensions while maintaining all of the information (an average sensitivity of 81.6%, specificity 88.3%, and accuracy of 84.9%). We utilized these two models, and the SVM (radial nuclear SVM with PCA) showed a reasonable level of performance (an average sensitivity of 81.6%, specificity of 88.3%, and accuracy of 84.9%).

Analysis of viral interaction networks

Understanding how HCV interacts with the human body will aid in our understanding of the virus's potential to control a variety of biological processes. During viral infections, several human proteins are targeted by various viral proteins. So we next constructed a network representing the 1115 HCV-human proteins interaction with nodes corresponding to different isolates of HCV proteins and 956 human factors derived from the databases that part of interactions have shown in Figure 2. We investigated the possibility of an interaction network of HCV different isolates (like 1b con1, 1a H77, and 2a JFH) proteins with human proteins (Figure 2). Also, the interactive network of interaction human proteins with

Table 2. Prediction performance of the proposed method in the 10-fold cross-validation

			Accuracy	Sensitivity	Specificity	interacting protein pairs	Non-interacting protein pairs
K-Nearest Neighbors	With PCA	Test	0.8834	0.843	0.9238	1115	1115
		Train	0.9008	0.8733	0.9283		
	Without PCA	Test	0.8879	0.8475	0.9283	1115	1115
		Train	0.9058	0.8823	0.9294		
Support vector machine (Linear nuclear)	With PCA	Test	0.8274	0.8251	0.8296	1115	1115
		Train	0.9619	0.9496	0.9742		
	Without PCA	Test	0.7937	0.8117	0.7758	1115	1115
		Train	0.9697	0.963	0.9765		
Support vector machine (Radial nuclear)	With PCA	Test	0.8498	0.8161	0.8834	1115	1115
		Train	0.9283	0.9193	0.9372		
	Without PCA	Test	0.8834	0.8744	0.8924	1115	1115
		Train	0.9148	0.9193	0.9103		
Random foresty	With PCA	Test	0.8879	0.8879	0.8879	1115	1115
		Train	1	1	1		
	Without PCA	Test	0.8946	0.8879	0.9013	1115	1115
		Train	0.9983	1	0.9966		

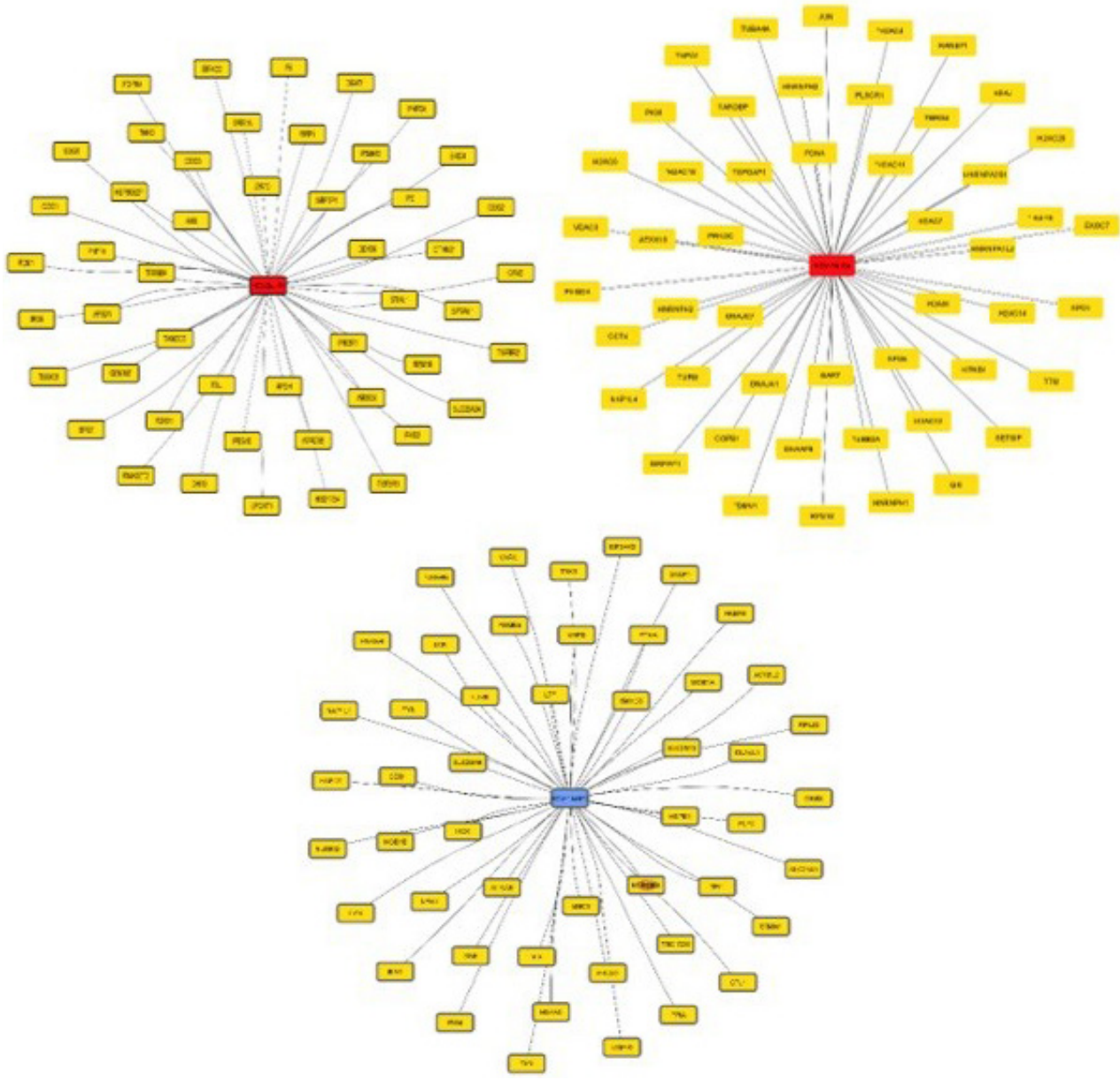

Figure 2. A glimpse of 1b con1, 1a H77, and 2a JFH HCV-human PPI networks. Central HCV-human square indicates HCV proteins and edges indicate HCV -human protein interactions

different HCV isolates has been shown in Figure 3.

Interestingly some of the identified human proteins like cellular tumor antigen p53 (TP53), HSP90AA1 (Heat Shock Protein 90 Alpha Family Class A Member 1), ACTB (Actin Beta), CTNNB1 (Catenin Beta 1), EP300 (E1A Binding Protein P300), HSPA8 (Heat Shock Protein Family A (Hsp70) Member 8), HSPA5 (Heat Shock Protein Family A (Hsp70) Member 5), STAT3 (Signal Transducer And Activator Of Transcription 3) HSP90AB1 (Heat Shock Protein 90 Alpha Family Class B Member 1), and HRAS (HRas Proto-Oncogene, GTPase) are particularly affected by HCV infection (figure 3 with red square have shown, supplementary data). our results were in agreement with the experimental finding reported

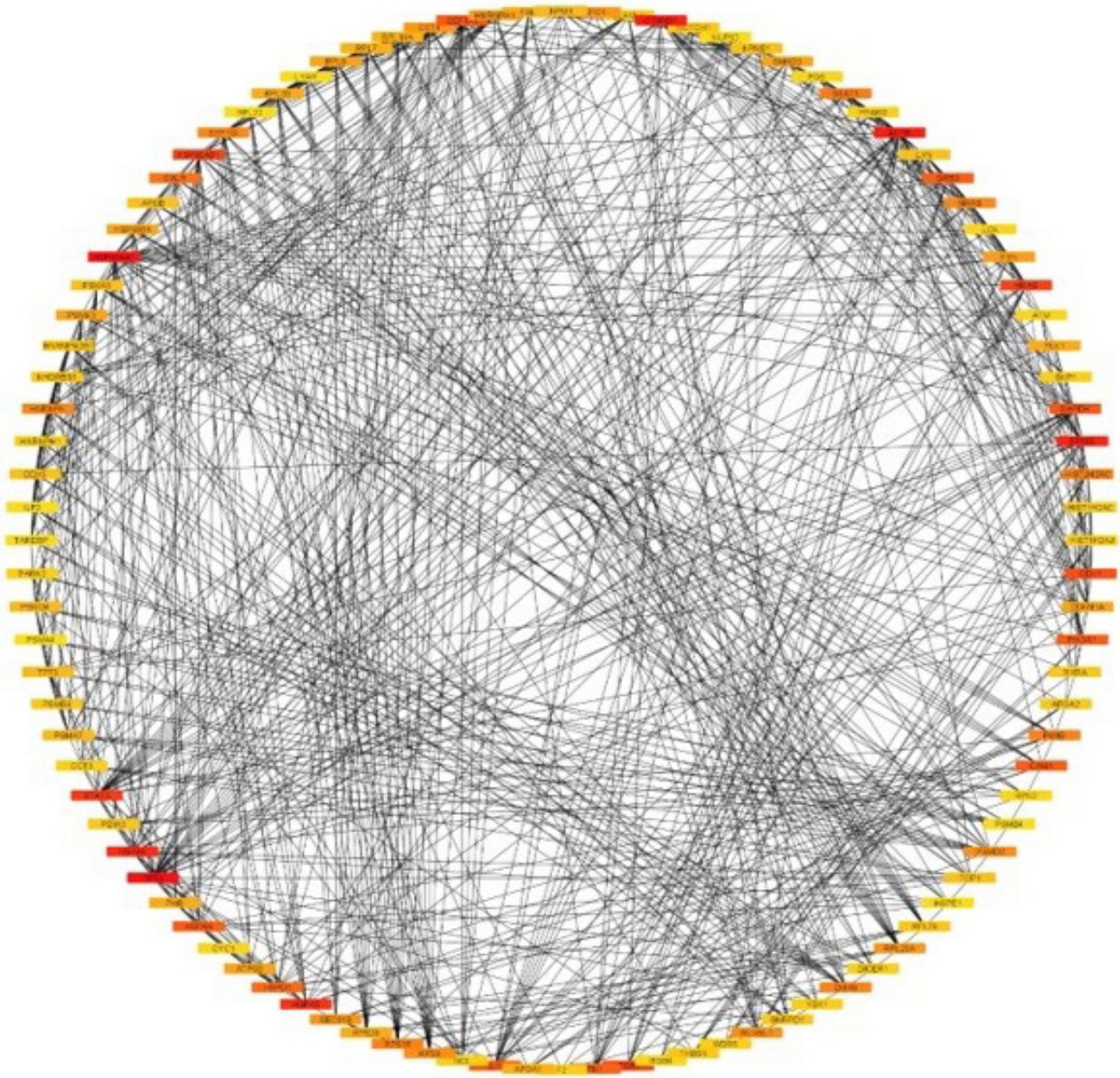

Figure 3. A glimpse simulation of the interactive network of activated human proteins with different HCV isolates has been drawn using Cytoscape software

by Chen and colleagues who showed that the Hsp70 family of proteins could participate in multiple steps of the HCV lifecycle 5. Also p53 tumor antigen has been shown disrupted in HCV-infected cells 6. HSP90 7 controls cell proliferation, motility, angiogenesis, signal transmission, and stress adaptation. HSP90AA1 and TOM34 were upregulated in HCV-induced hepatocellular cancer 8. The HRas signal transduction promotes hepatitis C virus cell entry by triggering the assembly of the host tetraspanin receptor complex 9. Yoshida and co-workers reported that the core protein of HCV can directly interact with and trigger STAT3 through phosphorylation of the critical tyrosine residue 10. Taken together, the relationship between these human proteins determined that the above-mentioned proteins play a critical role in the PPI network between HCV and humans.

Enrichment analysis

To enhance the number of human proteins known to interact with HCV, researchers turned to DAVID (37), an annotation, visualization, and discovery database. GO categories and pathways that were enriched are summarized in table 3 (38). Substantial enrichment of a term was defined as Benjamini-adjusted P values less than 0.005. Significant enrichment was defined as Benjamini-adjusted P values less than 0.005. Protein processing in the ER and measles infection were other GO keywords and pathways that were examined in this research. A viral process, protein folding (GO:0006457), and cell-cell adhesion (GO:0016032) are all examples (GO:0098609). The functional diversity of host proteins was uncovered using hierarchical clustering based on GO biological process (GO-BP) classifications, which have previously been identified as critical characteristics of proteins that interact with HCV. As with previous findings, a large number of interactors were discovered to play critical roles in cell proliferation, cell cycle, DNA replication/repair, and signal transduction, implying that these interactors are likely involved in processes other than a viral invasion that contribute to carcinogenesis (3,4,11,12). Additionally, the KEGG pathway analysis identifies several diseases that may develop in the human body as a consequence of HCV infection. The following table shows the principal

Table 3. 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
Pathway	Biological process	viral process(GO:0016032)	7.60E-15
		protein folding (GO:0006457)	8.85E-13
		cell-cell adhesion (GO:0098609)	1.86E-12
		protein stabilization (GO:0050821)	1.37E-08
		response to endoplasmic reticulum stress (GO:0034976)	1.10E-05
		platelet degranulation (GO:0002576)	1.34E-05
	Molecular function	protein binding (GO:0005515)	8.89E-52
		poly(A) RNA binding (GO:0044822)	2.54E-33
		unfolded protein binding (GO:0051082)	4.06E-14
		cadherin binding involved in cell-cell adhesion (GO:0098641)	5.12E-13
		identical protein binding (GO:0042802)	4.35E-11
		enzyme binding (GO:0019899)	4.40E-09
	Cellular component	Membrane (GO:0016020)	3.37E-41
		Extracellular exosome (GO:0070062)	1.23E-33
		Cytosol (GO:0005829)	5.51E-25
		Extracellular matrix (GO:0031012)	7.78E-21
		Melanosome (GO:0042470)	1.85E-14
		focal adhesion (GO:0005925)	4.71E-14
		Cytoplasm (GO:0005737)	1.12E-13
	KEGG	Protein processing in the endoplasmic reticulum (hsa04141)	5.94E-14
		Measles(hsa05162)	5.10E-05
		Hepatitis B (hsa05161)	6.09E-04
		Phagosome (hsa04145)	0.001649694
		Pathogenic Escherichia coli infection (hsa05130)	0.001649694
		Herpes simplex infection (hsa05168)	0.001649694
		Proteasome (hsa03050)	0.001649694
		Viral carcinogenesis (hsa05203)	0.001649694

pathways for the predicted human proteins that were congruent with Yamashita and colleagues' 13 findings. This study unambiguously proves that human proteins play a role in the direct or indirect interaction between virus illnesses and human proteins. Additionally, Table 4 (39-42) and Figure 4 highlight the enriched domains and

Table 4. 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
SMRT	EGF_CA (SM00179)	2.87E-09
	EGF (SM00181)	6.19E-08
PROSITE	Aspartic acid and asparagine hydroxylation site (PS00010)	3.11E-08
	Calcium-binding EGF-like domain signature (PS01187)	3.11E-08
	EGF-like domain signatures and profile (PS01186)	5.58E-07
	EGF-like domain signatures and profile (PS50026)	9.29E-06
	EGF-like domain signatures and profile (PS00022)	9.29E-06
	Endoplasmic reticulum targeting sequence (PS00014)	1.95E-04
PFAM	Calcium-binding EGF domain (PF07645)	2.86E-06
	Complement Clr-like EGF-like (PF12662)	1.40E-04
	Thioredoxin (PF00085)	0.001403487
INTERPRO	EGF-like calcium-binding (IPR001881)	4.85E-09
	EGF-like calcium-binding, conserved site (IPR018097)	2.00E-08
	EGF-type aspartate/asparagine hydroxylation site (IPR000152)	2.56E-08
	EGF-like, conserved site (IPR013032)	9.22E-07
	Epidermal growth factor-like domain(IPR000742)	2.14E-06
	Complement Clr-like EGF domain (IPR026823)	6.32E-05

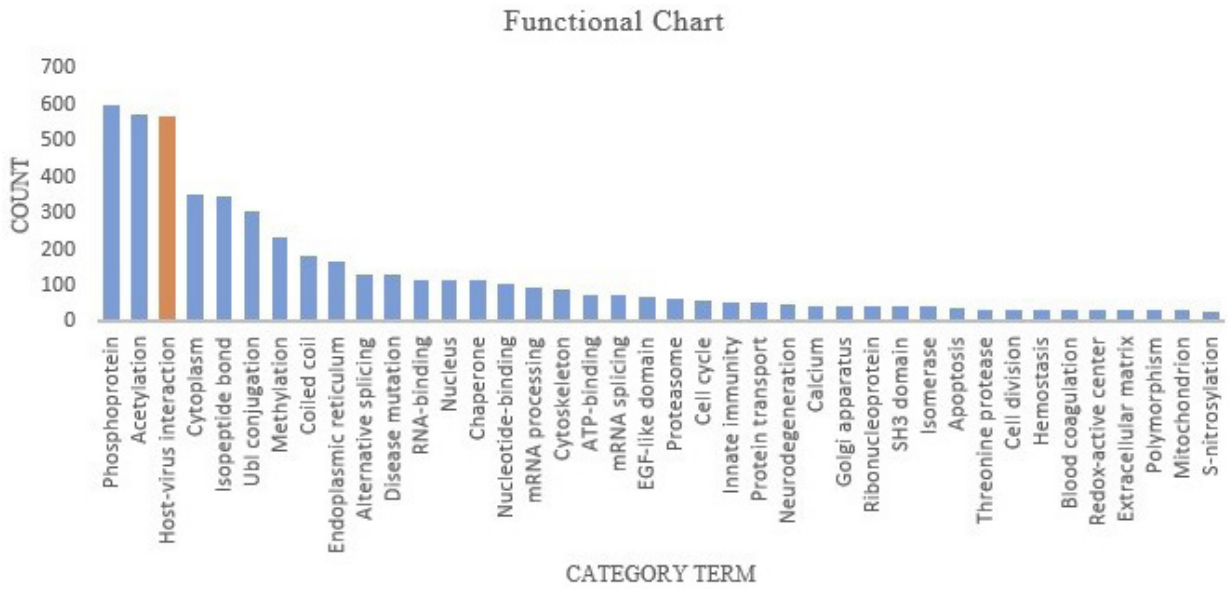

Figure 4. Functional chart of HCV targets. A subset of human proteins that interacted with different HCV isolates to their cellular functions based on Gene Set Enrichment Analysis (a term was considered significantly enriched if the Benjamini corrected P value was less than 0.005)

functional categories in the predicted human proteins interacting with HCV collection. Numerous domains highlighted in Table 4 have been experimentally validated (1,2,14,15). The findings also reveal that human proteins have improved distinguishing properties that might be used in future experiments and computational studies [Supplementary data].

CONCLUSION

Ensemble learning was utilized in this work to predict the interactions of human and hepatitis C virus (HCV) proteins. Encoding protein pairings was accomplished via the use of four distinct descriptions. As foundation classifiers, four unique classifiers were used: random forest, KNN, and support vector machine (SVM) using radial and liner nuclear features. The results demonstrate that our method, the SVM (radial nuclear), performs satisfactorily in a 10-fold cross-validation analysis on our benchmark dataset with PCA (on average, the sensitivity of 81.6%, specificity of 88.3%, and accuracy of 84.9%), indicating that it has a reasonably high performance in representing better patterns from PPIs. Also, the interaction networks of HCV-human proteins, enriched pathways, gene ontology, and enriched domains developed a more complete and comprehensive understanding of HCV infection and also provided insight into HCV manipulation of pathways.

AVAILABILITY OF DATA AND MATERIALS

The authors confirm that the data supporting the findings of this study are available within the article or its supplementary materials.

CONFLICT OF INTERESTS

The authors declare no conflict of interest related to this work.

CONSENT FOR PUBLICATION

Not applicable

ETHICAL APPROVAL

This article does not contain any studies involving animals performed by any of the authors.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

1. Lavanchy D. Evolving epidemiology of hepatitis C virus. *Clin Microbiol Infect* 2011;17(2):107-15. doi: 10.1111/j.1469-0691.2010.03432.x
2. World Health Organization. Hepatitis C--global prevalence (update). *Wkly Epidemiol Rec* 1999;74(49):425-7.
3. World Health Organization. Global surveillance and control of hepatitis C. Report of a WHO Consultation organized in collaboration with the Viral Hepatitis Prevention Board, Antwerp, Belgium. *J Viral Hepat* 1999;6(1):35-47.
4. Alter MJ. Epidemiology of hepatitis C virus infection. *World J Gastroenterol* 2007;13(17):2436-41. doi: 10.3748/wjg.v13.i17.2436
5. Plauzolles A, Lucas M, Gaudieri S. Influence of host resistance on viral adaptation: hepatitis C virus as a case study. *Infect Drug Resist* 2015;8:63-74. doi: 10.2147/idr.s49891
6. Lai CK, Jeng KS, Machida K, Lai MM. Association of hepatitis C virus replication complexes with microtubules and actin filaments is dependent on the interaction of NS3 and NS5A. *J Virol* 2008;82(17):8838-48. doi: 10.1128/jvi.00398-08
7. Dimitrova M, Imbert I, Kieny MP, Schuster C. Protein-protein interactions between hepatitis C virus nonstructural proteins. *J Virol* 2003;77(9):5401-14. doi: 10.1128/jvi.77.9.5401-5414.2003
8. Penin F. Structural biology of hepatitis C virus. *Clin Liver Dis* 2003;7(1):1-21. doi: 10.1016/s1089-3261(02)00066-1
9. Flajolet M, Rotondo G, Daviet L, Bergametti F, Inchauspé G, Tiollais P, et al. A genomic approach of the hepatitis C virus generates a protein interaction map. *Gene* 2000;242(1-2):369-79. doi: 10.1016/s0378-1119(99)00511-9
10. Gonzalez MW, Kann MG. Chapter 4: protein interactions and disease. *PLoS Comput Biol* 2012;8(12):e1002819. doi: 10.1371/journal.pcbi.1002819
11. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 2005;122(6):957-68. doi: 10.1016/j.cell.2005.08.029
12. Pei D, Xu J, Zhuang Q, Tse HF, Esteban MA. Induced pluripotent stem cell technology in regenerative medicine and biology. In: Kasper C, van Griensven M, Pörtner R, eds. *Bioreactor Systems for Tissue Engineering II: Strategies for the Expansion and Directed Differentiation of Stem Cells*. Berlin, Heidelberg: Springer; 2010. p. 127-41. doi: 10.1007/10_2010_72
13. Wuchty S. Computational prediction of host-parasite protein interactions between *P. falciparum* and *H. sapiens*. *PLoS One* 2011;6(11):e26960. doi: 10.1371/journal.pone.0026960
14. Krishnadev O, Srinivasan N. Prediction of protein-protein interactions between human host and a pathogen and its application to three pathogenic bacteria. *Int J Biol Macromol*

2011;48(4):613-9. doi: 10.1016/j.ijbiomac.2011.01.030
15. Doolittle JM, Gomez SM. Mapping protein interactions between Dengue virus and its human and insect hosts. *PLoS Negl Trop Dis* 2011;5(2):e954. doi: 10.1371/journal.pntd.0000954
16. 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;2:27. doi: 10.1186/1755-8794-2-27
17. Dyer MD, Murali TM, Sobral BW. Computational prediction of host-pathogen protein-protein interactions. *Bioinformatics* 2007;23(13):i159-66. doi: 10.1093/bioinformatics/btm208
18. Cui G, Fang C, Han K. Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinformatics* 2012;13(Suppl 7):S5. doi: 10.1186/1471-2105-13-s7-s5
19. Dyer MD, Murali TM, Sobral BW. Supervised learning and prediction of physical interactions between human and HIV proteins. *Infect Genet Evol* 2011;11(5):917-23. doi: 10.1016/j.meegid.2011.02.022
20. Tripathi LP, Kataoka C, Taguwa S, Moriishi K, Mori Y, Matsuura Y, et al. Network based analysis of hepatitis C virus core and NS4B protein interactions. *Mol Biosyst* 2010;6(12):2539-53. doi: 10.1039/c0mb00103a
21. Roohvand F, Maillard P, Lavergne JP, Boulant S, Walic M, Andréo U, et al. Initiation of hepatitis C virus infection requires the dynamic microtubule network: role of the viral nucleocapsid protein. *J Biol Chem* 2009;284(20):13778-91. doi: 10.1074/jbc.M807873200
22. de Chassey B, Navratil V, Tafforeau L, Hiet MS, Aublin-Gex A, Agaugué S, et al. Hepatitis C virus infection protein network. *Mol Syst Biol* 2008;4:230. doi: 10.1038/msb.2008.66
23. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, et al. The IntAct molecular interaction database in 2010. *Nucleic Acids Res* 2010;38(Database issue):D525-31. doi: 10.1093/nar/gkp878
24. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 2010;26(13):1608-15. doi: 10.1093/bioinformatics/btq249
25. Park Y, Marcotte EM. Revisiting the negative example sampling problem for predicting protein-protein interactions. *Bioinformatics* 2011;27(21):3024-8. doi: 10.1093/bioinformatics/btr514
26. Gupta S, Sharma AK, Shastri V, Madhu MK, Sharma VK. Prediction of anti-inflammatory proteins/peptides: an insilico approach. *J Transl Med* 2017;15(1):7. doi: 10.1186/s12967-016-1103-6
27. Chen K, Kurgan L, Rahbari M. Prediction of protein crystallization using collocation of amino acid pairs. *Biochem Biophys Res Commun* 2007;355(3):764-9. doi: 10.1016/j.bbrc.2007.02.040
28. Chen Z, Zhou Y, Zhang Z, Song J. Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features. *Brief Bioinform* 2015;16(4):640-57. doi: 10.1093/bib/bbu031
29. Hasan MM, Zhou Y, Lu X, Li J, Song J, Zhang Z. Computational identification of protein pupylation sites by using profile-based composition of k-spaced amino acid pairs. *PLoS One* 2015;10(6):e0129635. doi: 10.1371/journal.pone.0129635
30. Chen K, Kurgan LA, Ruan J. Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. *BMC Struct Biol* 2007;7(1):25. doi: 10.1186/1472-6807-7-25
31. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, et al. Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A* 2007;104(11):4337-41. doi: 10.1073/pnas.0607879104
32. Zhang S, Ye F, Yuan X. Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via PSSM. *J Biomol Struct Dyn* 2012;29(6):634-42. doi: 10.1080/07391102.2011.672627
33. You Z, Wang S, Gui J, Zhang S. A novel hybrid method of gene selection and its application on tumor classification. In: Huang DS, Wunsch DC, Levine DS, Jo KH, eds. *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence: 4th International Conference on Intelligent Computing, ICIC 2008 Shanghai, China, 2008 Proceedings 4*. Berlin, Heidelberg: Springer; 2008. p. 1055-68. doi: 10.1007/978-3-540-85984-0_127
34. 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):3147-54. doi: 10.1039/c4mb00410h
35. Brito AF, Pinney JW. Protein-protein interactions in virus-host systems. *Front Microbiol* 2017;8:1557. doi: 10.3389/fmicb.2017.01557
36. Farooq QUA, Khan FF. Construction and analysis of a comprehensive protein interaction network of HCV with its host *Homo sapiens*. *BMC Infect Dis* 2019;19(1):367. doi: 10.1186/s12879-019-4000-9
37. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;4(1):44-57. doi: 10.1038/nprot.2008.211
38. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 2014;42(Database issue):D199-205. doi: 10.1093/nar/gkt1076

39. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, et al. The PROSITE database. *Nucleic Acids Res* 2006;34(Database issue):D227-30. doi: 10.1093/nar/gkj063
40. Letunic I, Doerks T, Bork P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res* 2012;40(Database issue):D302-5. doi: 10.1093/nar/gkr931
41. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, et al. The Pfam protein families database. *Nucleic Acids Res* 2008;36(Database issue):D281-8. doi: 10.1093/nar/gkm960
42. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res* 2001;29(1):37-40. doi: 10.1093/nar/29.1.37