J. Inst. Eng. India Ser. B (December 2021) 102(6):1293–1302
https://doi.org/10.1007/s40031-021-00569-7

CASE STUDY

Finding Prediction of Interaction Between SARS-CoV-2 and Human Protein: A Data-Driven Approach

Moumita Ghosh[1] · Pritam Sil[1] · Anirban Roy[2] · Rohmatul Fajriyah[3] · Kartick Chandra Mondal[1]

Received: 13 August 2020 / Accepted: 26 February 2021 / Published online: 4 May 2021

Abstract COVID-19 pandemic defined a worldwide health crisis into a humanitarian crisis. Amid this global emergency, human civilization is under enormous strain since no proper therapeutic method is discovered yet. A wave of research effort has been put toward the invention of therapeutics and vaccines against COVID-19. Contrarily, the spread of this fatal virus has already infected millions of people and claimed many lives all over the world. Computational biology can attempt to understand the protein–protein interactions between the viral protein and host protein. Therefore, potential viral–host protein interactions can be identified which is known as crucial information toward the discovery of drugs. In this study, an approach was presented for predicting novel interactions from maximal biclusters. Additionally, the predicted interactions are verified from biological perspectives. For this, a study was conducted on the gene ontology and KEGG-pathway in relation to the newly predicted interactions.

Keywords Biclusters · Association Rule · Protein–Protein Interactions · COVID-19

✉ Kartick Chandra Mondal
kartickjgec@gmail.com

[1] Department of Information Technology, Jadavpur University, Kolkata, India

[2] Department of Environment, West Bengal Biodiversity Board, Kolkata, India

[3] Department of Statistics, Islamic University of Indonesia, Yogyakarta, Indonesia

Introduction

Protein–protein interactions (PPI) denote a complex network of reactions that take the responsibility to synchronize and execute the biological process at the cell level in all organisms. PPI dataset is of immense importance in molecular and system biology. Due to its power of revealing the infection mechanism by viral protein, nowadays it plays a crucial role in the study of drug discovery. Thus, treatment optimization is enhanced. A large number of experimental and computational approaches [1] have been attempted for drug discovery following the PPI dataset as the computational approaches can infer the predicted interactions [2]. Hence, analyzing the PPI dataset of SARS-CoV-2 and human would be useful.

Being computer science researchers, the authors could show an efficient algorithmic approach so that the prediction of the protein–protein interaction could be done in an efficient way. Not only that, its has been expanded the scope by incorporating a way to validate the predictions made. Any interaction dataset is easily represented by binary dataset that is easier to store in a compact form. Hence, it was worthy to use Bimax [3] biclustering approach that could be used to generate biclusters [4]. Further, algorithmic approach is able to identify the association rules [5] from generated biclusters.

A very few research [6, 7] have been done specifically in case SARS-CoV-2 or COVID-19 disease. The investigators [8] presents a three-layer network model through which Alphainfluenzavirus proteins (having similarity with SARS-CoV-2 proteins) are considered as an intermediary to predict protein interaction between SARS-CoV-2 and human proteins. A random walk model is proposed in the literature [9] to identify the SARS-CoV-2 pathogenic mechanism on viral–host protein interaction network.

The literature has shown that many analyses have been done over the PPI network following the classification-based methodology. An ensemble voting classifier model based upon SVM and Random Forest Technique is proposed in the researches [10] to predict PPIs between the SARS-CoV-2 and human proteins. Classification-based approach suffers from the subjective judgment which is the underlying feature for the designing of a classifier. Specific to the context, for the PPI dataset, a classifier needs both positive and negative interaction data. Though positive cases are validated experimentally, there is no evidence of negative interactions. It is true that based on the quantity of positive data, the classifier becomes reliable. At the same time, the selection of negative data is also important for performance consideration. Thus, a good classifier needs proper exploitation of both positive and negative interaction data. This kind of barrier directs us to follow the approach of rule mining-based methodology which has been successfully implemented in the research [2]. Besides, multiple groups of researchers have studied the HIV-human protein interaction and multiple efforts have already made toward this direction of drug discovery. The researches [2, 11, 12] are a few notable contributions among many.

So, summarizing the above discussion, in this study, bimax biclustering approach has been followed on the manually curated protein–protein interaction dataset of SARS-COV-2 and human protein. It has been shown the procedure used for predicting interactions from biclusters and from the generated rules as well. Moreover, the prediction has been related to gene ontology- and KEGG-pathway-based study using two well-known bioinformatics tool.

Approach

In order to predict new interactions from the existing interaction set, we make use of the association rule mining [13] approach. It consists of two main sub-parts. These are finding frequent item-sets (FI) or patterns (FP) and generating association rules from those patterns. As finding FP is computationally expensive, it has been decided to focus only on the maximal frequent patterns [3]. All the supersets of such patterns are infrequent, but information loss arises in the case of the subsets. This is because all subsets are frequent, but supporting objects are not available. To combat this information loss while minimizing the computational cost, frequent closed itemset [2] (FCI)/ bicluster mining could be the most favorable way. A FI is called closed if there exists no superset S of FI such that supporting objects of S is equal to the supporting objects of FI. Thus, using the set of FCI, all FI across the column attributes along with their supporting row objects can be extracted. In consequence, this approach reduces the number of rules to be presented to the users without losing any information.

In this regard, it can be concluded that finding maximal biclusters from a binary dataset is synonymous to retrieving the FCI in frequent pattern mining problem [14]. Thus, the first subproblem is converted into the problem of maximal bicluster generation. The step of generation of maximal biclusters is followed by the step of rule generation and new interaction prediction. These two are the main subtasks here as shown in Fig. 1.

Biclustering [4] is a way to compute the cohesiveness among the elements of a dataset. Bimax [3] is a well-known biclustering algorithm for extracting the maximal biclusters from the binary datasets. It follows a simple divide–conquer approach and extracts all inclusion maximal biclusters, basically, submatrices of all 1s'. An inclusion maximal bicluster indicates that the bicluster is not contained completely in any other bicluster. Here bimax was used as it accomplishes results within comparable execution time and memory usage along with the other best-performing algorithms [3]. The obtained biclusters has been used to predict new interactions and further to generate rules. Additionally, these rules also do prediction for new interactions. A straight forward approach is used to generate the rules from the generated biclusters or maximal closed frequent pattern as explained in [2]. Pictorial representation of the process is shown in Fig. 1. Here, it has been considered and generated only a non-redundant set of association rules for the necessary knowledge prediction.

Preparation of the SARS-CoV-2-Human PPI Dataset

The experimental dataset is created by combining PPI interactions from two different sources, viz. BioGRID repository [15] and a research article [6]. BioGRID is a conventional biological repository for protein–protein interaction, genetic interaction, etc. It was found a less amount of experimentally proved data of interaction in the data repository to extract sufficient predicted information. For this, an extra step for prediction in the data preprocessing step is carried out along with the rule mining stage. To this end, for assuring reliability, the final set of predicted interactions are justified in later phases by gene ontology and KEGG-pathway enrichment analysis. Semantic similarity of gene ontology terms of two proteins

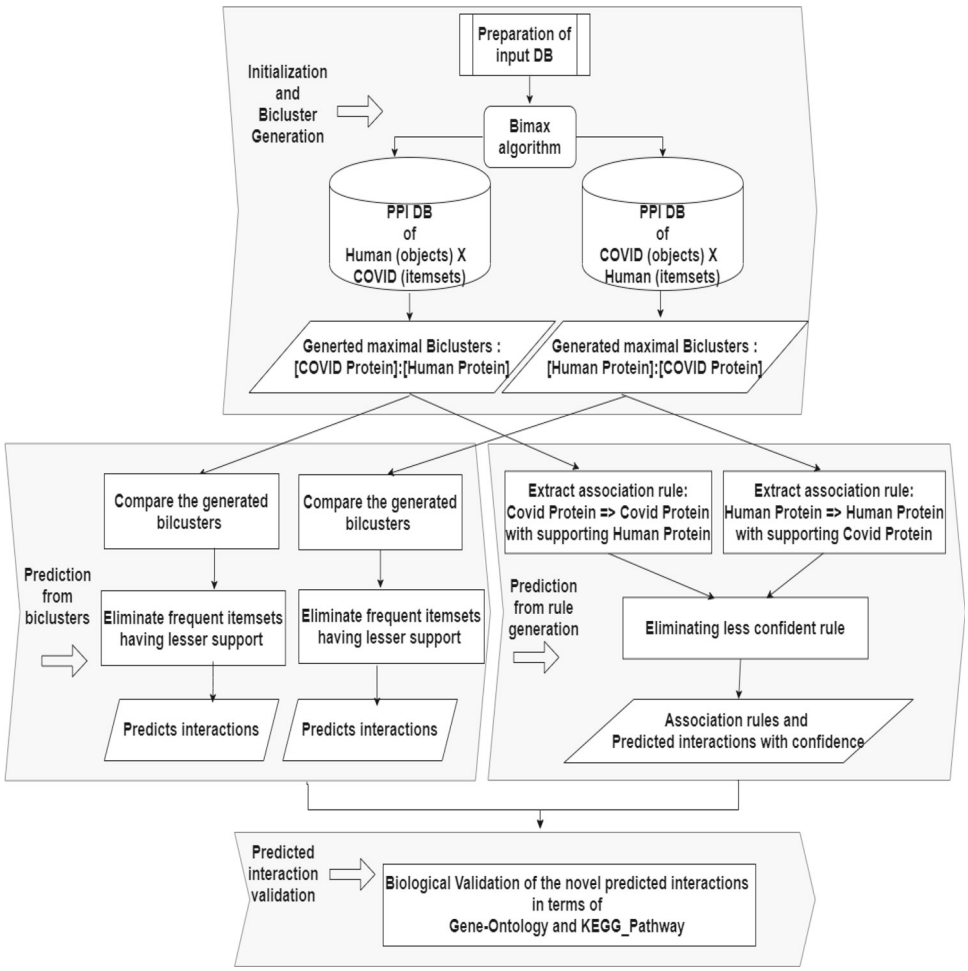

Fig. 1 Outline of the approach used for the experimentation performed

reveals the functional similarity which directs toward the probability of interaction.

It has been anticipated that the interactions may be possible by the rule of transitivity. The data was generated of intra organism SARS-CoV-2 protein interactions. Thus, it was assumed that a SARS-CoV-2 protein A interacts with other SARS-CoV-2 proteins X, Y, Z, and human proteins as well. Moreover, X, Y, and Z interact with some other human proteins. Hence, SARS-CoV-2 protein A may have the likelihood of interaction with the human proteins that have already known interaction with X, Y, and Z. Explain it by assuming there is a pathway. One protein would work in the downstream function of another. Unknown molecules or proteins may assemble with Protein A to form a complex which may interact with the human protein. Eventually, all these predicted interactions exploited along with the known interactions to form frequent closed patterns. Then, after removing the infrequent patterns, the final set of predicted interactions was listed.

Our dataset consists of 28 SARS-CoV-2 proteins and 346 human proteins and a total of 1255 interaction data. In the dataset, 1 represents an interaction between the human protein and viral protein. This binary dataset is treated as an input to the bimax algorithm. The input dataset is given as supplementary material along with this article.

Result and Discussion

SARS-CoV-2-Human PPI dataset is denoted as HxC where the rows represent the human proteins and the columns represent the viral proteins. In this case, the clustered itemsets of SARS-CoV-2 proteins that have a similar interaction pattern for a subset of objects of human proteins. It was also used the transposed form of the SARS-CoV-2-Human PPI dataset which is represented by the CxH. The aim of this transposed data is to obtain the clusters along with the human proteins for a subset of objects of SARS-CoV-2 proteins. From the statistics represented in Table 1, it could be justified that to cover all the possibilities, it was needed to experiment on both the original and the transposed datasets as there is a difference

Table 1 Statistics for the generated outputs

Dataset	Minimum support	# Biclusters	Non-redundant rules > 0.7 Confidence	# Predicted Interactions
HxC_OLD	2	10	0	0
	3	6	0	0
	4	4	0	0
CxH_OLD	2	10	0	0
	3	1	0	0
HxC	2	34	5	48
	4	24	5	48
	6	22	5	48
	8	17	5	48
	11	17	5	48
	13	12	5	48
	15	11	5	48
	20	10	5	48
	30	10	5	48
	40	10	3	39
CxH	2	34	7	8
	4	22	7	8
	6	8	4	6
	8	4	2	2

in the obtained results of the biclusters in both cases. Here, each bicluster specifies that with a subset of human proteins, a subset of SARS-CoV-2 proteins interacts and vice versa.

Both HxC and CxH were used as input to the bimax algorithm and a summary of generated biclusters is found in Table 1 showing the details with varying support counts. It has listed the interaction information for two sets of both of the datasets, i.e., before (HxC_OLD, CxH_OLD) and after(HxC, CxH) applying the rule of transition on the original known protein–protein interaction information. Figure 2 shows the graphical representation of Table 1 (for the HxC and the CxH dataset) where the number of biclusters changes significantly with varying minimum support while the number of rules and predicted interactions are less prone to change. It can be observed that the number of generated biclusters is decreasing with the increasing value of minimum support. Here, the minimum support value corresponds to the minimum number of rows to be considered. The minimum number of columns is always taken as 2 for generating all possible frequent closed itemsets (frequent 1 itemsets are not considered as they are unable to generate any rule).

Interactions Prediction from Biclusters

Interactions can be predicted from the obtained list of biclusters. Let us take an example of HxC dataset and consider two obtained biclusters in the form of <IS_CP1 : IS_HP1> and <IS_CP2 : IS_HP2> where <IS_CPi : IS_HPi> represents itemsets denoting i^{th} bicluster consisting of SARS-CoV-2 protein (CP) and human protein (HP). Here, a bicluster could be thought of as a cluster of elements (CP) that exhibits the same nature for a set of supporting objects (HP). If the similarity between any two clusters is more than the user specified threshold value, then we could make an attempt to predict new probable interactions for the remaining unmatched members of each cluster under experiment with the unmatched supporting objects of the other cluster. The similarity is measured by intersection operation between the two clusters. For predicting new interactions, we take the advantage of the set difference operator. $|IS\_CP1 - IS\_CP2| \overset{PredictsInteractionWith}{\longrightarrow} |IS\_HP2 - IS\_HP1|$. Similarly, $|IS\_CP2 - IS\_CP1| \overset{PredictsInteractionWith}{\longrightarrow} |IS\_HP1 - IS\_HP2|$.

Following this way, we obtain a list of predicted interactions, was obtained and Table 2 shows two sets of examples of such predicted lists. From the row 1a. and 1b., it can be derived that, having 80% similarity between the two clusters (for 1a., out of 5 elements in cluster, common elements with 1b. are ORF8, NSP7, NSP8, NSP12, so, similarity percentage of 1a. with 1b. can be represented by 4/5), it can be claimed that ORF9C may have the interaction with GLA and AKAP8L. In the last row, nothing can be predicted as <N, NSP4> is a proper subset of <N, ORF3A, NSP4>.

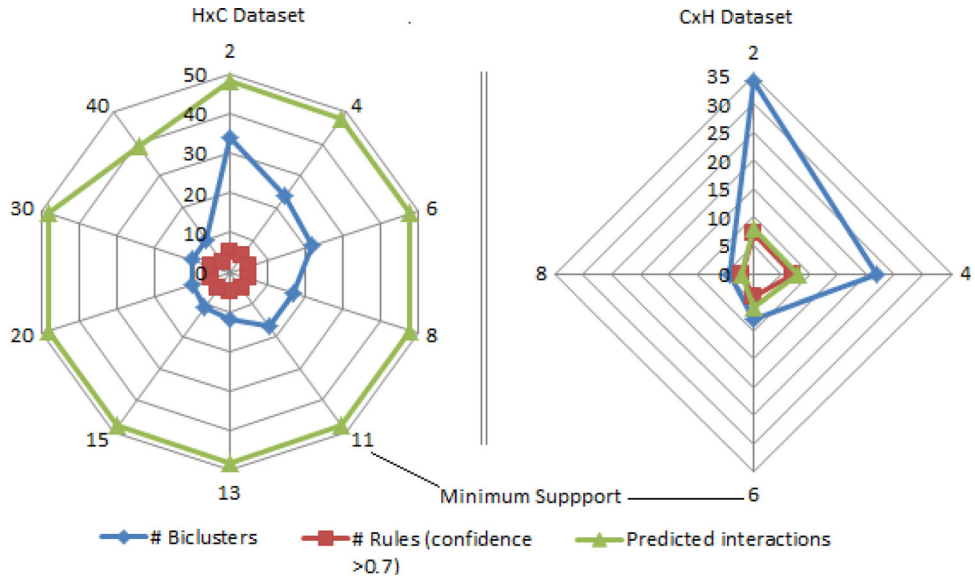

Fig. 2 Radar graph representation for the variation in the number of generated biclusters, rules, and predicted interactions with respect to minimum support

Table 2 Predicted interactions from the biclusters: taking similarity threshold 0.65

S. no.	Simi larity	Obtained biclusters	SARS-CoV-2 protein	Human protein
1a.	4/5, i.e., 80%	<ORF8 ORF9C NSP7 NSP8 NSP12 : FAM162A LMAN2 HS2ST1 NAT14 RALA EMC1>	ORF9C $\xrightarrow{Interacts}$	GLA AKAP8L
1b.	4/5, i.e., 80%	<M ORF8 NSP7 NSP8 NSP12 : GLA HS2ST1 AKAP8L>	M $\xrightarrow{Interacts}$	FAM162A LMAN2 NAT14 RALA EMC1
2a.	2/3, i.e., 66.6%	<N, ORF3A, NSP4 : RAB14 ALG11 DDX21 G3BP2 LARP1 MOV10 PABPC1 PABPC4 RBM28 RPL36 RRP9 UPF1 AP3B1 BRD2 BRD4 CWC27 SLC44A2 ZC3H18 ALG5 ARL6IP6 CLCC1 HMOX1 SUN2 TRIM59 VPS11 VPS39 CSNK2A2 CSNK2B FAM98A G3BP1 SNIP1 C19orf52 DNAJC11 IDE TIMM10 TIMM10B TIMM9 TIMM29 >	ORF3A $\xrightarrow{Interacts}$	NUP210
2b.	2/2, i.e., 100%	< N, NSP4 : RAB14 ALG11 NUP210 DDX21 G3BP2 LARP1 MOV10 PABPC1 PABPC4 RBM28 RPL36 RRP9 UPF1 AP3B1 BRD2 BRD4 CWC27 SLC44A2 ZC3H18 ALG5 ARL6IP6 CLCC1 HMOX1 SUN2 TRIM59 VPS11 VPS39 CSNK2A2 CSNK2B FAM98A G3BP1 SNIP1 C19orf52 DNAJC11 IDE TIMM10 TIMM10B TIMM9 TIMM29 >	Nothing can be predicted	—

Prediction by Generating Rules from the Biclusters

This section establishes the approach that we have already mentioned in "Approach." CxH dataset gives rules between human proteins, and HxC dataset gives rules between SARS-CoV-2 proteins which are true for corresponding SARS-CoV-2 and human protein, respectively. The obtained rules was filtered out in two steps. First, the rules based upon the support and confidence values. Being a sparse dataset, experimentally, it has been fixed, the support values to a lower threshold as follows. For the CxH dataset, the minimum support count is kept to 4 with a 70% minimum confidence level. Similarly, for the HxC dataset, the minimum support count is 11 with the same minimum confidence level as 70%. Next, the redundancy is eliminated taking the most general rules. Then, the interaction prediction is performed. From the final results, the HxC dataset predicts 48 unique and novel interactions. Similarly, from CxH dataset, 8 unique and novel interactions are predicted. To keep the prediction unambiguous, discard the interactions with a lower confidence level in case of the similar predicted interactions. The predicted interactions obtained from both the datasets are merged, and a PPI network is drawn as shown in Fig. 3. It has been three

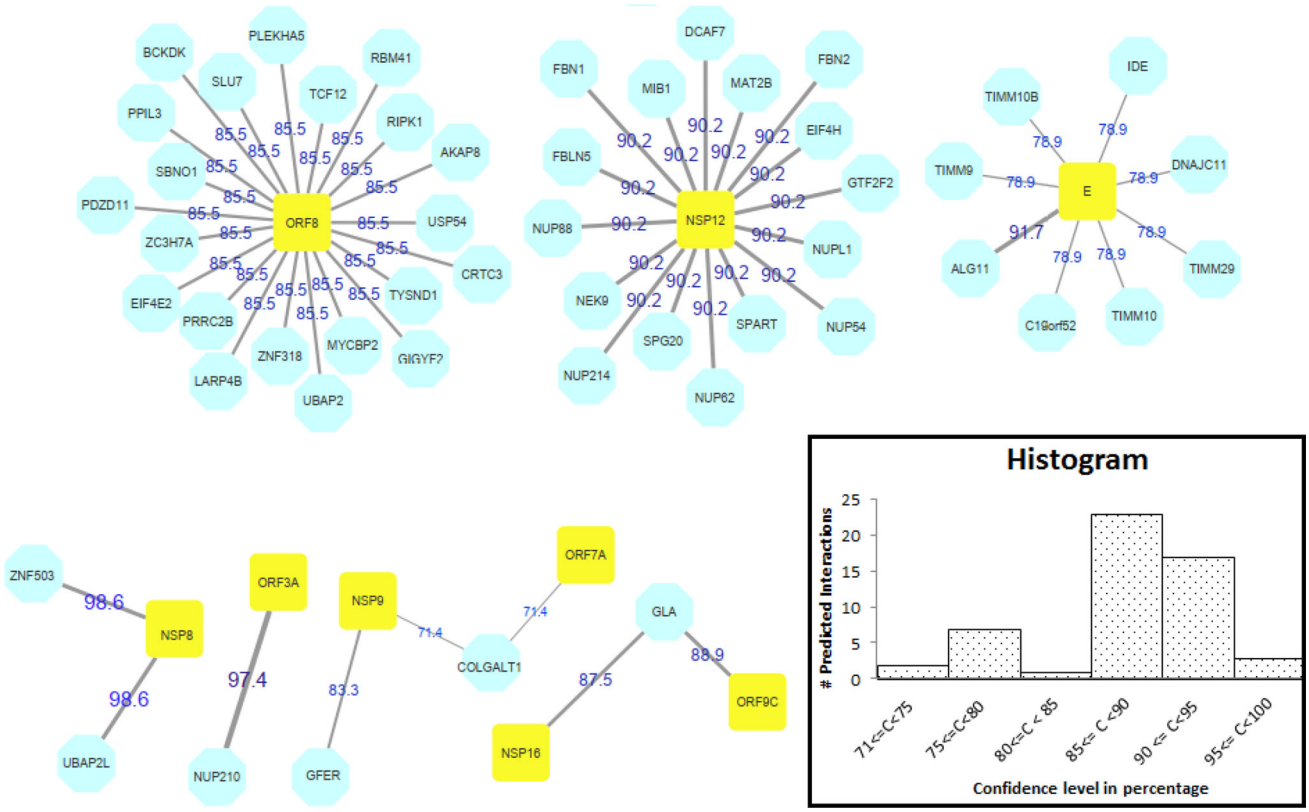

Fig. 3 Network for the predicted PPIs along with histogram for the number of predicted interactions at varying confidence level

common interactions, and thus, finally 53 unique novel interactions are obtained from the dataset. The confidence level for each predicted interaction lies within 85–90% in most of the cases followed by 90–95% interval. For this, it has been shown a histogram in Fig. 3 along with the interaction network. It can be seen from the figure that SARS-CoV-2 proteins ORF8, NSP12, and E act as network hub. All SARS-CoV-2 proteins are highlighted in yellow color, whereas the human proteins are in sky blue color. The edge width representing an interaction is getting wider along with the increasing confidence level of the interaction. Edge level is showing the value of confidence level for each interaction.

Gene Ontology-based justification of the predicted interactions

To justify the prediction, relevance with the help of biological interpretation was investigated. For this, DAVID (http://david.abcc.ncifcrf.gov) was used, a freely available online bioinformatics repository that provides functional annotations for a large set of genes. Among the multiple information extracted from this tool, it was opted for gene ontology (GO)-based study and KEGG-pathway. From all the three domains (Biological Process, Cellular Component, and Molecular Function) covered by GO, it was found the GO terms. Moreover, it has been extracted the non-redundant informative GO terms, based upon the p-values (taken from DAVID generated result), by using another tool REVIGO (http://revigo.irb.hr/). It reveals the outliers from the list of submitted GO-terms via checking the semantic similarity and outputs a sorted list based upon dispensability value for each GO-term. More unique terms are having lesser dispensability.

For the Predicted Interactions Obtained from the Biclusters

As shown in Table 2, ORF9B is predicted to have interactions with 24 human proteins. Table 3 shows GO-terms for verifying the predictions. It can be seen from the molecular function of predicted human proteins that these are related to identical binding activities. Study has shown that ORF9B also has a functional property of *membrane binding*.

For the Predicted Interactions Obtained from the Rules

Among the predicted interactions, SARS-CoV-2 protein ORF8 is found to have maximum predicted interactions with 21 human proteins. Followed by this, 16 and 8 numbers of human proteins have been predicted to interact

Table 3 Significant GO terms found in the human proteins that are predicted to interact with SARS-CoV-2 Protein ORF9B

GO-id	Term	% of Proteins	*P* value	Dispen-sability
GO Category: Biological Process (*P* value < 1*E*–01, Dispensability < 0.6)				
GO:1903955	positive regulation of protein targeting to mitochondrion	12.5	5.11*E*–3	0
GO:0000184	Nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	12.5	9.57*E*–3	0.019
GO:0006413	Translational initiation	12.5	1.25*E*–3	0.106
GO:0007265	Ras protein signal transduction	8.33	8.40*E*–2	0.183
GO:0016192	Vesicle-mediated transport	12.5	1.52*E*–2	0.216
GO:0006364	rRNA processing	12.5	2.89*E*–2	0.335
GO:0034058	Endosomal vesicle fusion	8.33	1.12*E*–2	0.347
GO:0045070	Positive regulation of viral genome replication	8.33	3.20E-2	0.445
GO:0008333	Endosome to lysosome transport	8.33	4.76*E*–2	0.453
GO:0016239	Positive regulation of macroautophagy	8.33	2.83*E*–2	0.468
GO:0045727	Positive regulation of translation	8.33	6.42*E*–2	0.516
GO Category: Cellular Component (*P* value < 1*E*–01, Dispensability < 0.6)				
GO:0010494	Cytoplasmic stress granule	12.5	9.35E-04	0
GO:0016020	Membrane	29.17	5.05*E*–2	0
GO:0005622	Intracellular	25	2.30*E*–2	0.075
GO:0030123	AP-3 adaptor complex	8.33	1.3*E*–2	0.227
GO:0005829	Cytosol	41.67	1.51*E*–2	0.327
GO:0005765	Lysosomal membrane	12.5	4.63*E*–2	0.33
GO:0005730	Nucleolus	20.83	2.10*E*–2	0.377
GO:0030897	HOPS complex	8.33	1.75*E*–2	0.417
GO:0031519	PcG protein complex	8.33	3.35*E*–2	0.491
GO:0030529	Intracellular ribonucleoprotein complex	16.66	6.45*E*–04	0.54
GO Category: Molecular Function (*P* value < 1*E*–01, Dispensability < 0.05)				
GO:0044822	Poly(A) RNA binding	54.17	1.26*E*–09	0
GO:0000166	Nucleotide binding	25	5.61*E*–05	0
GO:0004004	ATP-dependent RNA helicase activity	12.5	2.92*E*–3	0
GO:0003729	mRNA binding	12.5	1.04*E*–2	0
GO:0008494	Translation activator activity	8.33	1.11*E*–2	0
GO:0008143	Poly(A) binding	8.33	1.60*E*–2	0
GO:0005515	Protein binding	70.83	2.10*E*–2	0
GO:0003676	Nucleic acid binding	20.83	3.11*E*–2	0

with SARS-CoV-2 proteins NSP12 and E, respectively. Tables 4, 5, and 6 sum up the informative GO-terms verifying the identical biological activities for viral proteins ORF8, NSP12, and E, respectively.

While the activities are highly cohesive within each table data, each individual table has a distinct set of functions. Table 4 depicts significance GO terms like *membrane*, *poly(A) RNA binding*, etc., indicating the involvement of many human proteins in these. It is an important observation as these proteins are expected to have interaction with ORF8 that plays the main role in host–virus interaction. SARS-CoV-2 protein NSP12 is a multifunctional protein and mainly involved in the transcription and replication of viral RNAs. Table 5 is found to have many such common GO-terms related to RNA functions in the biological process. Similarly, from Table 6, it appears that many human proteins are involved in the molecular function of *protein transporter activity*. It says that these proteins are highly involved in transporting molecules across biological membrane. Hence, the prediction of the human proteins that interact with viral protein E is intuitive as it is a small membrane protein having a major role in the assembly of virions.

Table 4 Significant GO terms found in the human proteins that are predicted to interact with SARS-CoV-2 Protein ORF8

GO-id	Term	% of Proteins	*P* value	Dispen-sability
GO Category: Biological Process (*P* value < 1*E*–01, Dispensability < 0.3)				
GO:0000398	mRNA splicing, via spliceosome	14.28	2.57*E*–02	0
GO:0017148	Negative regulation of translation	9.52	6.36*E*–02	0.205
GO Category: Cellular Component (*P* value < 1*E*–01, Dispensability < 0.2)				
GO:0016020	Membrane	33.33	1.16*E*–2	0
GO:0071013	Catalytic step 2 spliceosome	9.52	8.24*E*–2	0
GO:0005737	Cytoplasm	42.85	8.30*E*–2	0.15
GO Category: Molecular Function (*P* value < 1*E*–01, Dispensability < 0.05)				
GO:0044822	Poly(A) RNA binding	33.33	1.12*E*–2	0

Table 5 Significant GO terms found in the human proteins that are predicted to interact with SARS-CoV-2 Protein NSP12

GO-id	Term	% of Proteins	*P* value	Dispen-sability
GO Category: Biological Process (*P* value < 1*E*–04, Dispensability < 0.05)				
GO:0007077	Mitotic nuclear envelope disassembly	35.71	2.88*E*–08	0
GO:0006409	tRNA export from nucleus	28.57	1.77*E*–06	0
GO:0010827	Regulation of glucose transport	28.57	1.95*E*–06	0
GO:0075733	Intracellular transport of virus	28.57	7.39*E*–06	0
GO:1900034	Regulation of cellular response to heat	28.57	2.37*E*–05	0
GO:0031047	Gene silencing by RNA	28.57	7.66*E*–05	0
GO:0016925	Protein sumoylation	28.57	8.96*E*–05	0
GO Category: Cellular Component (*P* value < 1*E*–01, Dispensability < 0.75)				
GO:0044613	Nuclear pore central transport channel	21.43	3.65*E*–05	0
GO:0001527	Microfibril	14.285	7.11*E*–03	0
GO:0005829	Cytosol	42.85	7.06*E*–02	0.1
GO:0031012	Extracellular matrix	21.42	1.82*E*–02	0.5
GO:0005643	Nuclear pore	14.28	5.01*E*–02	0.7
GO:0005578	Proteinaceous extracellular matrix	21.42	1.51*E*–02	0.7
GO category: Molecular Function (*P* value <1*E*–01, Dispensability < 0.05)				
GO:0017056	Structural constituent of nuclear pore	21.43	1.14*E*–04	0
GO:0005515	Protein binding	92.86	2.65*E*–03	0
GO:0005487	Nucleocytoplasmic transporter activity	14.29	1.68*E*–02	0
GO:0004386	Helicase activity	14.29	6.35*E*–02	0
GO:0005178	Integrin binding	14.29	7.80*E*–02	0.04
GO:0030023	Extracellular matrix constituent conferring elasticity	14.29	3.84*E*–03	0.31
GO:0005201	Extracellular matrix structural constituent	14.29	5.04*E*–02	0.4

Justification on the Predicted Interactions using KEGG-pathway

Along with the GO study, it has been examined the KEGG-pathway obtained from DAVID tool. KEGG-pathway has importance in bioinformatics research in understanding genomes, biological pathways, disease, drugs, etc.

For the Predicted Interactions Obtained from the Biclusters

Here, the example of KEGG-pathway enrichment for ORF9B was also addressed. The tool reveals that the human proteins that are predicted to have interactions with ORF9B are related to the *Metabolic pathways*, pathways to *Measles*, *Herpes simplex infection*. Many of the proteins

Table 6 Significant GO terms found in the human proteins that are predicted to interact with SARS-CoV-2 Protein E

GO-id	Term	% of Proteins	*P* value	Dispen-sability
GO Category: Biological Process (*P* value < 1*E*–01, Dispensability < 0.7)				
GO:0006626	Protein targeting to mitochondrion	42.86	2.38*E*–05	0
GO:0007605	Sensory perception of sound	28.57	3.12*E*–02	0
GO:0072321	Chaperone-mediated protein transport	28.57	1.90*E*–03	0.42
GO:0015031	Protein transport	28.57	9.08*E*–02	0.68
GO Category: Cellular Component (*P* value < 1*E*–03, Dispensability < 0.7)				
GO:0042719	Mitochondrial intermembrane space protein transporter complex	42.86	9.03*E*–07	0
GO:0005739	Mitochondrion	71.43	3.77*E*–04	0.28
GO:0005743	Mitochondrial inner membrane	71.43	4.88*E*–06	0.65
GO:0042721	Mitochondrial inner membrane protein insertion complex	28.57	9.87*E*–04	0.68
GO Category: Molecular Function (*P* value < 1*E*–01, Dispensability < 0.5)				
GO:0042803	Protein homodimerization activity	42.86	2.50*E*–2	0
GO:0008565	Protein transporter activity	28.57	2.53*E*–2	0
GO:0005215	Transporter activity	28.57	6.97*E*–2	0
GO:0008270	Zinc ion binding	42.86	6.0*E*–2	0.06
GO:0051087	Chaperone binding	28.57	2.94*E*–2	0.46

are related to the *B-signaling pathway*, *RNA degradation*, etc.

For the Predicted Interactions Obtained from the Rules

From the KEGG-pathway obtained using the tool, it has come by the probable interaction of ORF8 with human protein which indicates that multiple cellular activities may lead to Human T-Cell Leukemia Virus Infection. Human proteins that are predicted to interact with ORF8 are involved in the pathway of *HTLV-I infection*, *Apoptosis*, *Hepatitis C*, *Epstein-Barr virus infection*, *Insulin signaling pathway*, etc. For the predicted interactions with NSP12, the pathways are *Metabolic pathways*, *RNA transport*, etc. Similarly, SARS-CoV-2 protein E is expected to interact with human protein ALG11, IDE that are found to have involvement in *Metabolic pathways*, *Alzheimer's disease*, etc.

Conclusions

The main aim here is to help in accelerating the procedure of designing the drug and hence an improved medication for COVID-19. For this, the simple binary SARS-CoV-2-Human PPI dataset has been exploited. Formation of biclusters, association rule generation, and the procedure for discovering novel interactions are shown here. Also, the predicted interactions are interpreted biologically for finding their relevance. It has been seen that multiple human proteins that have predicted interaction with a single viral protein share common biological activities. The present study has left scope for consideration of multiple interaction types. Directions of the interactions (viral to host or host to viral) are also crucial information to be examined and working with.

References

1. H. Umbrin, S. Latif. A survey on protein protein interactions (ppi) methods, databases, challenges and future directions. In *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pages 1–6. IEEE, (2018)
2. K. C. Mondal, N. Pasquier, A. Mukhopadhyay, C. da Costa - Pereira, U. Maulik, A. GB Tettamanzi. Prediction of protein interactions on HIV-1-human PPI data using a novel closure-based integrated approach. In *International Conference on Bioinformatics Models, Methods and Algorithms*, pages 164–173. SciTePress, (2012)
3. A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, E. Zitzler, A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics **22**(9), 1122–1129 (2006)
4. S.C. Madeira, A.L. Oliveira, Biclustering algorithms for biological data analysis: a survey. IEEE/ACM transactions on computational biology and bioinformatics **1**(1), 24–45 (2004)
5. K. C. Mondal. *Algorithms for Data Mining and Bio-informatics*. PhD thesis, University of Nice Sophia Antipolis, (2013)
6. D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, K. M. White, M. J. O'Meara, V. V. Rezelj, J. Z. Guo, D. L. Swaney, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature **583**, 459–468 (2020)

7. A.A. Khan, Z. Khan, COVID-2019 associated overexpressed Prevotella proteins mediated host-pathogen interactions and their role in coronavirus outbreak. Bioinformatics **36**(13), 4065–4069 (2020)
8. B. Khorsand, A. Savadi, M. Naghibzadeh. SARS-CoV-2-human protein-protein interaction network. Informatics in Medicine Unlocked **20**, 1–10 (2020)
9. Y. Zhang, T. Zeng, L. Chen, S. Ding, T. Huang, Y. Cai. Identification of COVID-19 Infection-Related Human Genes Based on a Random Walk Model in a Virus–Human Protein Interaction Network. BioMed research international **2020**, 1–7 (2020)
10. L. Dey, S. Chakraborty, A. Mukhopadhyay, Machine learning techniques for sequence-based prediction of viral-host interactions between SARS-CoV-2 and human proteins. Biomedical journal **43**(5), 438–450 (2020)
11. D. Pal, A. S. Mondal, K. C. Mondal. Knowledge discovery from HIV-1-human PPIs assimilating interaction keywords. In *2016 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, pages 1–8. IEEE, (2016)
12. D. Pal, K. C. Mondal. Predicting novel interactions from HIV-1-Human PPI data integrated with protein signatures and GO annotations. *International Journal of Bioinformatics Research and Applications (IJBRA)*, (2020)
13. R. Agarwal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference*, pages 487–499, (1994)
14. K. C. Mondal, N. Pasquier, A. Mukhopadhyay, U. Maulik, S. Bandhopadyay. A new approach for association rule mining and bi-clustering using formal concept analysis. In *International conference on Machine Learning and Data Mining in Pattern Recognition*, pages 86–101. Springer, (2012)
15. C. Stark, B. J. Breitkreutz, L. Reguly, T.and Boucher, M. Breitkreutz, A.and Tyers. BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(Database issue):D535—D539, (2006)