Available online at www.sciencedirect.com

ScienceDirect

Biomedical Journal

journal homepage: www.elsevier.com/locate/bj

Original Article

Machine learning techniques for sequence-based prediction of viral–host interactions between SARS-CoV-2 and human proteins

Lopamudra Dey, Sanjay Chakraborty, Anirban Mukhopadhyay *

Department of Computer Science & Engineering, Heritage Institute of Technology, Kolkata, India
Department of Information Technology, Techno Main, Saltlake, Kolkata, India
Department of. Computer Science & Engineering, University of Kalyani, Kalyani, India

ARTICLE INFO

Article history:
Received 13 May 2020
Accepted 5 August 2020
Available online 3 September 2020

Keywords:
COVID-19
SARS-CoV-2
Protein–protein interaction
Supervised classification
Machine learning
Classifier ensemble

ABSTRACT

Background: COVID-19 (Coronavirus Disease-19), a disease caused by the SARS-CoV-2 virus, has been declared as a pandemic by the World Health Organization on March 11, 2020. Over 15 million people have already been affected worldwide by COVID-19, resulting in more than 0.6 million deaths. Protein–protein interactions (PPIs) play a key role in the cellular process of SARS-CoV-2 virus infection in the human body. Recently a study has reported some SARS-CoV-2 proteins that interact with several human proteins while many potential interactions remain to be identified.

Method: In this article, various machine learning models are built to predict the PPIs between the virus and human proteins that are further validated using biological experiments. The classification models are prepared based on different sequence-based features of human proteins like amino acid composition, pseudo amino acid composition, and conjoint triad.

Result: We have built an ensemble voting classifier using SVM^{Radial}, $SVM^{Polynomial}$, and Random Forest technique that gives a greater accuracy, precision, specificity, recall, and F1 score compared to all other models used in the work. A total of 1326 potential human target proteins of SARS-CoV-2 have been predicted by the proposed ensemble model and validated using gene ontology and KEGG pathway enrichment analysis. Several repurposable drugs targeting the predicted interactions are also reported.

Conclusion: This study may encourage the identification of potential targets for more effective anti-COVID drug discovery.

According to the World Health Organization (WHO), the coronavirus disease (COVID-19) pandemic, caused by a novel strain of coronavirus called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus infection, is one of the most crucial diseases in the current scenario. It has infected over 15 million people from more than 200 countries

* *Corresponding author*. Department of Computer Science & Engineering, University of Kalyani, Kalyani, Nadia, West Bengal-741235, India.
E-mail address: anirban@klyuniv.ac.in (A. Mukhopadhyay).
Peer review under responsibility of Chang Gung University.
https://doi.org/10.1016/j.bj.2020.08.003

At a glance of commentary

Scientific background on the subject

A comprehension of how SARS-CoV-2 virus proteins interact with the host cells for survival and reproduction is essential for drug exploitation. Protein-Protein Interaction (PPI) is one way the viruses interact with their hosts. Identifying PPIs between the virus and the host proteins help explain how these virus proteins replicate and cause the disease.

What this study adds to the field

This study may encourage the identification of potential target human proteins of the SARS-CoV-2 virus. A number of repurposable drugs of these predicted human proteins are also reported, which may accelerate anti-COVID drug discovery.

while causing death of more than 0.6 million people. The disease has created immense pressure and tension in the worldwide healthcare systems. At the end of the year 2019, Wuhan city of China reported the first case of the novel coronavirus infection. Now from Asia to Europe and America, its deadly effect is threatening the whole world [1]. Genomic analysis showed that SARS-CoV-2 is phylogenetically related to SARS-like bat viruses. Hence, bats could be the possible source of the viral replication [2]. Pangolins have also been identified as a potential intermediate host of novel coronavirus [3]. The usual symptoms of COVID-19 affected patients are pneumonia, shortness of breath, cough and cold, fever, and multiple organ failure [4]. The genetic characteristics of SARS-CoV-2 should be well understood to fight against this virus. It is a single-stranded RNA virus consisting of approximately 27–32 kb with particle size ranging from 65 to 125 nm in diameter [3]. The world healthcare systems are rigorously searching for a vaccine to mitigate the spread of the virus. Besides that, they isolate the infected patients along with some general medicine as immediate treatment and care.

SARS-CoV-2 comprises four main structural proteins including spike (S) glycoprotein, small envelope (E) glycoprotein, membrane (M) glycoprotein, and nucleocapsid (N) protein, in addition to many accessory proteins [5]. A comprehension of how these virus proteins interact with the host cells for survival and reproduction is essential for drug exploitation. Protein–Protein Interaction (PPI) is one way the viruses interact with their hosts. Identifying PPIs between the virus and the host proteins helps explain how the virus proteins work and how they replicate and cause the disease. Over the past decades, experimental strategies for recognizing PPIs have been established. Nevertheless, these experimental high-throughput screens are primarily used to classify intra-species PPIs while inter-species interactomes remained largely understudied. In comparison, laboratory identification of PPIs is usually time-consuming, laborious, and difficult to achieve complete protein interactomes. Therefore, efficient computational methods for PPI prediction are used to bridge the gap by presenting experimentally testable hypotheses and removing protein pairs having a low probability of interaction to reduce the selection of PPI candidates. Computational techniques have been popularly used for predicting viral–host interactions previously [6–8].

Few works have already pursued a broad range of applications of Artificial Intelligence (AI) and Machine Learning (ML) to cover medical challenges and outbreak prediction of COVID-19 pandemic, described in Ref. [9,10]. Kassani et al. have used publicly available COVID-19 dataset of chest X-ray [3] and Barstugan et al. analyzed Computerized Tomography (CT) images in Ref. [11] for automatic COVID-19 classification using deep learning method. Horry et al. [12] have also followed the same path of X-ray based COVID-19 detection using Artificial Intelligence and pre-trained deep learning models. However, Barstugan et al. used only CT images for COVID-19 classification using a *k*-fold Support vector machine (SVM) classifier [4,11]. These methods have some drawbacks. There is a possibility to raise the issue of over-fitting in deep learning due to the limited number of trained images [3,12]. Besides that, it also has limitations in classifying more challenging instances with vague, low contrast boundaries, and the presence of artifacts [13]. These approaches are time-consuming, carrying extra cost, space, and overhead. There is a portability issue of collecting sufficient input X-ray and CT images of COVID-19 patients for the learning process. Ozkaya et al. proposed fusing, and ranking deep features to detect COVID-19 [14]. In this work, they generated two (16×16 and 32×32) sub-datasets of 150 CT images. After improving the performance by deep feature fusion and ranking method, SVM has been applied for classification. Apart from image-based analysis and prediction, some works have been done on the optimistic drug discovery of COVID-19. Edison et al. introduced a Vaxign reverse vaccinology tool and the newly developed Vaxign-ML machine learning tool to predict COVID-19 vaccine candidates [15]. In an article [16], the authors have used network-based toolset to COVID-19 for recovering the primary pulmonary manifestations of the virus in the lung as well as observed comorbidities associated with cardiovascular diseases. They predicted that the virus can manifest to some special tissues such as the reproductive system, and brain regions using network proximity, diffusion, and AI-based metrics techniques. However, similar kind of works on data-driven drug repositioning framework discovery using machine learning and statistical analysis approaches have been proposed in Refs. [17,18]. It helps systematically integrate large-scale knowledge graph, literature, and transcriptome data to discover the potential drug candidates against SARS-CoV-2 [17]. In Ref. [19], a novel combination of machine learning, bioinformatics, and supercomputing has been used to predict antibody structures capable of targeting the SARS-CoV-2 receptor binding domain. The potential result of this article [19] suggests that their predicted antibody mutants may bind to the SARS-CoV-2 RBD and nullify the virus. Batra et al. [18] explored the knowledge of small-molecule treatment against COVID-19. It also provided a pipeline to perform high-throughput computational modeling and ensemble docking simulations for screening of COVID-19 therapeutic agents.

In this article, we have tried to predict the target human proteins of the SARS-CoV-2 virus based on their protein sequences combining amino acid composition, pseudo amino

acid composition, and conjoint triad features using machine learning techniques. The problem has been posed as a two-class classification problem, where the two classes correspond to the interacting and non-interacting proteins of the virus and the host, respectively. This is the first work in this domain as per our knowledge. Initially, we have employed the Learning Vector Quantization (LVQ) technique for feature subset selection as a preprocessing step. Subsequently, after feature reduction, we have used some popular supervised learning algorithms such as Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF) and K-Nearest Neighbor (KNN) along with a deep multi-layer perceptron model and ensemble techniques (Voting classifier, XGBoost, AdaBoost) for classification and prediction. We have used 10-fold cross-validation, repeated 10 times strategy for the supervised learning process. In terms of accuracy perspective on the test dataset, the voting classifier ensemble technique performs better than the other algorithms. Therefore, we have predicted the 1326 new potential human protein targets of the SARS-CoV-2 virus using an ensemble Algorithm. Gene ontology and pathway enrichment for these predicted interactions are investigated. Moreover, we have reported a number of repurposable drugs which target the predicted interactions.

Materials and methods

In this section, we describe the data sets used for our work followed by the methodologies employed.

Datasets

This section describes the different data sets used in this study.

SARS-CoV-2-human PPI database

In [5], Gordon et al. prepared a SARS-CoV-2-human protein–protein interactions (PPIs) database between human proteins and novel coronavirus proteins using affinity-purification mass spectrometry (AP-MS). The database contains 332 unique interactions between 332 human proteins and four structural and as well as 20 accessory coronavirus proteins. The degree distribution of the SARS-CoV-2 proteins in the SARS-CoV-2-human PPI network is given in Fig. 1. These experimentally validated 332 human proteins are used to construct positive training and testing datasets in our study. A summary of the PPI network of SARS-CoV-2-human is shown in Fig. 2 using Cytoscape [20]. The total network is given in Supplementary File S5.

Negative dataset

There is no "gold standard" for planning a negative dataset in the PPI network since non-interacting protein (negative sample) pairs are not experimentally established. Therefore, negative samples must be chosen with care, which may adversely influence the accuracy of predicted PPIs. There are two main sampling methods, namely random pairing, and subcellular localization. Most of the studies have generated non-interacting proteins at random and then eliminates the pairs used in the positive examples [7,21]. Some studies construct the negative samples having different subcellular localization compared to the positive samples [22]. However, both of these methods are not completely reliable. In the case of random sampling, it may incorrectly take a significant number of positive samples as negative samples and produce different accuracy for different random pairing. In Ref. [21] Ben-Hur et al. proved that in PPI prediction, subcellular localization based methods may generate biased accuracy.

We have prepared the negative samples in this article by selecting human proteins from the HPRD database release 9 [23] that are not present in the positive dataset and that have a low degree in the human PPI network. Viral protein pathway-based research showed that they appeared to target higher-degree human proteins rather than lower-degree proteins [24]. With the aid of Cytoscape, we determined the degree of each human protein in the HPRD database and ordered it in ascending order. Then the negative samples are prepared taking into account the minimal degree of human proteins. This degree-based

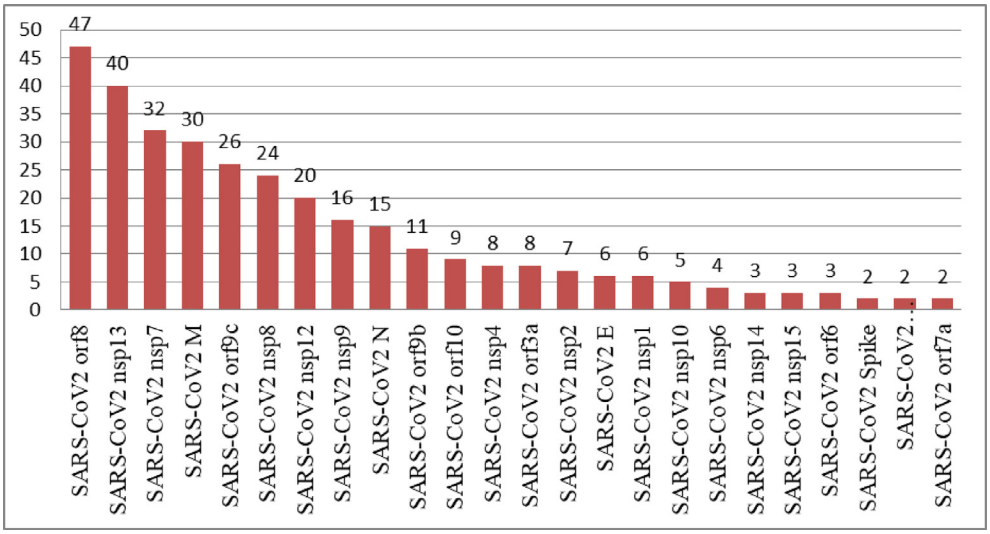

Fig. 1 SARS-CoV-2 proteins' frequency of interactions with human proteins.

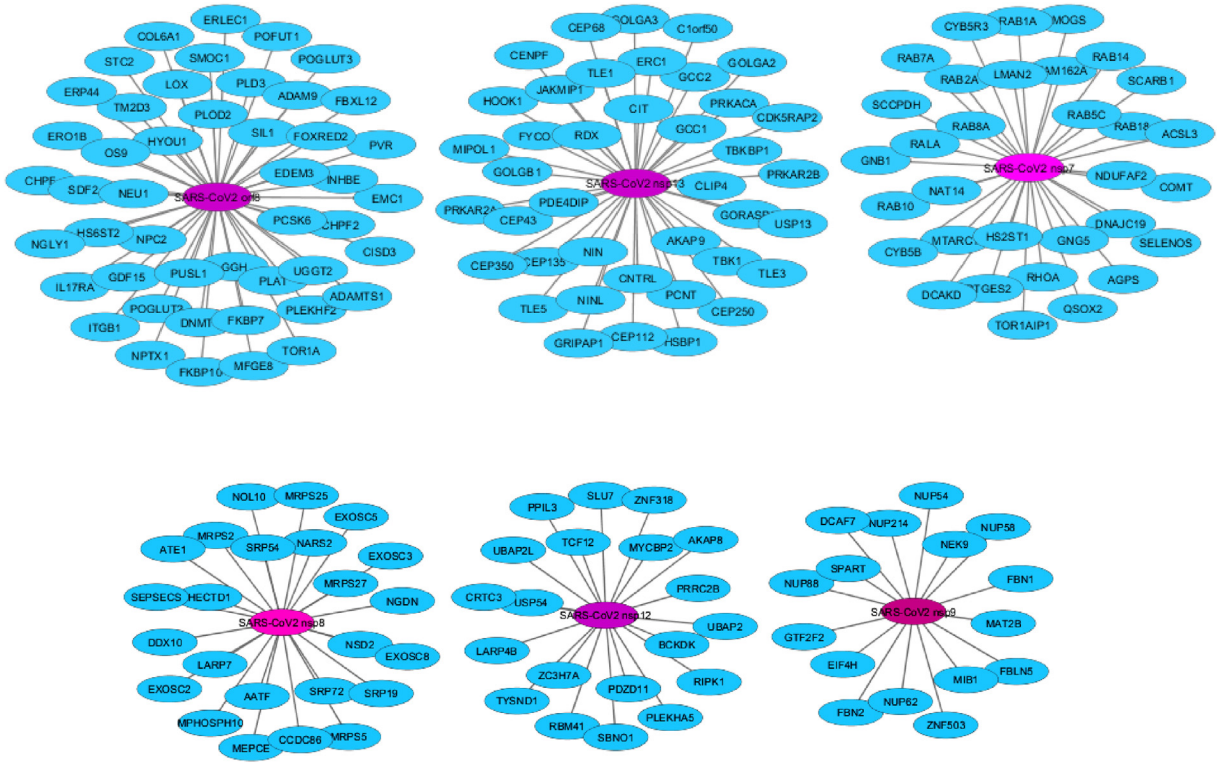

Fig. 2 A glimpse of SARS-CoV-2-human PPI network. Purple ovals indicate SARS-CoV-2 proteins, blue ovals indicate human proteins and edges indicate SARS-CoV-2-human protein interactions.

method of negative sample generation achieves better accuracy on the training dataset compared to the random pairing and subcellular localization [Table 1]. In order to prevent statistical differences, the same scale is assumed for the positive and negative sample, i.e., the ratio 1:1.

Independent dataset

An independent dataset is prepared to predict unknown human proteins that may indulge in protein–protein interaction with corona viral proteins. All the human proteins of the HPRD database (around 9155 proteins) except the proteins that are present in the positive and negative datasets are considered as the independent dataset.

Table 1 Comparison of accuracy of all supervised learning algorithms on 1:1 positive:negative training dataset considering random sampling, subcellular localization and degree distribution of preparing negative samples.

Algorithms	Degree Distributaion Accuracy	Random Sampling Accuracy	Subcellular Localization Accuracy
SVM^{Radial}	68.97	57.76	54.51
SVM^{Linear}	59.16	52.13	54.47
$SVM^{Polynomial}$	67.16	56.06	53.09
KNN	67.09	52.12	58.90
NB	61.38	53.08	54.78
RF	67.28	57.78	55.07
XGBoost	51.53	49.56	52.13
AdaBoost	49.53	56.78	55.67
DMLP (epochs = 50, Batch-Size = 10)	70.91	57.78	64.13

Supervised machine learning algorithms

A set of well-known supervised machine learning algorithms [25], such as SVM, Naive Bayes, Random forest, Deep Learning, KNN, and ensemble techniques are used for predicting PPIs. We have implemented all these algorithms in Python frameworks.

Support Vector Machine (SVM)

Support Vector Machines (SVM) is one of the most powerful supervised learning algorithms which is based on the concept of hyperplane and it is a generalization of a simple and intuitive classifier called the maximal margin classifier. For a given training sample, the Algorithm generates an optimal hyperplane that maximizes the margin between data points of different classes. For a two-class dataset, the nearest objects from each class should be well separated from the decision boundary [26,27]. SVM supports the concept of soft margin classifier because it can be violated by some of the training observations [28]. It can be represented as,

$$\text{Maximize}_{\beta 0,\beta 1,\ldots,\beta P} M, \sum_{i=1}^{P} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1-\varepsilon_i), \varepsilon_i \geq 0, \sum_{i=1}^{n} \varepsilon_i \leq C. \quad (1)$$

In equation (1), C is a nonnegative tuning parameter, M is the width of the margin and the optimization problem

chooses β_0, β_1, …, β_p to maximize M. ε_0, ε_1, …, ε_n are slack variables that allow individual observations to be on the wrong side of the margin or the hyperplane. x* are the observations [28].

K-Nearest Neighbor(KNN)

K-Nearest Neighbor (KNN) Algorithm is one of the simplest supervised learning algorithm used for both classification and regression problems. It is also called as the lazy learner as it uses all the data for training while doing prediction [26,27]. We have to choose the value of K first ($K \leq \sqrt{n}$), where n is the size of the data. Then it calculates the distance between test data and each row of training data with the help of any distance metric function (use Hamming distance for categorical data). Now, it sorts them in ascending order according to the calculated distance values. Then top K rows from the sorted array are chosen and finally, it will assign a class to the test point based on the most frequent class of these rows (Majority of voting) [29].

Naive Bayes (NB)

Naive Bayes is a probabilistic classifier, based on Bayes' theorem. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. So, each feature makes an independent and equal contribution to the outcome. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

In equation (2), $P(A|B)$ is the posterior probability, $P(B|A)$ is the maximum likelihood and $P(A)$ is the prior probability. Basically, we are trying to find the probability of event A, given the event B is true. Event B is also termed as evidence. In Naive Bayes perspective, $P(y|X)$, y is class variable and X is a dependent feature vector (of size n) where, $X = (x_1, x_2, x_3, \ldots\ldots, x_n)$. Therefore, we can rewrite the Naive Bayes equation for a set of independent features as,

$$P(y|x_1, x_2, x_3, \ldots, x_n) = \frac{P(x_1|y)P(x_2|y)\ldots P(x_n|y)P(y)}{P(x_1)P(x_2)\ldots P(x_n)} \tag{3}$$

Now, we need to create a classifier model. For this, we find the probability of a given set of inputs for all possible values of the class variable y and pick up the output with maximum probability. This can be expressed mathematically as,

$$y = \underset{y}{\operatorname{argmax}}\, P(y)\Pi_{i=1}^{n} P(x_1|y) \tag{4}$$

In equation (4), $P(y)$ is also called class probability and $P(x_i|y)$ is called conditional probability. Naive Bayes has been successfully used to predict the protein–protein interactions and binding sites of DNA/RNA. For PPI, the Algorithm generates binary classification output based on the protein sequence vector [26,28].

Ensemble techniques

Ensemble modeling is a powerful way to improve the performance of the model. In order to improve the performance of the model, it combines several base models into an optimal predictive model. There are two kinds of ensemble techniques Bagging and Boosting.

- In Bagging (Bootstrapping and Aggregation) technique, we create multiple bootstrapped subsamples (row sampling with replacement) from the original one and apply the Decision Tree learning model on each of the bootstrapped subsamples. After each subsample Decision Tree has been formed, an Algorithm is used to aggregate over the Decision Trees to form the most efficient predictor. In order to choose the most efficient predictor, it follows the majority of the voting classifier technique. This Bagging concept is also applicable to various sets of classifier models. In our work, we have applied the majority voting classification technique on combining SVM-polynomial, SVM-radial and RF models to achieve the best accuracy among all models [28,29]. A single decision tree classification gives low bias and high variance whereas a Random Forest classifier gives low bias and low variance. This bagging algorithm is also known as the Random Forest classifier.
- Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. The Adaptive boosting (AdaBoost) Algorithm helps us to combine multiple "weak classifiers" into a single "strong classifier". The weak learners in AdaBoost are decision trees with a single split, called decision stumps. AdaBoost works by putting more weight on difficult to classify instances and less on those already handled well. It is used for both classification and regression purposes [28,29].
- The Extreme Gradient Boosting (XGBoost) Algorithm is an implementation of gradient boosted decision trees designed for speed and performance. It has a wide range of applications, portable, flexible implementation and cloud integration of this model is easy. XGBoost is an ensemble tree method that applies the principle of boosting weak learners (CARTs generally) using the gradient descent architecture [28]. XGBoost algorithm provides high bias and low variance [29].

Deep multi-layered perceptron (DMLP)

MLP is a multiple feedforward artificial neural network that maps input vectors to output vectors [30]. It can be represented by a directed graph with multiple node layers, where the bottom and top layers are input and output layers respectively and others are hidden layers. DMLP is a fully connected network where every node in the upper level has connections with all the nodes in the lower level. Each node represents a neuron (or processing unit) with a nonlinear activation function except for input layer nodes. Multiple hidden layers are allowed that makes it deep neural [31]. In our work, we have used an adaptive learning rate optimization Algorithm which is designed specifically for training our deep neural network [32]. We have created a five hidden layers MLP network along with the ReLU (Rectified Linear Unit) activation function and 'sigmoid' function at the output layer. We have fitted our model with varying epochs, varying batch-sizes, and binary-cross-entropy as a loss function.

Sequence-based features

A total of 3 sets of sequence-based features, namely, amino acid composition, conjoint triad, and pseudo amino acid composition of the human proteins are considered to train the machine learning models. The FASTA sequences of human proteins are gathered from UniProt and the values of these features are extracted from protr [33]. These 3 feature sets are described below.

Amino acid composition (AAC)
The composition of amino acids explains the percentage of a type of amino acid found in a protein chain. This is one of the simple and effective predictive functions of PPIs. The arrangement of amino acids reflects only the abundance in a sequence of each amino acid [26].

Conjoint triad (CT)
Many studies have used conjoint triad to completely explain the essential PPI details to represent the properties of amino acid [34,35]. In the conjoint triad system, the 20 amino acids are divided into seven classes based on their amounts of dipoles and side chains [Table 2]. Every amino acid of a protein chain is then replaced by the number of clusters.

Pseudo-amino-acid composition (PseAAC)
Chou first proposed the composition of pseudo-amino acids in 2001 [36]. AAC only displays the frequency of each amino acid in a sequence, but information on sequence order is lost. Compared to AAC, PseAAC finds protein sequence order along with amino acid compositions [37,38]. The order of sequence can be extracted from sequence similarity variables such as hydrophobicity, hydrophilicity, and side-chain mass.

Feature selection

The performance of prediction depends on two aspects. They are feature extraction and performance of the classifier. Feature selection is one of the important tasks in classification algorithms. The datasets used for classification contain a large number of features. Most of them are either partially or completely irrelevant and redundant to the classifier. Therefore, feature selection is used to select a number of important features so that it can achieve acceptable classification accuracy compared to considering all features. In this work, we have used a well-known Kohonen's Learning Vector Quantization (LVQ) technique to identify important features and remove redundant features from the original dataset [39]. LVQ is a simple, universal, and efficient learning classifier. It is very similar to the k-Nearest Neighbors. In an N-dimensional feature space, LVQ tries to approximate optimal decision borders between different classes with a number of labeled codebook vectors. The Euclidean distance measure helps to label the closest codebook vector of an example vector. A feature selection strategy using LVQ is evaluated to optimize the hypothesis margin of LVQ classification through minimizing its loss function [40] and it generates importance score of each feature based on the class identification.

Table 2 The seven clusters of amino acids based on their dipoles and side-chain volumes.

Cluster number	Protein groups
Cluster 1	A, G, V
Cluster 2	I, L, F, P
Cluster 3	Y, M, T, S
Cluster 4	H, N, Q, W
Cluster 5	R, K
Cluster 6	D, E
Cluster 7	C

Methodology

Our proposed methodology is represented in terms of algorithmic steps below.

Algorithm. Steps of Proposed PPI Prediction Methodology between SARS-CoV-2 Virus and Human

Input:

I_1. Collect the human proteins that interact with SARS-CoV-2 virus [5] (Positive dataset).

I_2. Collect non-interacting human proteins from HPRD using the concept of degree distribution (Negative dataset).

Output: Predicted potential human target proteins.

Procedure:

1. Combine I_1 and I_2 to construct the training dataset.
2. Construct the feature vector using AAC, CT, and PseAAC.
3. Collect reduced feature-subset using Learning Vector Quantization (LVQ) feature selection technique.
4. Apply SVM, RF, KNN, NB and DMLP classifiers for training the model and do comparative analysis for the prediction of new sets of interactions.
5. Identify target human proteins using majority voting-based ensemble classifier.
6. Assess the predictions using literature search, Gene Ontology (GO) enrichment, KEGG pathway analysis.
7. Predict repurposable drugs targeting the identified human proteins.

The pictorial representation of the PPI prediction methodology has been shown in Fig. 3.

Performance measures

The performance of each classifier is evaluated with 10-fold cross-validation in 10 repeat runs to acquire average values. Different measurements like accuracy, Kappa, sensitivity, specificity, precision, and F1 score are calculated considering 1:1 positive and negative datasets. These parameters are defined as per equations (5)–(10).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} * 100\% \quad (5)$$

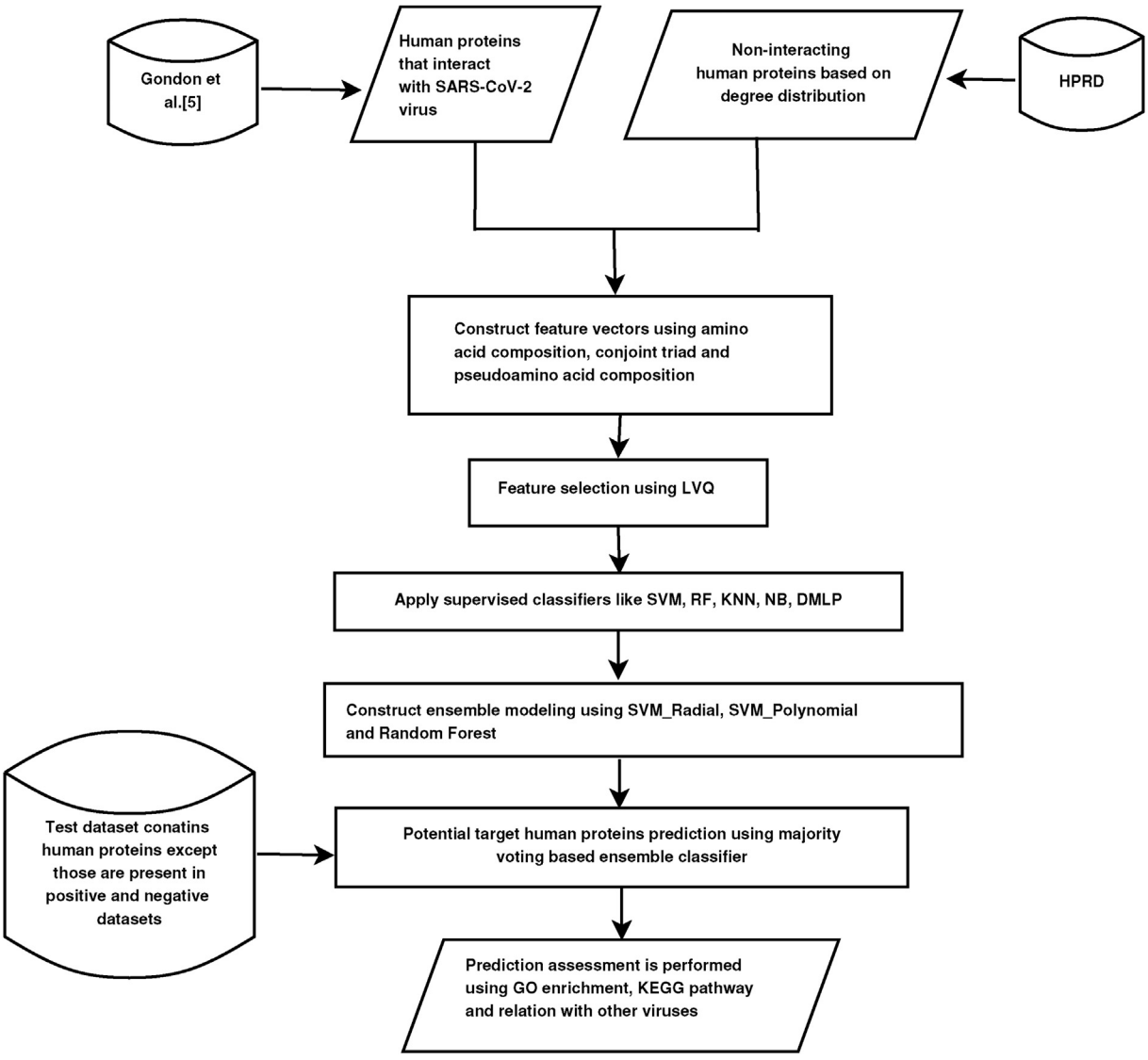

Fig. 3 Block diagram of protein–protein interaction prediction methodology.

$$Kappa = \frac{Observed\ accuracy - Expected\ accuracy}{1 - Expected\ accuracy}, \quad (6)$$

where,

$$Expected\ accuracy = \frac{((TN+FP)*(TN+FN)) + ((FN+TP)*(FP+TP))}{(TP+FP+TN+FN)*(TP+FP+TN+FN)}$$

$$Specifity = \frac{TN}{FP+TN}*100\%, \quad (7)$$

$$Precision = \frac{TP}{TP+FP}*100\%, \quad (8)$$

$$Recall = \frac{TP}{TP+FN}*100\%, \quad (9)$$

$$F1\ score = \frac{2*(Recall*Precision)}{(Recall+Precision)}*100\% \quad (10)$$

where, TP, FP, FN, and TN respectively denote the numbers of true positives, false positives, false negatives, and true negatives.

Results

Performance of the classifiers

In this work, we have started with 3 types of features (amino acid composition, conjoint triad, and pseudo amino acid composition) of human proteins which results in a 413-dimensional feature vector. The importance of all these features are calculated using LVQ. Afterthat, the knee point is evaluated to detect a large change in the importance score and extract the most important features. We have extracted 38 significant features from 413 features of the original dataset by applying the procedure.

Table 3 shows the comparison of performance between selected best 38 features vs all 413 features using all the used

Table 3 Comparison of cross-validation performance between all features vs selected best 38 features using all supervised learning algorithms on 1:1 positive and negative training dataset. The best accuracy values for each classifier are highlighted in boldface.

Method	All features		Selected features	
	Accuracy	Kappa	Accuracy	Kappa
SVM^{Radial}	68.97	36.93	**69.45**	36.90
SVM^{Linear}	59.16	23.34	**65.86**	31.79
$SVM^{Polynomial}$	67.16	35.93	**67.78**	34.58
KNN	**67.09**	34.19	59.27	28.53
NB	61.38	29.88	**62.23**	30.91
RF	67.28	34.50	**68.72**	36.90
XGBoost	51.53	23.01	**59.13**	28.23
AdaBoost	49.53	20.31	**53.84**	24.02
DMLP(epochs = 50,Batch-Size = 10)	68.91	36.61	**70.51**	38.72

supervised learning algorithms. All these values have been calculated by 10 fold cross-validation repeated 10 times and then the average of them is reported. In case of DMLP classifier, the batch size of 32 or 25 is acceptable, with epochs = 100 unless the dataset is very large. For large datasets, a batch size of 10 with epochs between 50 and 100 can be considered. In our case, our COVID-19 PPI dataset is quite large in terms of the number of features (413-dimensional feature vectors) for the training dataset. Therefore, we have used epochs = 50 and batch-size = 10. After the feature selection, we kept the same set of epochs and batch-size for showing parity in the result analysis. It can be seen from the Table that these 38 important features achieved higher accuracy than considering all the features used together for all classifiers except KNN. KNN classification averages the labels of *K*-Nearest neighbor samples to come to a decision. However, when the number of neighbors reduces due to feature selection, it reduces the accuracy sometimes.

Performance of the classifiers using blind datasets

The blind dataset is used to prevent bias of the classifiers. 124 proteins (62 interacting + 62 non-interacting) from the training dataset are treated as test datasets. The rest of the proteins are used as the training dataset. The performance of the classifier models using a blind dataset is given in Table 4. It can be seen from the table that SVM^{Radial}, $SVM^{Polynomial}$, and Random Forest method achieved better accuracy, specificity, and F1 score over other classifiers. Although, DMLP gives better accuracy on the training datasets, on the test dataset it can't do well compared to the other models. Therefore, to increase the prediction accuracy and minimize the false positives we have built a majority voting based Ensemble model using these three methods SVM^{Radial}, $SVM^{Polynomial}$ and Random Forest) to predict unknown SARS-CoV-2 target proteins of human. It can be seen that the ensemble method achieved better accuracy (72.33%), recall (71.67%), specificity (74.41%), precision (72.41%), and F1 score (72.03%) than all the other classifier models.

Prediction of potential human protein targets of SARS-CoV-2

On both positive and negative 1:1 dataset we measured AAC, conjoint triad, and PseAAC features. This dataset is used as the training dataset. All the human proteins of the HPRD database, which are not present in the positive and negative dataset, are considered for prediction analysis. We have applied the Ensemble method on these large numbers of human proteins (9155) of the HPRD dataset and predicted 3603 potential interacting human proteins. The prediction results are listed in Supplementary File S1 along with their average probability prediction score. However, standard SVM and Random Forest assume that the probability threshold for both classes is equal, i.e., 0.5 for binary classification problem. In this work, we have changed the probability threshold from 0.5 to 0.7 so that we can predict all high probability human target proteins and avoid possible false positives. A higher threshold indicates higher confidence in predicting the positive class. The average probability of the predicted human proteins varies from 0.926 to 0.461. Our motivation is to predict high-throughput target human proteins. If we consider the threshold of the average probability 0.9, then we get only 10 target human proteins. For probability factor 0.8, we obtain 340 proteins. Therefore, we have set the probability value to 0.7, so that a reasonable number of target human proteins can

Table 4 Comparison of performance of all supervised learning algorithms on blind dataset.

Algorithms	Accuracy	Recall	Specificity	Precision	F1-Score
SVM^{Radial}	69.67	58.06	73.33	62.85	67.68
SVM^{Linear}	63.93	58.06	70	61.76	65.63
$SVM^{Polynomial}$	68.03	56.64	80	64	70
KNN	64.17	66.13	56.67	61.81	59.12
NB	65.03	65	56.45	66	65.18
RF	68.93	66.13	70	66.67	68.29
XGBoost	61.2	63	55.23	61	63.29
AdaBoost	54.3	59.84	60	60	60.17
DMLP(epochs = 50,Batch-Size = 10)	63.47	60	57.9	61.01	60.53
Ensemble Technique	**72.33**	**71.67**	**74.41**	**72.41**	**72.03**

Table 5 The top 10 high-degree predicted target human proteins with their degrees and average prediction scores.

Protein Name	Degree	Average Prediction Score
THEM4	168	0.707725
OMG	156	0.724195
RTN4RL1	154	0.772267
ANXA4	135	0.846931
TTC3	131	0.723774
MYO1A	130	0.749535
TEX10	88	0.797231
NOSIP	81	0.755477
PCSK1N	76	0.711644
MYH11	74	0.841619

be predicted. Using this procedure, the potential number of target human proteins comes down to 1326 from 3603 (File S2). We have also calculated the degree of these predicted human proteins with respect to all the proteins of the HPRD database using Cytoscape and included in the supplementary files. The degree value varies from 1 to 168. The top 10 high-degree proteins with the degree and prediction score are listed in Table 5.

Discussion

Gene ontology (GO) term enrichment

GO is one of the most used annotation systems to obtain the biological relevance of high-throughput experiments. To examine the functional characteristics of these predicted 1326 high-probability human target proteins, GO term analysis is done on the biological process (BP), cellular component (CC), and molecular function (MF) categories. It has been noticed that the proteins that have similar cell locations or involved in some biological process or molecular function, are likely to interact with each other. We have collected GO annotations of all predicted human proteins from DAVID 6.8 [41] having corrected p-values less than 0.05 to validate the predictions. We have found that some of the most enriched GO-CC terms of the predicted human proteins are cytosol, extracellular exosome, cytoplasm, membrane, and nucleoplasm. The GO-CC term describes the different locations of a cell, at the levels of subcellular structures and macromolecular complexes. Being an RNA-based virus, the novel coronavirus attacks either the nucleus or cytoplasm of the host cell and causes a respiratory block in the lungs. Therefore, the proteins involving in these CC terms are likely to act as potential coronavirus targets.

Antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent, NIK/NF-kappaB signaling, regulation of the cellular amino acid metabolic process, positive regulation of ubiquitin-protein ligase activity involved in the regulation of mitotic cell cycle transition, negative regulation of ubiquitin-protein ligase activity involved in the mitotic cell cycle, etc. are five most enriched GO-BP terms collected from the DAVID server. Some of the most enriched GO-MF terms are like protein binding, ATP binding, GTP binding, cadherin binding involved in cell–cell adhesion, poly(A) RNA binding, etc. In Ref. [42], the authors examined that the coronavirus proteins especially the spike proteins bind with target human proteins and increase cell adhesion during infection. All the enriched GO terms along with involved human proteins and p-values are listed in Supplementary File S3.

Table 6 The significant KEGG pathways of the predicted human proteins.

KEGG Pathway	Protein Count	Predicted Human Proteins
Proteasome (p = 2.3615E-8)	19	PSMB10, SHFM1, PSMB8, PSMA2, PSMB4, PSMC5, PSMD12, PSMA6, PSMB1, PSMC4, PSMA5, PSMC3, PSMC2, PSMD1, PSMC1, PSMB2, POMP, PSMD4, PSMD7
Endocytosis (p = 2.5126E-7)	49	LDLR, CHMP4B, TSG101, CHMP5, CAPZA2, CHMP6, PIP5K1C, CLTC, SMAP1, PIP5KL1, VPS4B, SPG21, KIF5B, KIF5A, RAB4A, HLA-A, HLA-B, HLA-E, HLA-F, ARPC1A, ARPC1B, RAB11FIP5, RAB11FIP3, CHMP1B, ACAP3, ACAP1, RAB5A, SH3GL1, VPS29, SNX5, SNX2, SNX1, ARPC4, HSPA1A, SNX4, ARPC5, ARFGEF2, CHMP2B, SH3GLB1, RAB11A, EHD1, EHD2, RAB31, ARF1, RAB35, ARF3, RAB22A, VPS28, DNM1
Biosynthesis of antibiotics (p = 7.0351E-5)	44	LDOA, HSD17B10, LDHB, LDHA, ADPGK, PGAM1, HK2, HK1, ASL, AGXT, PDHB, FDFT1, GOT1, IDH3G, HK3, ENO2, IDH2, GCSH, IDH1, ENO3, PDHA2, CAT, RPIA, PDHA1, HADH, ENO1, SHMT1, PFKL, AK1, SUCLG1, FDPS, IDH3B, ACLY, PFKM, IDH3A, NME5, ALDH7A1, PYCR2, NME2, PKLR, MVK, PRPS2, CBS, PRPS1
Carbon metabolism (p = 1.02674E-6)	29	ALDOA, ADPGK, GLUD1, PGAM1, HK2, HK1, AGXT, PDHB, GOT1, IDH3G, HK3, IDH2, ENO2, IDH1, ENO3, PDHA2, CAT, RPIA, PDHA1, ENO1, SHMT1, PFKL, SUCLG1, IDH3B, PFKM, IDH3A, PKLR, PRPS2, PRPS1
Biosynthesis of amino acids (p = 5.6641E-6)	21	ALDOA, SHMT1, PFKL, PGAM1, IDH3B, PFKM, ASL, IDH3A, PYCR2, GOT1, IDH3G, PKLR, ENO2, IDH2, ENO3, IDH1, RPIA, PRPS2, CBS, ENO1, PRPS1
Glycolysis/Gluconeogenesis (p = 6.9568E-6)	20	ALDOA, LDHB, LDHA, PFKL, ADPGK, HK2, PGAM1, HK1, PFKM, PDHB, ALDH3A1, G6PC, ALDH7A1, HK3, PKLR, ENO2, ENO3, PDHA2, PDHA1, ENO1
Central carbon metabolism in cancer (p = 6.3733E-4)	16	PFKL, MET, HK2, PGAM1, RAF1, HK1, SIRT6, PFKM, PDHB, SLC16A3, SLC1A5, HK3, PDGFRB, PDHA2, MTOR, PDHA1

Table 7 The predicted human proteins that interact with the proteins of other viruses.

Virus	Number of overlapping proteins	Database Name	Number of human proteins present in the database	Reference
Dengue	174	DenvInt	480	[46]
HIV-1	1290	HIV-1 Human Interaction Database	4667	[47]
HCV	144	HCVpro	467	[48]
Ebola	16	Zhou et al.	60	[49]
Zika	5	ZikaBase	24	[50]
H1N1	160	Shapira et al.	617	[51]

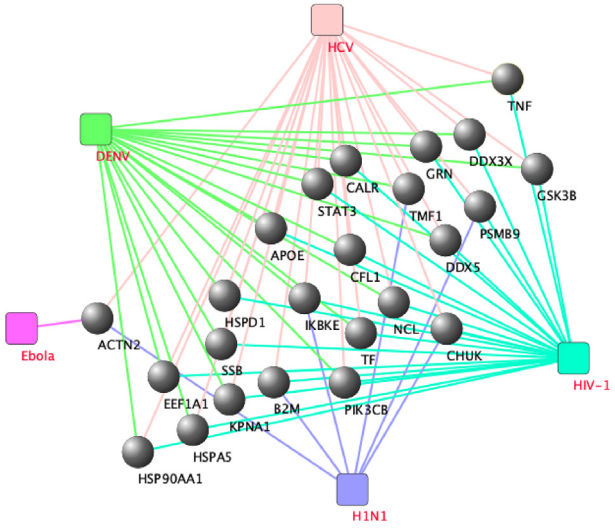

Fig. 4 The diagrammatic representation of the predicted target human proteins that interact with the proteins of multiple viruses. Black circles represent human proteins. Square boxes represent different viruses. Interactions of human proteins with different viruses are represented as edges colored as per the color of the boxes of respective viruses.

KEGG pathway analysis

Analysis of the KEGG pathway shows potential illnesses that can develop in the human body due to COVID-19 infection. All the significant pathways of 1326 predicted target human proteins along with the corrected p-values are listed in Table 6.

Viruses are well known to be exploiting the machinery of host cells for their own replication. The proteasome pathway is one of these intracellular processes that are hijacked by the viruses [43]. Myung et al. showed this proteasome pathway is involved with different viruses like retrovirus, human immunodeficiency virus type 1 (HIV-1), simian immunodeficiency virus (SIV), and Moloney murine leukemia virus (Mo-MuLV). Therefore, the probability of association of this pathway with novel coronavirus is very high. Endocytosis is a biological process of transporting particles, such as large molecules, parts of cells, and even whole cells, to bring into a cell. Experiments on the SARS-CoV-2 virus have shown that the endocytic pathway is the main pathways for regulating the entrance of CoVs into the host cells and thus the endocytic pathway, as well as the involved human proteins, can be extensively studied for the target of anti-viral therapies [44]. In Ref. [45], Wenzhong Liu et al. showed that because of the failure to regularly combine carbon dioxide and oxygen during SARS-CoV-2 virus infection, the lung cells have highly severe toxicity and inflammation, which ultimately results in ground-glass pulmonary images. According to our results, novel coronavirus can alter the synthesis of macromolecules and its growth rate and can cause carbon metabolism. The term biosynthetic pathway suggests manufacturing the antibiotics which can help to combat drug-resistant viruses and diseases. One of the main public health issues of recent times is the exponential growth of antibiotic-resistant pathogens. Therefore, the predicted human proteins involving this pathway need to be further studied to develop antibiotics for the SARS-CoV-2 virus. The amino acid plays an important role in the expression of the different viral functions including synthesis of viral coat proteins and the development of full infectious virions. Novel coronavirus may alter this amino acid composition and cause several other diseases in the human body. We can conclude from the above pieces of evidence that the predicted human proteins involving these pathways are strongly involved in coronavirus infection and experimental validation is required to establish direct interactions.

Interaction between predicted human proteins and other viruses

In this section, we have tried to find out the relation between our predicted 3603 human proteins with the other viruses. Our research included six specific human-pathogenic RNA viruses, namely, Dengue, HIV-1, HCV, Ebola, Zika, and H1N1. We have found various pieces of evidence that many predicted proteins interact with different medically important viruses from published literature. Table 7 shows the number of overlapping predicted human proteins and experimentally verified human proteins with other viruses. We found a large number of predicted human proteins interact with more than one virus. For example, human protein IKBKE has the highest degree of 4. It interacts with four viruses, namely, Dengue virus (DENV), HIV-1, HCV, and H1N1. A small PPI network of the predicted target humans that interact with at least three viruses is shown in Fig. 4. From Table 7 it can be seen that a majority of the predicted human proteins of the SARS-CoV-2 virus overlaps with the HIV-1 virus. Recently, a study

Table 8 List of drugs associated with the predicted target human proteins that interact with the proteins of at least 3 different viruses.

Sl. No.	Drugs Name	Human Protein Name
1.	Remicade, Etanercept, Adalimumab, Thalidomide, Inamrinone, Golimumab, Certolizumab Pegol, Chloroquine, Glucosamine, Clenbuterol	TNF
2.	Atorvastatin, Cetrorelix	HSPD1
3.	Melatonin, Tretinoin, Gentamicin, Tenecteplase	CALR
4.	Aspirin, Fluorouracil	HSPA5
5.	Amlexanox, Procaine	IKBKE
6.	Thyroglobulin, Amikacin, Pembrolizumab	B2M
7.	Carfilzomib, Bortezomib, Ixazomib Citrate	PSMB9
8.	Rifabutin	HSP90AA1
9.	Lovastatin, Zinc Sulfate, Doxorubicin, Prasterone, Progesterone, Octreotide, Epinephrine, Dactinomycin, Nandrolone Phenpropionate, Candicidin	PIK3CB
10.	Lithium Citrate Hydrate, Lithium Carbonate, Fluoxetine	GSK3B
11.	Albumin Human, Prednisone, Tretinoin, Ganciclovir, Triamcinolone, Irbesartan, Vitamin E, Lorazepam, Soybean Oil, Gonadotropin, Chorionic	APOE
12.	Mesalamine, Aminosalicylic Acid, Sulfasalazine, Acetylcysteine, Ascorbate	CHUK

revealed that the human protein called RBBP6 aids in the fight against Ebola by interfering with its replication cycle. This RBBP6 along with 16 other predicted human target proteins overlaps with ebola related human proteins. The rest of the human protein's name and their association with the different viruses are listed in Supplementary File S4.

No effective drug has yet been discovered targeting SARS-CoV-2 virus. The traditional mechanism for drug development and its approval is much more expensive and time-consuming. On the other hand, repurposing the existing drugs is an alternative method for identification of effective drugs. It can substantially shorten the time and minimize costs relative to new drug development. The human proteins that interact with the proteins of multiple viruses can be good candidate for as targets for drug repurposing. We have found 30 predicted human proteins that interact with the proteins of at least 3 viruses. The U.S. Food and Drug Administration (FDA) approved drugs that interact with these human proteins are queried using DGIdb (http://dgidb.org/) and reported in Table 8. Most of these drugs are used in cancer-inhibiting microtubules treatment, and other drug classes are used for diseases like Atherosclerosis, Crohn's disease, antiinflammatory agent, tumor necrosis factor, blood circulation and infection.

Conclusion

In this study, various sequence-based features and computational approaches are used for the first time in predicting the potential human targets of the SARS-CoV-2 virus. We have used the LVQ Algorithm for feature selection. HIV, influenza, and other viruses are well researched in various literature works compared to coronavirus proteins as SARS-CoV-2 which has emerged recently. HIV, the most well-studied virus, is believed to have around 1000 direct interactions with human proteins and 3000 indirect interactions. Both HIV-1 and SARS-CoV-2 are enveloped RNA viruses. The SARS-CoV-2 virus contains a large number of proteins, even more than HIV-1. Therefore, it can be assumed that SARS-CoV-2 is a virus expected to have a large number of interactions with human proteins. However, due to the lack of experimentally validated SARS-CoV-2-human PPIs, most of the possible interactions are currently unknown. In this paper, we have predicted the 1326 potential target human proteins considering a high probability factor (70%) using ensemble modeling. We have also analyzed the GO terms and KEGG pathway of these predicted human proteins and some of the pathways are also supported by recent literature. Some repurposable drugs that target the predicted interactions have also been reported. We hope that this study will help biologists recognize possible associations between novel coronavirus and human proteins and facilitate the development of anti-viral drugs.

Conflicts of interest

The authors declare that they have no conflict of interest.

Acknowledgement

Anirban Mukhopadhyay acknowledges the support received from the research project grant (Memo No: 355(Sanc.)/ST/P/S&T/6G-10/2018 dt. 08/03/2019) of Department of Science & Technology and Biotechnology, Government of West Bengal, India.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.bj.2020.08.003.

REFERENCES

[1] Ucar F, Korkmaz D. COVIDiagnosis-Net: deep Bayes-SqueezeNet based diagnostic of the coronavirus disease 2019 (COVID-19) from X-Ray images. Med Hypotheses 2020;140:109761.

[2] Ibrahim IM, Abdelmalek DH, Elshahat ME, Elfiky AA. COVID-19 spike-host cell receptor GRP78 binding site prediction. J Infect 2020;80:554–62.

[3] Kassani SH, Kassasni PH, Wesolowski MJ, Schneider KA, Deters R. Automatic detection of coronavirus disease (COVID-19) in X-ray and CT images: a machine learning-based approach. 2020 [Preprint]. 2020 [cited 2020 April 22]. arXiv:200410641.

[4] Kumar R, Arora R, Bansal V, Sahayasheela VJ, Buckchash H, Imran J, et al. Accurate prediction of COVID-19 using chest X-Ray images through deep feature learning model with SMOTE and machine learning classifiers. 2020. medRxiv 2020.04.13.20063461 [Preprint]. 2020 [cited 2020 April 17].

[5] Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 2020 April 17;583:459–68.

[6] Bandyopadhyay S, Ray S, Mukhopadhyay A, Maulik U. A review of in silico approaches for analysis and prediction of HIV-1-human protein–protein interactions. Briefings Bioinf 2015;16:830–51.

[7] Barman RK, Saha S, Das S. Prediction of interactions between viral and host proteins using supervised machine learning methods. PloS One 2014;9:e112034.

[8] Sen R, Nayak L, De RK. A review on host–pathogen interactions: classification and prediction. Eur J Clin Microbiol Infect Dis 2016;35:1581–99.

[9] Bullock J, Luccioni A, Pham KH, Lam CSN, Luengo-Oroz M. Mapping the landscape of artificial intelligence applications against COVID-19. [Preprint] 2020 [cited 2020 April 23]. arXiv:200311336.

[10] Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U, et al. COVID-19 outbreak prediction with machine learning. 2020 April 19. Available at SSRN 3580188.

[11] Barstugan M, Ozkaya U, Ozturk S. Coronavirus (covid-19) classification using CT images by machine learning methods. [Preprint] 2020 [cited 2020 March 20] arXiv:200309424.

[12] Horry MJ, Paul M, Ulhaq A, Pradhan B, Saha M, Shukla N, et al. X-Ray image based COVID-19 detection using pre-trained deep learning models. engrXiv 2020 April 22.

[13] Wang L, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images. [Preprint] 2020 [cited 2020 May 11]. arXiv:200309871.

[14] Ozkaya U, Ozturk S, Barstugan M. Coronavirus (COVID-19) classification using deep features fusion and ranking technique. [Preprint] 2020 [cited 2020 April 7]. arXiv:200403698.

[15] Ong E, Wong MU, Huffman A, He Y. COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. Front Immunol 2020;11:1581.

[16] Gysi DM, Valle ÍD, Zitnik M, Ameli A, Gan X, Varol O, et al. Network medicine framework for identifying drug repurposing opportunities for covid-19. [Preprint] 2020 [cited 2020 Aug. 9]. arXiv:200407229.

[17] Ge Y, Tian T, Huang S, Wan F, Li J, Li S, et al. A data-driven drug repositioning framework discovered a potential therapeutic agent targeting COVID-19. [Preprint] 2020 [cited 2020 March 12]. bioRxiv 2020.03.11.986836.

[18] Batra R, Chan H, Kamath G, Ramprasad R, Cherukara MJ, Sankaranarayanan S. Screening of therapeutic agents for COVID-19 using machine learning and ensemble docking simulations. J Phys Chem Lett 2020;11:7058–65.

[19] Desautels T, Zemla A, Lau E, Franco M, Faissol D. Rapid in silico design of antibodies targeting SARS-CoV-2 using machine learning and supercomputing. [Preprint] 2020 [cited 2020 April 10]. bioRxiv 2020.04.03.024885.

[20] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–504.

[21] Ben-Hur A, Noble WS, BioMed Central. Choosing negative examples for the prediction of protein-protein interactions. BMC Bioinf 2006;7:S2.

[22] Sun T, Zhou B, Lai L, Pei J. Sequence-based prediction of protein protein interaction using a deep-learning Algorithm. BMC Bioinf 2017;18:277.

[23] Keshava Prasad T, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human protein reference database—2009 update. Nucleic Acids Res 2008;37:D767–72.

[24] Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. Genome Biol 2010;11:R53.

[25] Sen PC, Hajra M, Ghosh M. Supervised classification algorithms in machine learning: a survey and review. Emerging Technology in modelling and graphics. Springer; 2020. p. 99–111.

[26] Dey L, Mukhopadhyay A. A classification-based Approach to prediction of Dengue virus and Human protein-protein interactions using amino Acid composition and conjoint triad features. 2019 IEEE Region 10 Symposium (TENSYMP), Kolkata, India. 2019. p. 373–8.

[27] Dey L, Chakraborty S, Biswas A, Bose B, Tiwari S. Sentiment analysis of review datasets using naive bayes and k-nn classifier. IJIEEB 2016;8:54–62.

[28] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media; 2009.

[29] Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 2012;42:463–84.

[30] Amid S, Mesri Gundoshmian T. Prediction of output energies for broiler production using linear regression, ANN (MLP, RBF), and ANFIS models. Environ Prog Sustain Energy 2017;36:577–85.

[31] Wan S, Liang Y, Zhang Y, Guizani M. Deep multi-layer perceptron classifier for behavior analysis to estimate Parkinson's disease severity using smartphones. IEEE Access 2018;6:36825–33.

[32] Heidari AA, Faris H, Mirjalili S, Aljarah I, Mafarja M. Ant lion optimizer: theory, literature review, and application in multi-layer perceptron neural networks. In: Nature-inspired optimizers. Springer; 2020. p. 23–46.

[33] Xiao N, Cao DS, Zhu MF, Xu QS. Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics 2015;31:1857–9.

[34] Wang H, Hu X. Accurate prediction of nuclear receptors with conjoint triad feature. BMC Bioinf 2015;16:402.

[35] Basit AH, Abbasi WA, Asif A, Minhas FUAA. Training large margin host-pathogen protein-protein interaction predictors. J Bioinf Comput Bio 2018;16:1850014.

[36] Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 2001;43:246–55.

[37] Tang Y, Wei C, Sumikoshi K, Nakamura S, Terada T, Kadota K, et al. Predicting protein–protein interactions using sequence homology and machine-learning methods. Res J Life Sci Bioinformatics, Pharm Chem Sci. 2017;3:1–26.
[38] Roy S, Martinez D, Platero H, Lane T, Werner-Washburne M. Exploiting amino acid composition for predicting protein-protein interactions. PloS One 2009;4:e7813.
[39] Pregenzer M, Pfurtscheller G, Flotzinger D. Automated feature selection with a distinction sensitive learning vector quantizer. Neurocomputing 1996;11:19–29.
[40] Hu Y, Liu W. A novel feature selection algorithm based on LVQ hypothesis margin. Neural Comput Appl 2014;24:1431–9.
[41] Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 2003;4:R60.
[42] Qing E, Hantak M, Perlman S, Gallagher T. Distinct roles for sialoside and protein receptors in coronavirus infection. mBio 2020;11:e02764–819.
[43] Myung J, Kim KB, Crews CM. The ubiquitin-proteasome pathway and proteasome inhibitors. Med Res Rev 2001;21:245–73.
[44] Yang N, Shen HM. Targeting the endocytic pathway and autophagy process as a novel therapeutic strategy in COVID-19. Int J Biol Sci 2020;16:1724.
[45] Liu W, Li H. COVID-19: attacks the 1-beta chain of hemoglobin and captures the porphyrin to inhibit human heme metabolism. [Preprint] 2020 [cited 2020 July 13]. Available from: https://doi.org/10.26434/chemrxiv.11938173.v5.
[46] Dey L, Mukhopadhyay A. DenvInt: a database of protein–protein interactions between dengue virus and its hosts. PLoS Neglected Trop Dis 2017;11:e0005879.
[47] Ako-Adjei D, Fu W, Wallin C, Katz KS, Song G, Darji D, et al. HIV-1, human interaction database: current status and new features. Nucleic Acids Res 2015;43:D566–70.
[48] Kwofie SK, Schaefer U, Sundararajan VS, Bajic VB, Christoffels A. HCVpro: hepatitis C virus protein interaction database. Infect Genet Evol 2011;11:1971–7.
[49] Zhou X, Park B, Choi D, Han K. A generalized approach to predicting protein-protein interactions between virus and host. BMC Genom 2018;19:568.
[50] Gurumayum S, Brahma R, Naorem LD, Muthaiyan M, Gopal J, Venkatesan A. ZikaBase: an integrated ZIKV-human interactome Map database. Virology 2018;514:203–10.
[51] Shapira SD, Gat-Viks I, Shum BO, Dricot A, de Grace MM, Wu L, et al. A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. Cell 2009;139:1255–67.