Informatics in Medicine Unlocked 20 (2020) 100413

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

SARS-CoV-2-human protein-protein interaction network

Babak Khorsand *, Abdorreza Savadi, Mahmoud Naghibzadeh

Department of Computer Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

ARTICLE INFO

Keywords:
COVID-19
SARS-CoV-2
Coronavirus
Protein-protein interaction
Host-pathogen protein interaction
Protein interaction prediction

ABSTRACT

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the novel coronavirus which caused the coronavirus disease 2019 pandemic and infected more than 12 million victims and resulted in over 560,000 deaths in 213 countries around the world. Having no symptoms in the first week of infection increases the rate of spreading the virus. The increasing rate of the number of infected individuals and its high mortality necessitates an immediate development of proper diagnostic methods and effective treatments. SARS-CoV-2, similar to other viruses, needs to interact with the host proteins to reach the host cells and replicate its genome. Consequently, virus-host protein-protein interaction (PPI) identification could be useful in predicting the behavior of the virus and the design of antiviral drugs. Identification of virus-host PPIs using experimental approaches are very time consuming and expensive. Computational approaches could be acceptable alternatives for many preliminary investigations. In this study, we developed a new method to predict SARS-CoV-2-human PPIs. Our model is a three-layer network in which the first layer contains the most similar Alphainfluenzavirus proteins to SARS-CoV-2 proteins. The second layer contains protein-protein interactions between Alphainfluenzavirus proteins and human proteins. The last layer reveals protein-protein interactions between SARS-CoV-2 proteins and human proteins by using the clustering coefficient network property on the first two layers. To further analyze the results of our prediction network, we investigated human proteins targeted by SARS-CoV-2 proteins and reported the most central human proteins in human PPI network. Moreover, differentially expressed genes of previous researches were investigated and PPIs of SARS-CoV-2-human network, the human proteins of which were related to upregulated genes, were reported.

1. Introduction

Coronaviruses (CoVs) are a big family of viruses that can cause diseases ranging from common cold to severe respiratory tract infections [1]. Moreover, some types of CoVs are zoonotic and they are transmittable from animals to human. Sever Acute Respiratory Syndrome (SARS-CoV) is one of the strains of coronavirus which came from civet cats [2] and horseshoe bat [3] and emerged in 2002/2003 in southern china and spread to 26 countries with 8096 infected cases leading to 774 deaths [4]. Middle East Respiratory Syndrome (MERS-CoV) is another one, which came from dromedary camel which was detected in Arabian Peninsula with 2494 infected cases leading to 858 deaths [5]. The latest version of CoV called Novel Coronavirus (SARS-CoV-2) emerged in Wuhan [6], a Chinese city with a population of 11 million, which causes coronavirus disease 2019 (COVID-19) [7]. SARS-CoV-2 probably came from bat [8] or minks [9]. COVID-19 was initially detected in December 2019 and contaminated 835 cases leading to 25 deaths up to Jan 22, 2020 [10]. The number of infections increased to 17400 cases leading to 362 deaths till Feb 2, 2020 [11] and more than 40000 cases leading to 800 deaths till Feb 10, 2020 [12]. The number of infections has increased at an exponential rate, with a doubling period of 1.8 days [13]. By gathering 180 reports from the world health organization (WHO) we built a database to show how COVID-19 spread to more than 12 million cases all over the world and killed more than 560,000 of its victims till July 10, 2020. Figs. 1–3 show how SARS-CoV-2 became a pandemic all over the world within four month and how it kills thousands of people in different countries every day.

CoVs are single stranded positive sense RNA (ssRNA+) viruses which belong to the order Nidovirales, the family Coronaviridae, and the subfamily Orthocoronavirinae. Alphacoronoavirus, Betacoronavirus, Deltacoronavirus, and Gammacoronavirus are its four different genera among which Betacoronavirus is the most pathogenic genus [14]. Betacoronavirus has five subgenera including Embecovirus, Sarbecovirus, Merbecovirus, Nobecovirus, and Hibecovirus [15].

* Corresponding author.
E-mail address: khorsand@mail.um.ac.ir (B. Khorsand).

https://doi.org/10.1016/j.imu.2020.100413
Received 20 June 2020; Received in revised form 11 July 2020; Accepted 10 August 2020
Available online 13 August 2020

Embecovirus, including types OC43 and HKU1, generally cause mild to moderate upper-respiratory tract illnesses, like the common cold. SARS-CoV and SARS-CoV-2 belong to Sarbecovirus (SV) subgenus, while MERS-CoV belongs to Merbecovirus subgenus.

Viruses are parasites which lack the capacity to live and reproduce outside of a host body. Protein-protein interaction (PPI) between viral proteins and host proteins is indispensable for viral proteins to reach the host cells. Consequently, identification of virus-host PPI network is the key to predict the behavior of viruses which in turn can be useful in designing antiviral drugs.

Biomolecular fluorescence complementation [16], co-immunoprecipitation [17], and yeast two-hybrid are some of the experimental methods for detecting virus-host PPI. These methods are both expensive and very time-consuming while computational methods are fast and inexpensive in prediction of virus-host PPIs.

Many studies confirm the practicality of computational methods in PPI prediction while further researches can improve the performance of these methods especially with respect to accuracy. Model training is one line of research in which computational models are trained using extracted features of experimentally developed PPIs. Different classifiers were used in different studies in the learning phase of the models. Sun used deep learning [18] while Dyer [19], Chatterjee [20], Mei [21], and Eid [22] applied support vector machine to predict virus-host PPIs. Nourani used Naïve Bayes [23]. Decision tree related classifier was the approach taken by Basit [24]. Random forest was the approach followed by both Yang [25] and Barman [26]. The approach followed by Leite [27], Zahiri [28], and Mei [29], were k-nearest neighbors, multilayer perceptron, and AdaBoost, respectively.

On the other hand, researchers work on different viruses to predict their PPI networks. Zhao predicted HIV1-human PPI network [28], Khorsand uncovered Alphainfluenzavirus-human PPI network [29], Ray predicted HCV-human PPI network [30], Chan revealed west nile virus-human PPI network [31], and Duran worked on herpesvirus-human PPI network [32].

Features to learn a model is another view to look at predicting virus-host PPI network. Alguwaizani used repeating patterns and amino acids composition [33], Kösesoy used location based encoding of amino acids [34], Mir used structure similarity [35], Guven-Maiorov used interface similarity [36], and Khorsand used network topology and gene ontology [29] in their researches towards learning PPI networks.

In the present study as it is shown in Fig. 4, we constructed SARS-CoV-2-human PPI network by building a novel three-layer network in which the first layer represents Alphainfluenzavirus-Human PPI network, the second layer shows the Alphainfluenzavirus-SARS-CoV-2 similarity network, and the third layer reveals SARS-CoV-2-human PPI network from clustering coefficient rule of the first two layers.

In the rest of this paper, in Section 2, by performing protein basic local alignment search tool (Balstp) on SARS-CoV-2 proteins, its orthologs were detected. As all of its orthologs belongs to SV, we decided to work on the whole SV family. A novel method is proposed to detect SV proteins' orthologs among Alphainfluenzavirus (AIV) proteins. And finally, by building a multi-layer network, we constructed SV-human PPI network from which SARS-CoV-2-human PPI network is extractable. In Section 3, some of SARS-CoV-2 features were reported which could be used in machine learning approach for predicting PPI network. Thereafter, by analyzing the constructed SARS-CoV-2-Human PPI network, the most central human proteins targeted by SARS-CoV-2 is reported. Eventually, differentially expressed genes of two available gene expression omnibus series (GSE) were investigated and PPIs of SARS-CoV-2-human PPI network which have human proteins among upregulated genes are marked. A short conclusion is placed in Section 4.

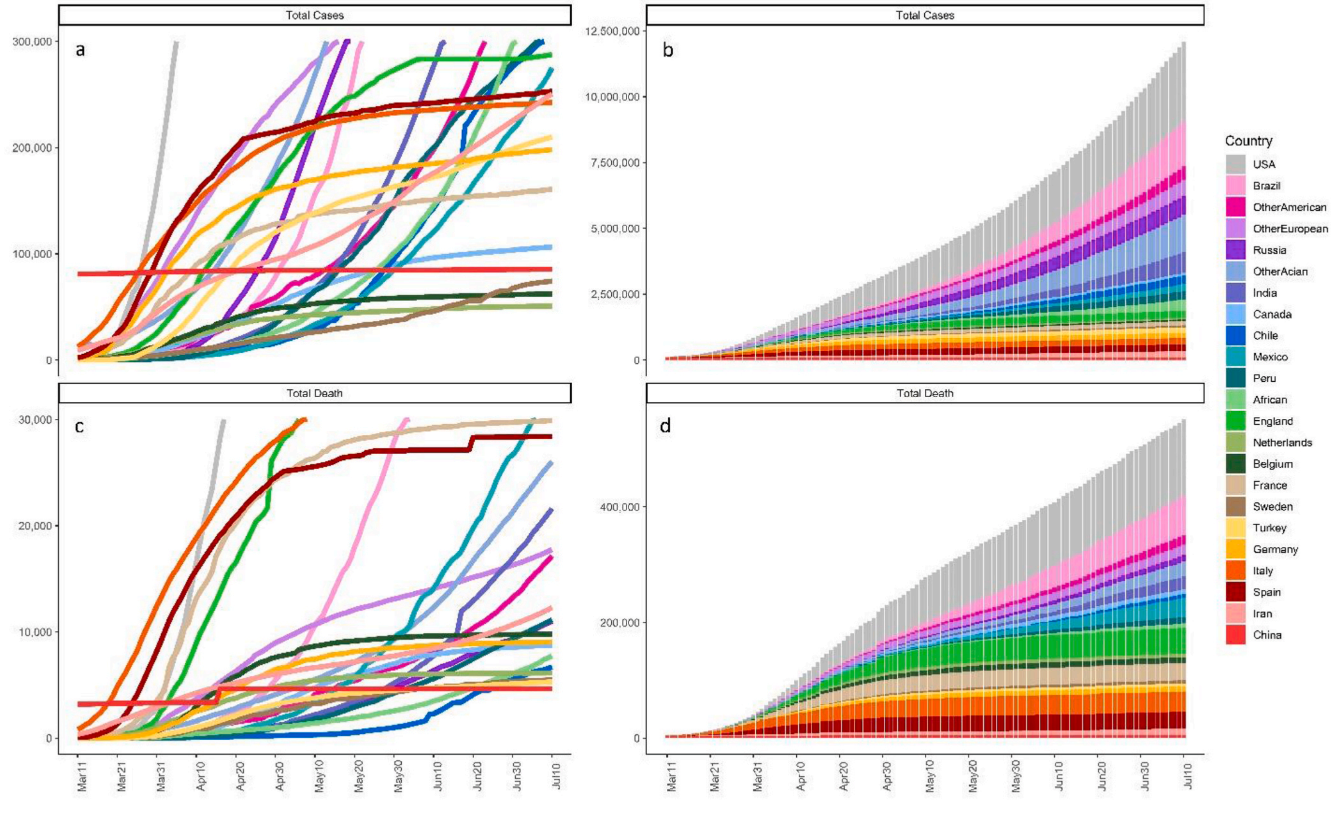

Fig. 1. Fig. 1-a shows the spreading speed of COVID-19 confirmed cases of different countries in reaching 300000 cases, while 1-c shows the time that each country needs to sacrifice 30000 cases by COVID-19. Fig. 1-b shows the total COVID-19 confirmed cases of different countries, and 1-d shows the total number of deaths caused by COVID-19.

2. Material and methods

SARS-CoV-2 has a genome with 29903 nucleotides [37] and 14 proteins. Blastp [38] was performed on all of these proteins to identify the virus proteins (VPs) with the highest sequence identity. For each of the SARS-CoV-2 proteins, the first three proteins with the highest scores are reported in Table 1.

SARS-CoV-2 and all of its homologs belong to SV, so we decided to work on the whole SV proteins. SV has 1772 VPs belonging to 158 different strains. Strains which contain at least three VPs were chosen. The 113 chosen strains (containing 1726 VPs) were clustered to 14 groups according to their sequence length (from VPs with sequence length of less than 50 residues up to VPs with sequence length of more than 6800 residues). Each of 14 groups, clustered into groups with at least 98% sequence identity. Thereafter, within each group, among VPs which have sequence identity of more than 98%, a representative VP (VP which has the highest sequence identity with the others) was chosen. After these two clustering steps, from 1726 VPs the total of 294 VPs were chosen.

As pathogenicity and transmissibility of SV are similar to those of one of the other families of respiratory viruses, AIV [39], and there are many experimental PPIs between AIV proteins and human proteins (HPs) in PPI databases such as Intact [40] and VirusMINT [41], we probed AIV proteins to detect SV orthologs among them in order to detect possible HP victims.

For constructing AIV-human PPI network, 11040 AIV-human PPIs were extracted from STRING [42], Intact [40], DIP [43], VirusMINT [41], and BioGRID [44]. Among these PPIs, 10878 PPIs belonging to H1N1, H3N2, and H5N1 subtypes were chosen. The selected PPIs were among 2966 HPs and 119 VPs of 45 different strains. The 119 chosen VPs were clustered into groups such that in each cluster each pair's mutual sequence identity was higher than 98%. In each group a representative with the highest average sequence identity scores in all other VPs of the respective group were chosen. This process led to the selection of 74 VPs from a collection of 119 AIV proteins.

For detecting orthologs of SV proteins in AIV proteins, four different scores were calculated between each of the 294 SV proteins and each of the 74 AIV proteins.

I. Primary structure similarity score: Smith-Waterman sequence alignment [45] was performed on all 21756 sequence pairs without any opening gap penalty but with −2 penalty score for the extension gap. Blocks substitution matrix 62 (BLOSUM62) was considered as scoring matrix. For each of the pairs, local alignment score was considered as its primary structure similarity, PS, score.

II. Secondary structure similarity score:

For simplicity, secondary structure (SS) of each protein was expressed by 3 types of SS namely helix, strand and coil. Three different scores, extracted from SS of each pair.

a. Ordinary form: Local alignment on SS of each of 21756 pairs without any opening gap penalty but with −2 penalty score for extension gaps. In similarity matrix for each match 3 credit points were considered as its alignment score. For mismatch of coil against helix or strand −1 and for mismatch of helix against strand −3 was considered as its alignment score. For each of the pairs, the sum of alignment scores was considered as SS similarity score called *SSO*.

b. Compact form: Local alignment of the compact form of SSs with the parameters of the previous step. The SS compact form was obtained simply by eliminating the consequent repeats of each SS types. As an example, with considering "HHHHHHCCCCHHHHHCCCCEEEEEEECCHHHHH" as the SS of a protein, its SS compact form will be "HCHCECH". For each of the pairs, sum of alignment scores was considered as SS similarity score called *SSC*.

c. Longest common substring: Maximum number of consecutive matches between each pair considered as its SS similarity score called *SSLCS*.

III. Accessibility similarity score:

Accessible surface area (ASA) of a residue is the surface area of that residue over which it contacts with solvent. Accessibility of a residue is the ASA of that residue in the folded protein over maximum possible ASA for that residue which is calculated by Tien [46]. For any residues of each protein, accessibility was calculated and each residue with accessibility score less than 0.25 was considered as buried. Accessibility score between 0.25 and 0.35 was considered as intermediate and more than 0.35 was considered as exposed. So, each protein was expressed by a sequence of 3 types of ASA including buried, intermediate, and exposed. Three different scores were extracted from ASA of each pair.

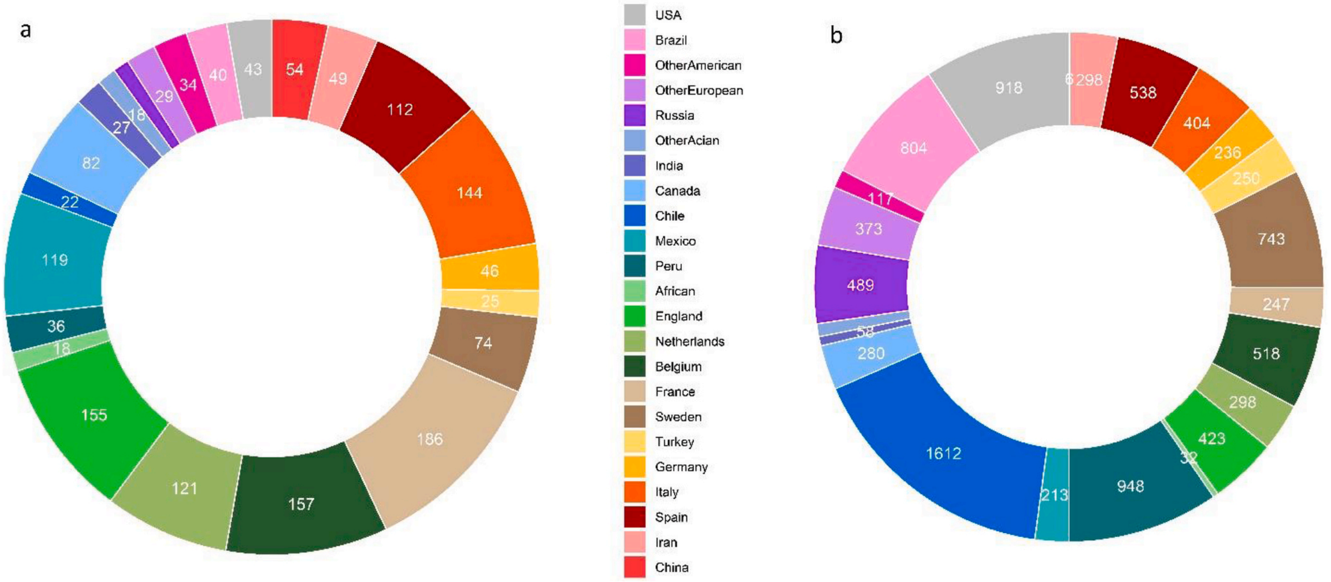

Fig. 2. Fig. 2-a shows the number of deaths from each 1000 COVID-19 cases (death rate). Fig. 2-b shows the number of COVID-19 cases per each 100000 individuals of the respective countries.

a. Ordinary form: Local alignment on ASA of each of 21756 pairs without any opening gap penalty but with −2 penalty score for extension gap. In similarity matrix for each match 3 credits were considered as its alignment score. For mismatch of intermediate against exposed 1 credit point and for intermediate against buried −1 was considered as its alignment score. For mismatch of exposed against buried −3 was considered as its alignment score. For each of the pairs, the sum of alignment scores was considered as ASA similarity score called *ASAO*.
b. Compact form: Local alignment on compact form of ASA with the parameters of previous step. ASA compact form was gained simply by eliminating the consequent repeats of each ASA types. As an example, with considering “EEEEEXXXXBBBBXXEEEEEE” as ASA of a protein, its ASA compact form will be “EXBXE”. For each of the pairs, the sum of the alignment score was considered as ASA similarity score called *ASAC*.

IV. Gene ontology semantic similarity score:

Semantic similarity of Gene ontology (GO) terms of two proteins can reveal functional similarity which increases the probability of the possible interaction between them. GO [47] comprises three domains: Molecular function (MF) which shows the biochemical activities of the gene products, cellular component (CC) which shows the active place of the gene in the cell, and biological process (BP) which shows the biological objectives in which the gene products participate. For each domain, there is a hierarchal directed acyclic graph in which each node represents a GO term. To find semantic similarity of GO terms of two proteins, Relevance method [48] was used which depends on the frequencies of each protein’s GO terms and GO terms’ frequencies of their common ancestor. For each of 21756 AIV-SV pairs, 3 semantic similarity scores were calculated called *BPS*, *MFS*, *CCS* and the mean of them were considered as GO semantic similarity score called *GOS*.

The final score is the harmonic mean of all the mentioned scores calculated by:

$$S_{c,i} = \frac{7+k}{(2/PS)+(1/SSO)+(1/SSC)+(1/SSLCS)+(1/ASAO)+(1/ASAC)+(k/GOS)}$$

As for some of the VPs, GO terms were not available, *k* was defined to show the number of domains having GO terms. To avoid division by zero problem, if there was not any valid GO term in all GO domains of a VP, *GOS* would be set to one and *k* would be set to zero.

Finally, a matrix of scores with 294 rows and 74 columns was created. The scoring matrix has the minimum score of 13, median score

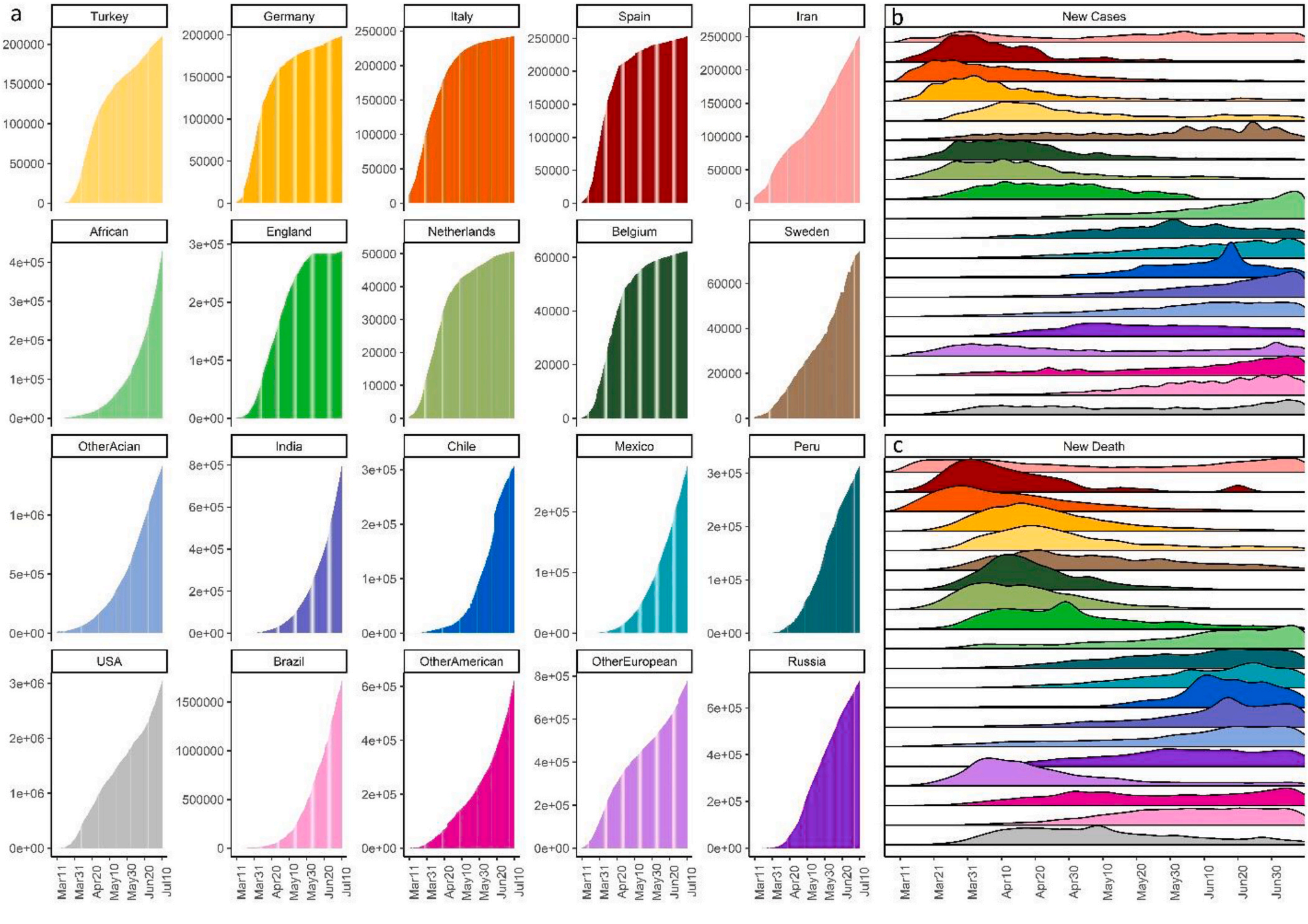

Fig. 3. Fig. 3-a shows the whole spreading distribution of COVID-19. Fig. 3-b and 3-c shows the daily spreading and death distributions, respectively.

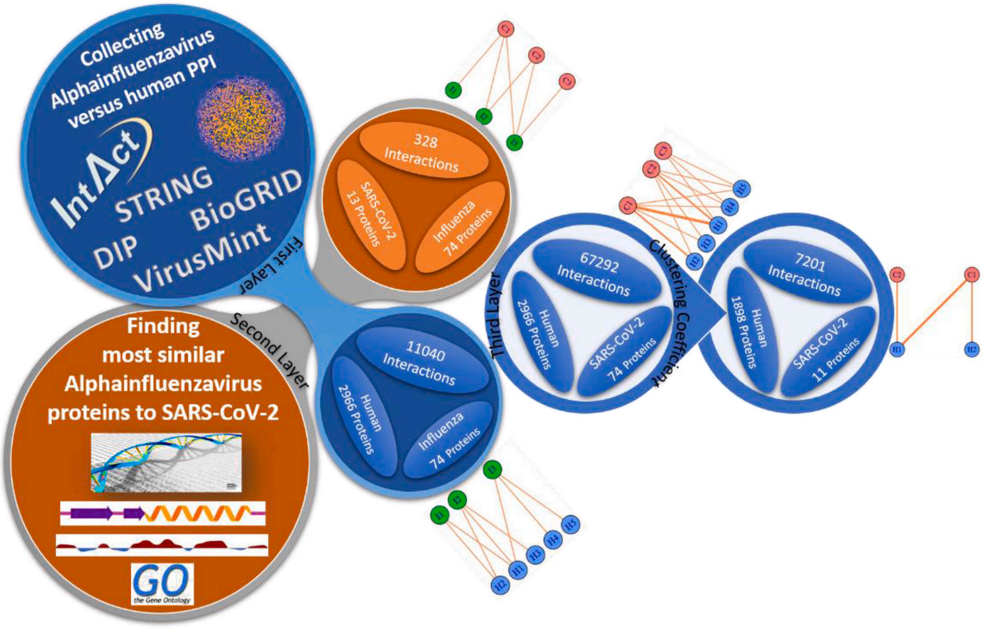

Fig. 4. Schematic view of predicting SARS-CoV-2-human protein-protein interaction network.

Table 1
SARS-CoV-2 orthologs.

ID	Protein	Length	Identity	Ortholog id	Protein	Taxid	Taxonomy
P0DTD1	Replicase polyprotein 1 ab	7096	95.7%	A0A2R3SV02	Non-structural polyprotein 1 ab	1508227	Bat SARS-like coronavirus
			95.3%	A0A2R3SUX5			
			86.2%	P0C6X7	Replicase polyprotein 1 ab	694009	Human SARS coronavirus
P0DTC1	Replicase polyprotein 1a	4405	95.6%	A0A2R3SV02	Non-structural polyprotein 1 ab	1508227	Bat SARS-like coronavirus
			95.1%	A0A2R3SUX5			
			80.7%	A0A0U1WJY1	Orf1ab polyprotein	1503302	BtRs-BetaCoV HuB2013
P0DTC2	Spike glycoprotein	1273	80.5%	A0A2R3SUW7	Spike protein	1508227	Bat SARS-like coronavirus
			79.8%	A0A2R3SUW9			
			77.1%	A0A0U2IWM2	Spike glycoprotein	1739625	SARS-like coronavirus
P0DTC3	Protein 3a	275	92.1%	A0A2R3SUX1		1508227	Bat SARS-like coronavirus
			90.9%	A0A2R3SUV9			
			76.1%	A0A023PTR5	Protein 3	1487703	Rhinolophus affinis coronavirus
P0DTC4	Envelope small membrane protein	75	100%	A0A2R3SUY7	Envelope small membrane protein	1508227	Bat SARS-like coronavirus
			94.7%	Q3I5J3	Envelope small membrane protein	349344	Bat coronavirus Rp3/2004
			94.7%	Q3LZW9	Envelope small membrane protein	442736	Bat coronavirus HKU3
P0DTC5	Membrane protein	222	98.6%	A0A2R3SUX3	Membrane protein	1508227	Bat SARS-like coronavirus
			91.7%	Q0Q472	Membrane protein	389167	Bat coronavirus 279/2005
			91.4%	Q3LZX9	Membrane protein	442736	Bat coronavirus HKU3
P0DTC6	Non- structural protein 6	61	93.4%	A0A2R3SUW5		1508227	Bat SARS-like coronavirus
			73.8%	U5WIP8		1415852	Bat SARS-like coronavirus WIV1
			73.8%	A0A0U2PPC8		1739625	SARS-like corona virus WIV16
P0DTC7	Protein 7a	121	88.4%	A0A2R3SUY1		1508227	Bat SARS-like coronavirus
			88.5%	Q3I5J0	Protein 7a	349344	Bat coronavirus Rp3/2004
			88.5%	Q3LZX7	Protein 7a	442736	Bat coronavirus HKU3
P0DTD8	Protein 7b	44	88.1%	A0A0U1WHL8		1503303	BtRs-BetaCoV YN2013
			85.7%	Q3I5I9	Protein 7b	349344	Bat coronavirus Rp3/2004
			85.7%	P0C5A9	Protein 7b	389167	Bat coronavirus 279/2005
P0DTC8	Non- structural protein 8	121	94.2%	A0A2R3SUZ9		1508227	Bat SARS-like coronavirus
			58.7%	D2DJX2		722424	SARS coronavirus Rs_672/2006
			58.7%	U5WI34		1415851	Bat SARS-like corona virus RsSHC014
P0DTC9	Nucleoprotein	419	94.3%	A0A2R3SUZ1	Nucleoprotein	1508227	Bat SARS-like coronavirus
			94.3%	A0A2R3SUX6			
			91.1%	A0A023PSY2	Nucleoprotein	1487703	Rhinolophus affinis coronavirus
P0DTD2	Protein 9b	97	76.5%	A0A023PUR2	Protein 13	1487703	Rhinolophus affinis coronavirus
			74.2%	Q3LZX3	Protein 9b	442736	Bat coronavirus HKU3
			74.2%	Q3LZU1	Protein N	338606	Bat coronavirus HKU3-3
P0DTD3	Protein 14	73	92.9%	A0A2R3SV09		1508227	Bat SARS-like coronavirus
			80.1%	Q3I5K1	ORF14	349342	Bat SARS coronavirus Rp1
			80.1%	Q3I5J8	ORF14	349343	Bat SARS coronavirus Rp2

of 49.5, third quantile score of 53.5 and maximum score of 72. We considered 53.5 (its third quantile score) as ortholog threshold. This threshold could be used for adjusting the size of predicted PPI network which has inverse relation with the size of PPI network. Increasing the threshold decreases the size of PPI network, while decreasing the threshold increases the number of PPIs. For each SV protein, AIV proteins with scores higher than 53.5 were considered as its orthologs.

For each SV protein, HPs' interactors of all its AIV orthologs were considered as its possible interactors. For each of the SV proteins and their HPs' interactors which were reached from the previous step, a bipartite graph was constructed.

In social networks, if A is friend of B and B is friend of G, there is a probable friendship relation between A and G. Now if A is friend of B, C, D, and E and B, C, D, and E are friends of G, the probability of friendship relation between A and G would be much higher. We use this rule for filtering the possible interactions constructed by the previous step.

Each SV protein may have several AIV orthologs, so each SV protein may be connected to an HP with multiple edges (each edge from one ortholog). To confine the number of predicted interactions, two filtration process were performed.

- Among all edges of the SV-human PPI network, single edges were eliminated.
- Multiple edges were converted to a simple edge and that edge was weighted by the sum of ortholog scores of its interactors minus whole of the median score (49.5). Edges with weight less than 10 were then eliminated.

As an example, shown in Fig. 5, consider I1, I2, and I3 as AIV proteins, C1, C2, and C3 as SV proteins and H1, H2, H3, H4, and H5 as Human proteins. Now consider I1, I2, and I3 are C1's orthologs with scores 58, 54, and 59 respectively. Moreover, consider I2 and I3 are C2's orthologs with scores 61 and 55. Finally, consider I3 is C3's ortholog with score 56. Suppose H1 and H2 were interactors of I1, and H1, H2, H3 were interactors of I2, and H1, H4, H5 were interactors of I3. The weight of edges is calculated in Table 2.

So, in the final SV-Human PPI network of our example, from 13 possible interactions (edges), just three interactions (C1–H1, C1–H2, and C2–H1) would be created.

Table 2
Calculation of edges' weight of the sample network.

Edge	Weight
C1–H1	(58–49.5) + (54–49.5) + (59–49.5) = 22.5
C1–H2	(58–49.5) + (54–49.5) = 13
C1–H3	(54–49.5) = 4.5
C1–H4	(59–49.5) = 9.5
C1–H5	(59–49.5) = 9.5
C2–H1	(54–49.5) + (59–49.5) = 14
C2–H2	(54–49.5) = 4.5
C2–H3	(54–49.5) = 4.5
C2–H4	(59–49.5) = 9.5
C2–H5	(59–49.5) = 9.5
C3–H1	(59–49.5) = 9.5
C3–H4	(59–49.5) = 9.5
C3–H5	(59–49.5) = 9.5

3. Results and discussion

Investigating SARS-CoV-2 genome and proteins leads to the following observations.

3.1. SARS-CoV-2 genome's GC-content is 38%

Probing SARS-CoV-2 genome, reveals that 29.9% of SARS-CoV-2 genome consists of adenine, 32% consists of thymine, 18.4% consists of cytosine, and 19.6% consists of guanine.

3.2. A, T, GT, TC, AC, GA, CAA, and GAA are microsatellites of SARS-CoV-2 with the most repeats

Tandem repeats, continuously repeated motifs, with the core subsequence less than 7 are called microsatellite which can help in structural analysis. All of SARS-CoV-2 microsatellites are extracted with FMSD [49] (fast microsatellite discovery method) and reported in Table 3.

3.3. Leucine, valine, alanine, threonine and serine are the most frequent amino acids of SARS-CoV-2 proteins

Fig. 6-a shows the frequency of each amino acid in SARS-CoV-2 proteins which is called amino acid composition. Fig. 6-b shows the frequency of each amino acid for all SARS-CoV-2 proteins.

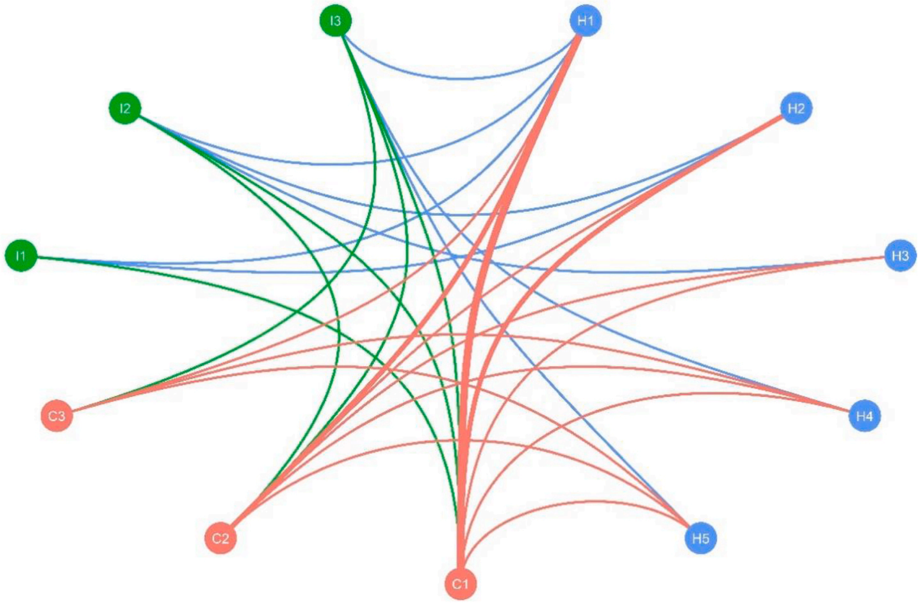

Fig. 5. A multipartite network in which the first layer shows the similarity between AIV proteins (I1, I2, and I3) and SV proteins (C1, C2, and C3) with green edges. The second layer shows the interactions between AIV proteins and human proteins (H1, H2, H3, H4, and H5) with blue edges. The third layer shows the possible interactions between SV proteins and human proteins with red edges. The thicknesses of the three red edge types shows that the ticker ones could be better candidates for SV-human PPI network. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 3
Microsatellites of SARS-CoV-2.

Core	Starts from	Repeats	Core	Starts from	Repeats	Core	Starts from	Repeats
A	29870	34	CTT	14756	3	TGT	25642	3
T	11074	8	TTC	22320	3	AAT	25757	3
TC	7813	5	AGT	23088	3	CGA	26191	3
GT	20486	5	AAG	3188	3	GTG	28556	3
AC	13162	4	GAT	3205	3	TGC	28934	3
GA	22954	4	CTT	4736	3	CAA	28987	3
GAA	3055	3	ATG	11366	3	CTG	29021	3
GAA	3073	3	ATC	11910	3	AAG	29389	3
TTC	626	3	TGA	13895	3			

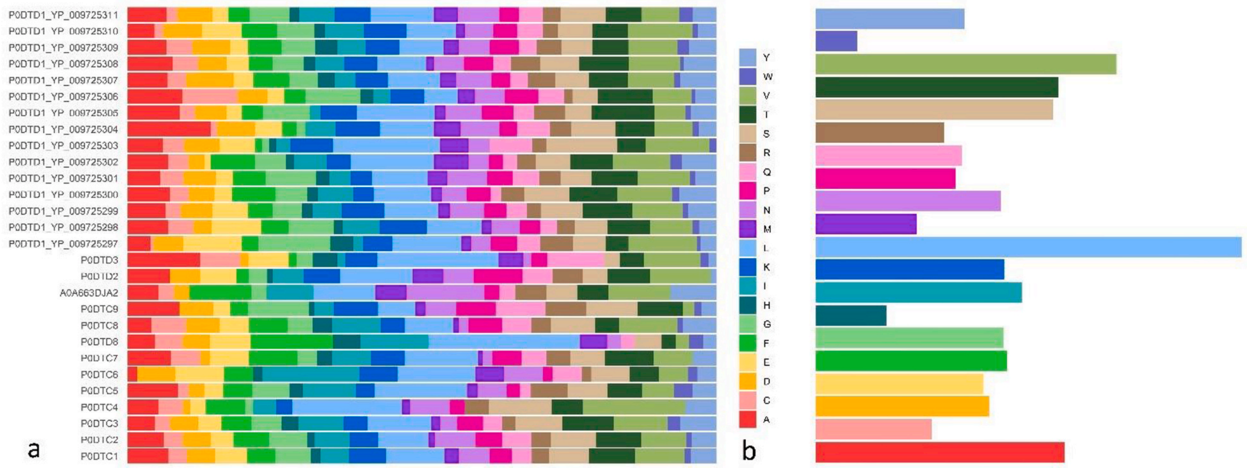

Fig. 6. Fig. 6-a shows the amino acid composition of SARS-CoV-2 proteins. Fig. 6-b shows the whole amino acid distribution of SARS-CoV-2 proteins.

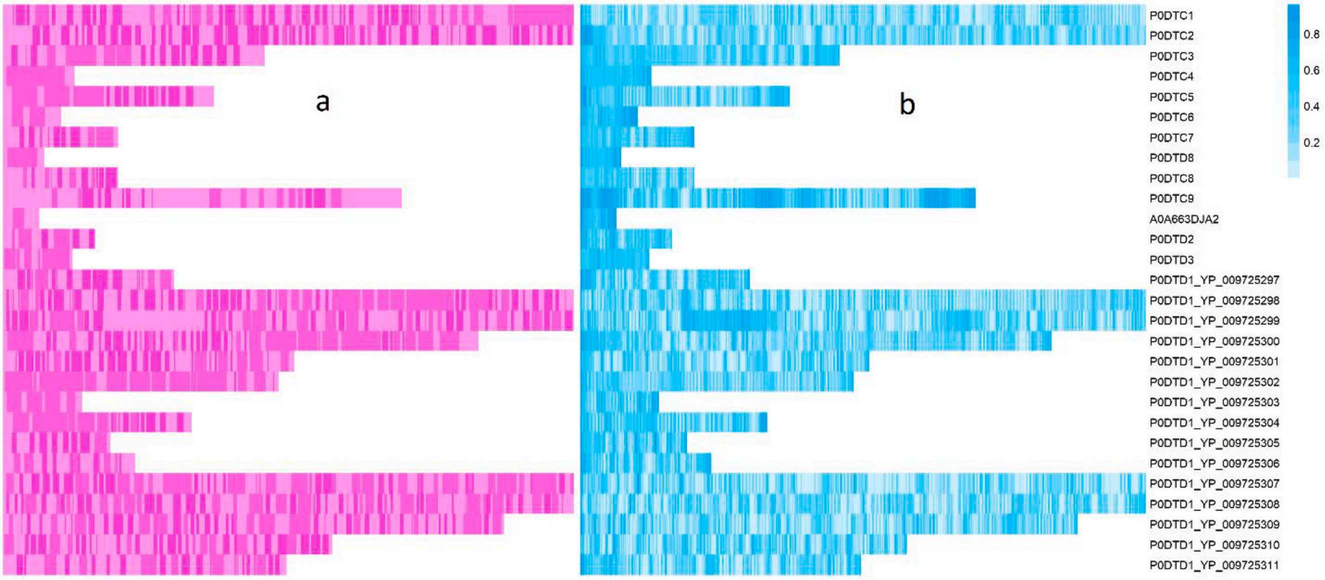

Fig. 7. Fig. 7-a shows the secondary structure of SARS-CoV-2 proteins. Light, medium, and dark pink declare coil, helix, and extended structures, respectively. Fig. 7-b shows the accessible surface area of SARS-CoV-2 proteins. Two lowest colors of gradient represent buried residues and the other colors of gradient show exposed residues. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3.4. Most of SARS-CoV-2 amino acids prefer to be exposed and have helix or coil structure

Secondary structures of SARA-CoV-2 proteins which could be used as feature vector in machine learning approach were extracted. Fig. 7-a shows the position of coils, helix, and extended in each of SARS-CoV-2 proteins. Moreover, accessible surface area was calculated for each of SARS-CoV-2 proteins. Values less than 0.2 were considered as buried, while values more than 0.2 were considered as exposed. As it is shown in Fig. 7-b, accessible surface area was depicted for all the amino acids of each SARS-CoV-2 proteins.

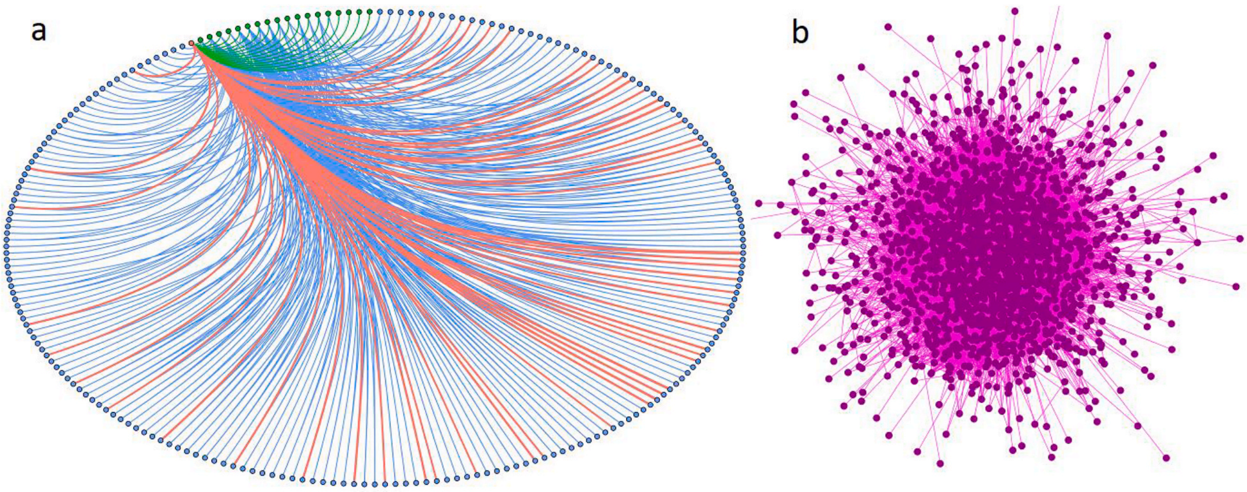

Fig. 8. Fig. 8-a shows *P0DTC7-human* protein-protein interaction network. *P0DTC7* proteins are shown with red nodes. Its IAV orthologs are shown with green nodes which have 389 interactions with 218 HPs (blue nodes). 49 HPs which have weights more than 10 and are connected to at least 2 IAV proteins were selected as final interactors of *P0DTC7* proteins and so the final *P0DTC7-human* PPI network is a bipartite network among the red nodes and 49 blue nodes which are connected by red edges. Fig. 8-b shows induced subgraph of human protein-protein interaction network of human proteins interacting with SARS-CoV-2 proteins. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3.5. SARS-CoV-2-human predicted PPI network has 7201 interactions between 11 VPs and 1898 HPs

P0DTC7 is one of the SARS-CoV-2 proteins. From 74 IAV proteins, 18 IAV proteins with similarity scores above 53 with *P0DTC7* were chosen as its orthologs which is shown with green color in Fig. 8a. These 18 AIV proteins have 389 interactions with 218 HPs which is shown with blue edges. Among these 218 HPs, 49 HPs were connected to at least two IAV proteins and have weights more than 10. So, they were chosen as the final HPs' interactors of *P0DTC7* and their PPI with *P0DTC7* is shown with red edges (see Fig. 8a).

Following the same strategy for all SV proteins leads to SV-human PPI network with 87894 interactions between 201 SV proteins and 2679 HPs. From the whole SV-human PPI network, 7201 interactions belong to SARS-CoV-2-human PPI network between 11 VPs and 1898 HPs.

3.6. Q86VP6, Q92905, Q13573, and P01106 are the most central nodes in human interactors of SARS-CoV-2

Human PPI network (HPPIN) has 261624 interactions between 19985 HPs. The induced subgraph of HPPIN of the 1898 HPs which

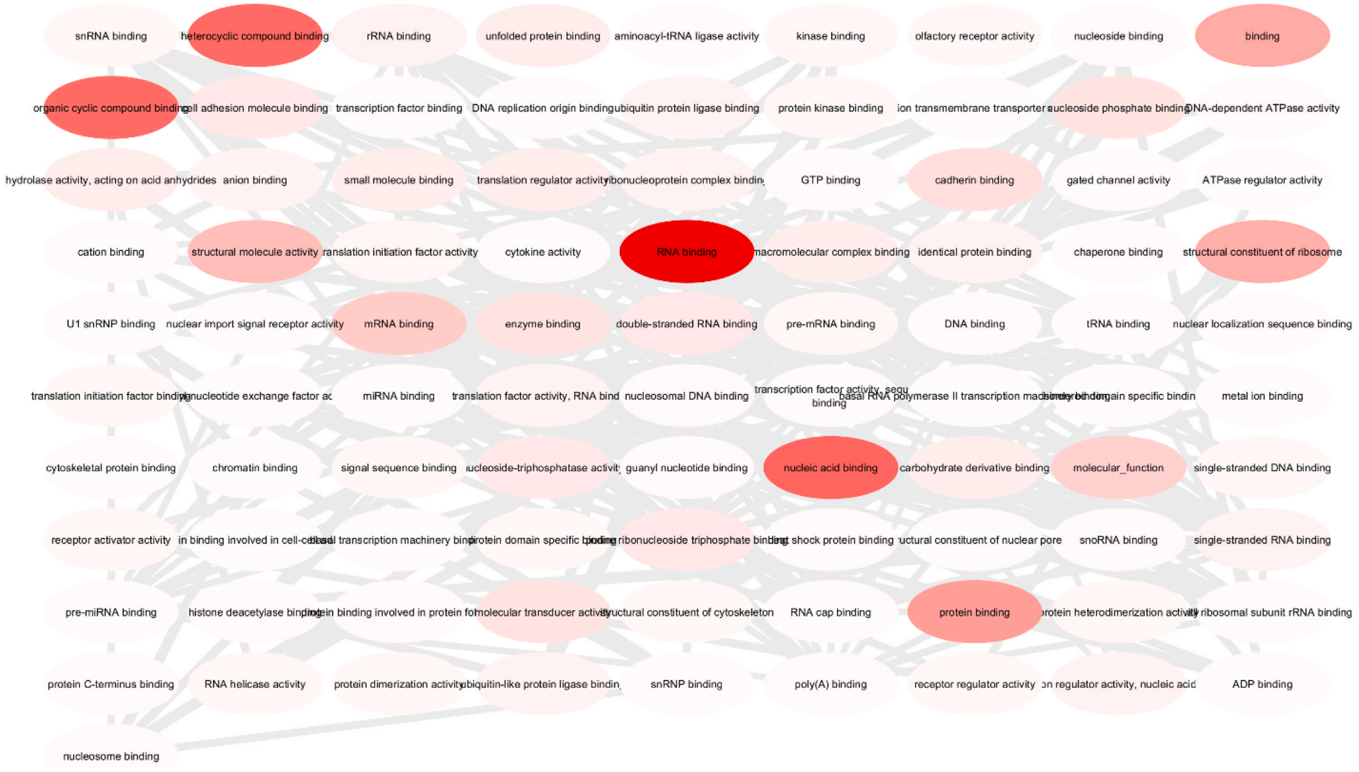
snRNA binding
heterocyclic compound binding
rRNA binding
unfolded protein binding
aminoacyl-tRNA ligase activity
kinase binding
olfactory receptor activity
nucleoside binding
binding
organic cyclic compound binding
cell adhesion molecule binding
transcription factor binding
DNA replication origin binding
ubiquitin protein ligase binding
protein kinase binding
ion transmembrane transporter
nucleoside phosphate binding
DNA-dependent ATPase activity
hydrolase activity, acting on acid anhydrides
anion binding
small molecule binding
translation regulator activity
ribonucleoprotein complex bindin
GTP binding
cadherin binding
gated channel activity
ATPase regulator activity
cation binding
structural molecule activity
ranslation initiation factor activity
cytokine activity
RNA binding
macromolecular complex binding
identical protein binding
chaperone binding
structural constituent of ribosome
U1 snRNP binding
nuclear import signal receptor activity
mRNA binding
enzyme binding
double-stranded RNA binding
pre-mRNA binding
DNA binding
tRNA binding
nuclear localization sequence binding
translation initiation factor binding
guanyl-nucleotide exchange factor ac
miRNA binding
translation factor activity, RNA bind
nucleosomal DNA binding
transcription factor activity, sequ binding
basal RNA polymerase II transcription machinery binding
protein domain specific bindin
metal ion binding
cytoskeletal protein binding
chromatin binding
signal sequence binding
nucleoside-triphosphatase activit
guanyl nucleotide binding
nucleic acid binding
carbohydrate derivative binding
molecular_function
single-stranded DNA binding
receptor activator activity
in binding involved in cell-cell
basal transcription machinery bin
protein domain specific binding
ribonucleoside triphosphate binding
heat shock protein binding
uctural constituent of nuclear pore
snoRNA binding
single-stranded RNA binding
pre-miRNA binding
histone deacetylase binding
protein binding involved in protein fo
molecular transducer activity
structural constituent of cytoskeleton
RNA cap binding
protein binding
protein heterodimerization activity
ribosomal subunit rRNA binding
protein C-terminus binding
RNA helicase activity
protein dimerization activity
ubiquitin-like protein ligase bindin
snRNP binding
poly(A) binding
receptor regulator activity
on regulator activity, nucleic acid
ADP binding
nucleosome binding

Fig. 9. 122 enriched molecular functions of human proteins targeted by SARS-CoV-2 proteins.

interact with SARS-CoV-2 is called HPPINS depicted in Fig. 8b.

HPPINS has 18342 interactions between 1756 HPs. By calculating different centrality measures, the following HPs were the most important nodes of the HPPINS which can be good candidates for experimental PPI test.

Q86VP6 (*cullin-associated NEDD8-dissociated protein 1*) with the degree of 392 (highest degree), closeness of 0.52, radiality of 0.84, and betweenness of 0.06 interacts with *P0DTC1, P0DTC4,* and *P0DTC6*.

Q92905 (*COP9 signalosome complex subunit 5*) with the degree of 365, closeness of 0.52, radiality of 0.84 (highest radiality) and betweenness of 0.06 interacts with *P0DTC4, P0DTC6, P0DTD2, P0DTD8,* and *A0A663DJA2*.

P01106 (*myc proto-oncogene protein*) with the degree of 311, closeness of 0.5, radiality of 0.83 and betweenness of 0.07 (highest betweenness) interacts with *A0A663DJA2*.

Q13573 (*SNW domain-containing protein 1*) with the degree of 307, closeness of 0.52 (highest closeness), radiality of 0.83 and betweenness of 0.045 interacts with *P0DTC4, P0DTD8,* and *P0DTD1*.

3.7. GO enrichment analysis

SARS-CoV-2-human PPI network has 1898 HPs out of which 1130 HPs interacts with at least two SARS-CoV-2 proteins. GO enrichment analysis were performed with PANTHER classification system [50] on these 1130 HPs in three separate classes for detecting enriched biological process, molecular function, and cellular component. 122 enriched molecular function, 199 enriched cellular components, and 748 enriched biological process with p-value less than 0.01 were extracted. Fig. 9 shows enriched molecular functions which is depicted with REVIGO [51].

3.8. 727 interactions of SARS-CoV-2-human PPI network belongs to 215 differentially expressed HPs

GSE150316 is an expression profiling by high throughput sequencing experiment on five COVID-19 positive patients and five negative control ones in five different organs. We compare the gene expression data of each organ between positive patients and negative controls and report genes with their log2 fold changes higher than one (over expressed at least two times), as differentially expressed genes (DEGs). Thereafter, we search these DEGs among HPs targeted by SARS-CoV-2 proteins in our predicted SV-human network and marked them. Twenty DEGs are detected in lung which make 255 interactions in SV-human PPI network. Ninety-five DEGs are detected in heart which make 2099 interactions in SV-human PPI network. Nine DEGs are detected in liver which make 104 interactions in SV-human PPI network. Six DEGs are detected in kidney which make 27 interactions in SV-human PPI network. And finally, thirty-five DEGs are detected in bowel which make 634 interactions in SV-human PPI network.

GSE1739 [52] is an expression profiling by array experiment on ten SARS patients and four negative controls. We compared the gene expression data of positive patients and negative controls and report genes with their log2 fold changes higher than one as DEGs. Thereafter, we search these DEGs among HPs of our predicted SV-human network and marked them. One hundred sixty-eight DEGs make 6319 interactions in SV-human PPI network.

Out of these 9072 interactions of SV-human PPI network, 727 interactions belong to SARS-CoV-2-human PPI network between 215 HPs and SARS-CoV-2 proteins. We marked all of these interactions in our database.

4. Conclusion

In the current study, by collecting 180 reports from the world health organization, we demonstrate how COVID-19 reached a pandemic state all over the world. Then, we investigated SARS-CoV-2 orthologs. As all of its orthologs belong to Sarbecovirus, we decided to work on the whole Sarbecovirus proteins. Initially, we clustered its proteins according to their length and eliminated proteins with high sequence identities within each group. Thereafter, we found similar Alphainfluenzavirus proteins by primary structure, secondary structure, accessibility, and gene ontology semantic similarities. And finally, we made a weighted Sarbecovirus-human protein-protein interaction network by connecting Sarbecovirus proteins to human proteins, which have interactions with at least two of their similar Alphainfluenzavirus proteins and weighted them according to their clustering coefficient. Our final dataset contains 87894 protein-protein interactions between Sarbecovirus and human proteins. The first 7201 interactions belong to SARS-CoV-2-human protein-protein interactions. The constructed weighted protein-protein interaction network is publicly available at http://bioinf.modares.ac.ir/software/complexnet/Corona/CoronaPPIN.txt.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Khani H, Tabarraei A, Moradi A. Survey of coronaviruses infection among patients with flu-like symptoms in the golestan province, Iran TT -. mljgoums Nov. 2018;12(6):1–4.

[2] Qu J, Wickramasinghe C. SARS, MERS and the sunspot cycle. Curr Sci 2017;113(8):1501.

[3] Luk HKH, Li X, Fung J, Lau SKP, Woo PCY. Molecular epidemiology, evolution and phylogeny of SARS coronavirus. Infect Genet Evol 2019;71:21–30.

[4] Organization WH. Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. 2003.

[5] Dudas G, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-human interface. Elife 2018;7:e31257.

[6] Huang C, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;395(10223):497–506.

[7] Remuzzi A, Remuzzi G. COVID-19 and Italy: what next? Lancet 2020.

[8] Zhou P, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020.

[9] Zhu H, et al. Host and Infectivity Prediction of Wuhan 2019 Novel Coronavirus Using Deep Learning Algorithm. 2020. bioRxiv.

[10] Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. Lancet 2020;395(10223):470–3.

[11] Bonilla-Aldana DK, et al. "Coronavirus infections reported by ProMED, february 2000–January 2020. Trav Med Infect Dis 2020:101575.

[12] Li G, De Clercq E. Therapeutic Options for the 2019 Novel Coronavirus (2019-nCoV). Nature Publishing Group; 2020.

[13] Cheng ZJ, Shan J. 2019 Novel coronavirus: where we are and what we know. Infection 2020.

[14] Chen Y, Liu Q, Guo D. Emerging coronaviruses: genome structure, replication, and pathogenesis. J Med Virol 2020.

[15] Su S, et al. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol 2016;24(6):490–502.

[16] Hu CD, Chinenov Y, Kerppola TK. Visualization of interactions among bZIP and Rel family proteins in living cells using bimolecular fluorescence complementation. Mol Cell 2002;9(4):789–98.

[17] Golemis E. Protein-protein interactions: a molecular cloning manual. CSHL Press; 2005.

[18] Sun T, Zhou B, Lai L, Pei J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinf 2017;18(1):1–8.

[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.

[20] Chatterjee P, Basu S, Kundu M, Nasipuri M, Plewczynski D. PPI_SVM: prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables. Cell Mol Biol Lett 2011;16(2):264–78.

[21] Mei S. Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins. PloS One 2013;8(11):e79606.

[22] Eid F-E, ElHefnawi M, Heath LS. DeNovo: virus-host sequence-based protein–protein interaction prediction. Bioinformatics 2016;32(8):1144–50.

[23] Nourani E, Khunjush F, Sevilgen FE. Virus–human protein–protein interaction prediction using Bayesian matrix factorization and projection techniques. Biocybern Biomed Eng 2018;38(3):574–85.

[24] Basit AH, Abbasi WA, Asif A, Gull S, Minhas FUAA. Training host-pathogen protein–protein interaction predictors. J Bioinf Comput Biol 2018;16(4):1850014.

[25] Yang X, Yang S, Li Q, Wuchty S, Zhang Z. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. Comput Struct Biotechnol J 2020;18:153–61.

[26] Barman RK, Saha S, Das S. Prediction of interactions between viral and host proteins using supervised machine learning methods. PloS One 2014;9(11): e112034.

[27] Leite DMC, Brochet X, Resch G, Que Y-A, Neves A, Peña-Reyes C. Computational prediction of host-pathogen interactions through omics data analysis and machine learning. In: International Conference on Bioinformatics and Biomedical Engineering; 2017. p. 360–71.

[28] Zahiri J, Khorsand B, Yousefi A, Kargar M, Zade RSH, Mahdevar G. AntAngioCOOL: Computational detection of anti-angiogenic peptides. J. Transl. Med. 2019;17(1).

[29] Khorsand B, Savadi A, Zahiri J, Naghibzadeh M. Alpha influenza virus infiltration prediction using virus-human protein-protein interaction network. Math. Biosci. Eng. 2020;17(4):3109–29.

[30] Ray S, Alberuni S, Maulik U. Computational prediction of HCV-human protein-protein interaction via topological analysis of HCV infected PPI modules. IEEE Trans Nanobiosci 2018;17(1):55–61.

[31] Chen J, Sun J, Liu X, Liu F, Liu R, Wang J. Structure-based prediction of West Nile virus-human protein–protein interactions. J Biomol Struct Dyn 2019;37(9): 2310–21.

[32] Durán AH, Greco TM, Vollmer B, Cristea IM, Grünewald K, Topf M. Protein interactions and consensus clustering analysis uncover insights into herpesvirus virion structure and function relationships. PLoS Biol 2019;17(6):e3000316.

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

[34] Kösesoy İ, Gök M, Öz C. "A new sequence based encoding for prediction of host–pathogen protein interactions. Comput Biol Chem 2019;78:170–7.

[35] Mir A, Naghibzadeh M, Saadati N. INDEX: incremental depth extension approach for protein–protein interaction networks alignment. Biosystems 2017;162:24–34.

[36] Guven-Maiorov E, Tsai C-J, Ma B, Nussinov R. Interface-based structural prediction of novel host-pathogen interactions. In: Computational methods in protein evolution. Springer; 2019. p. 317–35.

[37] Wu F, et al. Complete Genome Characterisation of a Novel Coronavirus Associated with Severe Human Respiratory Disease in Wuhan, China. bioRxiv. 2020. 01.24.919183, Jan. 2020.

[38] Consortium U. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47(D1):D506–15.

[39] Munster VJ, Koopmans M, van Doremalen N, van Riel D, de Wit E. A novel coronavirus emerging in China—key questions for impact assessment. N Engl J Med 2020;382(8):692–4.

[40] Kerrien S, et al. The IntAct molecular interaction database in 2012. Nucleic Acids Res 2011:gkr1088.

[41] Chatr-aryamontri A, et al. VirusMINT: a viral protein interaction database. Nucleic Acids Res Jan. 2009;37:D669–73. Database issue.

[42] Szklarczyk D, et al. "STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res 2014;43(D1):D447–52.

[43] Xenarios I, Salwinski L, Duan XJ, Higney P, Kim S-M, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 2002;30(1):303–5.

[44] Oughtred R, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Res Nov. 2018;47(D1):D529–41.

[45] Prasad DVV, Jaganathan S. "Improving the performance of smith–waterman sequence algorithm on gpu using shared memory for biological protein sequences. Cluster Comput 2019;22(4):9495–504.

[46] Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum allowed solvent accessibilites of residues in proteins. PloS One Nov. 2013;8(11). e80635–e80635.

[47] Consortium GO. Expansion of the gene ontology knowledgebase and resources. Nucleic Acids Res 2017;45(D1):D331–8.

[48] Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinf 2006;7 (1):302.

[49] Naghibzadeh M, Savari H, Savadi A, Saadati N, Mehrazin E. Developing an ultra-efficient microsatellite discoverer to find structural differences between SARS-CoV-1 and Covid-19. Inf Med 2020:100356. Unlocked.

[50] Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res 2019;47(D1):D419–26.

[51] Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PloS One 2011;6(7):e21800.

[52] Reghunathan R, et al. Expression profile of immune response genes in patients with severe acute respiratory syndrome. BMC Immunol 2005;6(1):2.