Intelligent Decision Technologies 16 (2022) 145–158
DOI 10.3233/IDT-200185
IOS Press

Graph analysis and clustering of proteins linked with COVID-19

J. Susymary[a,*] and P. Deepalakshmi[b]
[a]*Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankoil, Virdhunagar District, Tamil Nadu, India*
[b]*Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Virdhunagar District, Tamil Nadu, India*

Abstract. A remarkable number of scientific initiatives are in practice to encounter the new coronavirus epidemic (COVID-19). One of the biggest challenges faced by the COVID-19 researchers in the therapeutic field is the knowledge about the biological functions in disease-human interacting proteins. The detection of COVID-19 protein complexes, a group of proteins that possess the same biological functions, helps in better understanding of the biological processes in our body. The main contribution of this work is to cluster proteins that perform the same biological functions to increase the knowledge about the COVID-19 disease-human interacting proteins. The authors investigated proteins linked with COVID-19 disease by creating a disease-human protein-protein interaction graph. Topological means of graph analysis and graph clustering have been employed to group proteins that possess the same biological functions. These clusters will be the protein complexes that work together to carry out a specific biological function in a human cell. Moreover, through the cluster analysis, we can uncover previously unknown COVID-19 disease-human protein links that are beneficial for promising knowledge discovery. Also, the authors evaluated how the Markov Cluster algorithm, a graph-based algorithm finds interesting patterns of similar features from COVID-19 disease-human protein-protein interaction graph. The Markov Cluster algorithm results in six statistically significant protein clusters, including cluster (A): keratinization (3.50E-71), (B): regulation of cellular process (6.62E-05), (C): regulation of cell cycle (1.31E-27), (D): mitotic cell cycle (1.66E-06), (E): regulation of phosphoprotein phosphatase activity (1.15E-09), and (G): G2/M transition of mitotic cell cycle (3.03E-07).

Keywords: COVID-19, protein-protein interaction, graph clustering, Markov Cluster algorithm, protein clusters

1. Introduction

The current pandemic of COVID-19, a respiratory disease emerged in late 2019 has led to 1,051,635 confirmed cases and 56,985 fatalities in 208 countries with cases as of 4th April, 2020 [1]. The COVID-19 is induced by a new virus which causes severe acute respiratory syndrome – 2 (SARS-2) of the Coronaviridae family [2]. The history of coronavirus unfolds the episode of SARS-CoV in 2002 with 8000 confirmed cases and 10% fatality. Similarly, another event in Middle East respiratory syndrome coronavirus (MERS-CoV) in 2012 had 2500 confirmed cases and 36% fatality rate [3]. COVID-19 spreads with a higher fatality, and makes it hard to contain the disease and also escalates its pandemic potentiality [3]. From Latin, "corona" means crown. This remarkable pathogenic virus attack returns common flu to acute respiratory infections that can start long-term reduction in lung function, and proceed to death [4–7]. Common flu spreads more quickly, and the fatality rate is low [8]. It is crucial to develop an understanding of how coronavirus proteins attack human proteins during infection, to devise therapeutic strategies to counteract COVID-19. This knowledge will be useful and can be applied to develop drugs and repurpose currently used ones. So far, no antiviral medications

*Corresponding author: J. Susymary, Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankoil, Virdhunagar District, Tamil Nadu, 626126, India. E-mail: susymaryj@gmail.com.

ISSN 1872-4981/$35.00

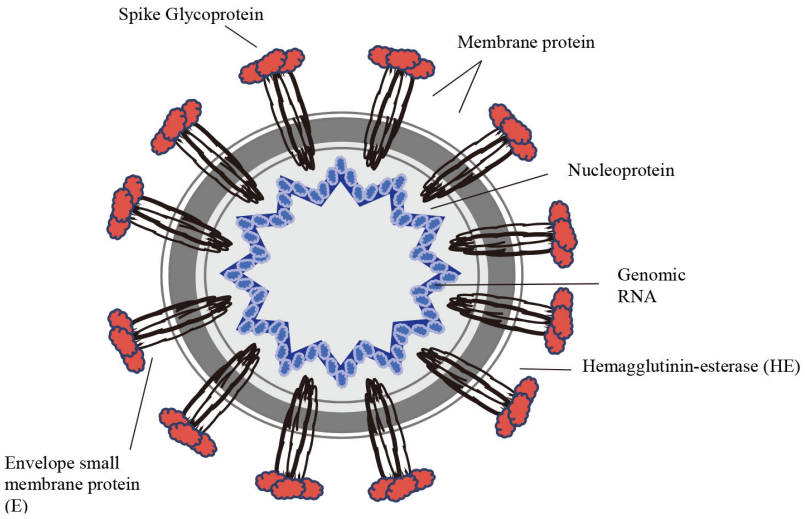

Fig. 1. The Virion structure.

have been developed for COVID-19. The COVID-19 genome encodes with 16 non-structural proteins NSP1 to NSP16, four structural proteins E, M, N, and S, and eleven accessory factors, ORF3A, ORF3B, ORF6, ORF7A, ORF8, ORF9B, ORF9C, ORF10, ORF1AB, ORF1A, and ORF7B [9]. Table 1 shows the known COVID-19 protein names and description. Researchers are examining the presumed coronavirus proteins to develop drugs and vaccines in this pandemic. The Virion structure shows spikes projecting from the envelope that looks like a crown (Fig. 1) [10].

Protein-Protein Interactions (PPIs) present many challenges for the identification of drug-like molecules [11]. Both traditional as well as some innovative strategies like graph-based methods provided valuable tools for the discovery of PPI modulators and its potential of targeting PPIs for therapeutic intervention. We cannot understand the biological processes happening inside our body without extensive analysis of COVID-19 disease-human protein interactions [12]. The main objective of this work is to find protein clusters linked with COVID-19 disease-human interactions. Clustering proteins with the same biological functions help biomedical researchers to explore knowledge about the natural process happens in our body due to this disease. We propose the analysis of the structural and non-structural proteins as the seeds to construct a protein-protein interaction (PPI) network associated with COVID-19 [9]. In a PPI network, vertices or nodes represent proteins, and edges represent interactions. Graph cluster analysis with a blend of topological properties of the PPI network provides an appropriate biological knowledge for a promising tool to understand the biological function of the protein groups.

Further sections of this paper are in the following order: in Section 2, the authors did a literature study on the creation and investigation of PPI networks, Section 3 discusses how a PPI data can be modelled as a graph and the modelling algorithm which we have used to create the COVID-19 – human PPI data, Section 4 explains the graph clustering algorithm, Markov Cluster (MCL), Section 5 discloses the cluster validity measures, Section 6 describes the methodology of our analysis, Section 7 canvas the results and discussion, Section 8 covers possibilities, limitations, and future study, and finally, we conclude our study in Section 9.

2. Literature study

The literature study helps to accomplish a theoretical base for the research problem by seeking the past events of the particular subject. In this literature section, we cut across the knowledge about graph clustering and topological analysis of PPI networks. Topological means of PPI network analysis provide a platform for exploring complex diseases [13–15]. Network-based computational models to analyse COVID-19 disease – human protein interaction network – has already been applied by researchers in drug and therapeutics [16]. Studies used significant gene bio-signatures as the seeds to build the PPI network and analyse disease dynamics through topological analysis of the PPI network [17]. Several models for protein complex detection from PPI

Table 1
Summary of graph clustering methods

Reference	Objective	Clustering method	Clustering algorithm	Application
[18]	Protein complex detection	Stochastic search method	HGCA	Biological networks
[20]	Community detection	Stochastic search method	Louvain cluster	Biological networks
[19]	Protein complex detection	Local neighbourhood density search	MCODE	Biological networks
	Protein complex detection	Flow-based simulation	MCL	Biological networks
[21]	Protein clusters	Flow-based simulation	FOA-MCL	Biological networks
[22]	Operon prediction	Flow-based simulation	MCL	Biological networks
[23]	Find essential proteins	Flow-based simulation	RWEP	Biological networks

network has been applied in graph-based clustering techniques [18,19].

Graph clustering using community detection algorithms follows a stochastic search method based on the seed vertex or edges [18,20]. A heuristic graph clustering algorithm (HGCA) based on different topological properties has been proposed for protein complex detection [18]. The algorithm used a weighted degree for edges and vertices. A cluster description model is then constructed based on the candidate vertex and the cluster. Based on this description model, the HGCA starts with the seed vertex and produces communities in a greedy manner. A study with six community detection algorithms was used to analyse two biological networks and evaluated the resultant communities [20]. Among the six algorithms, Louvain cluster algorithm is found to be the fastest with an agglomerative approach to maximise the modularity. Density-based local neighbourhood search method and flow simulation method are widely used in protein complex detection and analysing PPI networks [19,21–24]. A demonstration of how MCL and Molecular Complex Detection algorithm (MCODE) has been done to identify patterns from PPI data related to Alzheimer's disease [19]. A combination of MCL algorithm with Fruit Fly Optimization Algorithm (FOA), FOA-MCL has been used to find clusters formed by PPI network data of human immunodeficiency virus (HIV) [21]. An operon prediction model based on Markov clustering algorithm used some generic attribute information of genomes for graph [22]. The results show that the operon model has a better capability of operon prediction than classical operon prediction methods. The Random Walk Essential Proteins (RWEP) is a method that endorses random walks with restart that incorporate the topological and biological properties [23]. It has been applied to rule out protein essentiality in PPI networks.

Stochastic search clustering method is extensively used in PPI networks where the objects are considered as vertices. This method is time-consuming and, therefore, suitable for small networks. The flow-based method works on the principle of random walks and has a tendency to stay within clusters rather than between clusters. This method will not produce overlapping clusters. However, the process is time-consuming. Local neighbourhood density search method performs graph clustering by recognising seed proteins as individual clusters and then proceed greedily to add vertices.

Table 1 shows a summary of the reviewed graph clustering methods. Graph clustering and analysis using the clustering co-efficient as the validity measure provides a valuable tool for the partitioning of the PPI network [19,25,26]. However, the MCL algorithm is mainly designed for graphs and can be applied to biological applications [19,27]. Novel algorithms for analysing PPI networks by combining MCL and optimisation techniques have also been developed [21].

3. Protein-protein interaction graph

A PPI network is a collection of protein interactions, often deposited in online databases [28–30]. Since proteins interact with each other and also carries information signals from one protein to another, understanding critical biological processes in the human body will be difficult without an extensive analysis of PPI [30]. Analysing PPI networks helps to mine data that assure one to create biological systems with new properties, and protein complexes for therapeutic purposes [16,18,31–34].

Concerning topology, the PPI networks follow a small-world property and are scale-free networks [35, 36]. In small-world networks, it is possible to reach from a protein to any other protein in only a small number of steps. In scale-free networks, most proteins have a reduced number of interactions. Among the biological graph drawing algorithms, the force-directed layout model is very much flexible [37]. The algorithm can be used to calculate undirected graphs by using the information contained within the structure of the graph. There are several force-driven algorithms. Since the fundamental and aesthetic goal of optimisation of PPI graph drawing algorithms is the min-

imisation of crossings between edges, reducing the distance between edges and incorporating the properties such as small-world effect, scale-free network, the force-directed model is suitable for drawing a PPI graph [38,39].

Generally, a PPI network can be created as an undirected, unweighted graph $G = (V, E)$ where V is a group of proteins and E is a group of interactions between the proteins. The main data structure [40] used to store network representation is the adjacency matrix.

Let G be a graph such that $V(G) = \{v1, v2, \ldots, vn\}$, the adjacency matrix representation of G is a $n \times n$ matrix. If $a_{r,c}$ is the value in the matrix A, at row r and column c, then $a_{r,c} = 1$; if v_r is adjacent to v_c; otherwise, $a_{r,c} = 0$. Adjacency matrices require space of $\Theta\left(|n|^2\right)$.

The Kamada-Kawai [41] suggested that the number of edge crossings for a PPI graph is not a benchmark for a layout algorithm. This layout algorithm measures the total balance of the PPI graph as the square summation of the differences between the ideal distance and the actual distance for the entire vertices. The calculation is delineated in Eq. (1).

$$\textit{Stress}\,(x) = \sum_{i<j} w_{ij}\left(||x_i - x_j|| - d_{ij}\right)^2 \qquad (1)$$

For some pair of vertices, i and j, where d_{ij} is the standard distance amidst vertices, x is the set of coordinates and $w_{ij} = d_{ij}^{-\alpha}$. The Kamada-Kawai approach preserves the total balance of the PPI graph, and deliver layouts with minimum edge crossings by approximation and minimisation of stress.

3.1. Pseudocode for the Kamada-Kawai algorithm

Compute d_{ij} *for* $1 \leqslant i \neq j \leqslant n$;
Compute l_{ij} *for* $1 \leqslant i \neq j \leqslant n$;
Compute k_{ij} *for* $1 \leqslant i \neq j \leqslant n$;
Initialize $p_1, p_2, \ldots, p_n$;
While $\max_i \Delta_i > \epsilon\{$
Let p_m *be the particle satisfying* $\Delta_m = \max_i \Delta_i$;
While $(\Delta_m >)\{$
Compute δ_x *and* δ_y *by solving Eqs* (1) *and* (2);
$x_m = x_m + \delta_x$;
$y_m = y_m + \delta_y$; $\}\}$

Let $p_1, p_2, \ldots, p_n$ be the particles in a plane that matches the vertices $v_1, v_2, \ldots, v_n \in V$. d_{ij} is the distance amidst two vertices v_i and v_j. l_{ij} is the length of the shortest path between v_i and v_j. l_{ij} is defined as $l_{ij} = L \times d_{ij}$ where L is the seductive length of a single edge in the display plane. k_{ij} is the strength of the spring amidst p_i and p_j and is driven as $k_{ij} = \frac{K}{d_{ij}^2}$, where K is a constant.

The location of a particle in a plane is disclosed by x and y coordinate values. Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be the coordinate variables of particles $p_1, p_2, \ldots, p_n$ respectively. The energy E is delineated as in Eq. (2).

$$E = \sum_{i=I}^{n-I} \sum_{j=i+1} \frac{1}{2} k_{ij}\{(x_i - x_j)^2 + (y_i - y_j)^2 + l_{ij}^2 - 2l_{ij}\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}\} \qquad (2)$$

Δ_m is the local minimum and is calculated using Newton-Raphson method.

δ_x and δ_y can be calculated by solving the following Eqs (3) and (4).

$$\frac{\partial^2 E}{\partial x_m^2}\left(x_m^{(t)}, y_m^{(t)}\right)\delta_x + \frac{\partial^2 E}{\partial x_m \partial y_m}\left(x_m^{(t)}, y_m^{(t)}\right)\delta_y = -\frac{\partial E}{\partial x_m}\left(x_m^{(t)}, y_m^{(t)}\right) \qquad (3)$$

$$\frac{\partial^2 E}{\partial x_m \partial y_m}\left(x_m^{(t)}, y_m^{(t)}\right)\delta_x + \frac{\partial^2 E}{\partial y_m^2}\left(x_m^{(t)}, y_m^{(t)}\right)\delta_y = -\frac{\partial E}{\partial y_m}\left(x_m^{(t)}, y_m^{(t)}\right) \qquad (4)$$

In a graph, the least level of organisation is vertices and degree of vertices. Vertices are connected by edges to form motifs, sub-graphs of three or more vertices. Motifs are linked to form communities or complexes. A sub-graph is a graph formed from the disjoint union of complete graphs. Graph topology statistics include scale-free properties to fit power-law feature: average degree, degree distribution, small world properties that can be measured by clustering coefficient, average path length.

4. Graph clustering

Graph clustering is an augmenting area with a perspective to discover contemporary facts from complex data that can be pictured as a graph. The field of clustering has grown, and the number of clustering algorithms reported in biological applications is also high [19,21,42–44]. Graph clustering has two perspectives: intra-graph clustering and inter-graph clustering. Intra-graph clustering is the process of grouping objects within a single graph and inter-graph clustering method clusters between graphs. The intra-graph clustering method focus on both vector-based and graph-

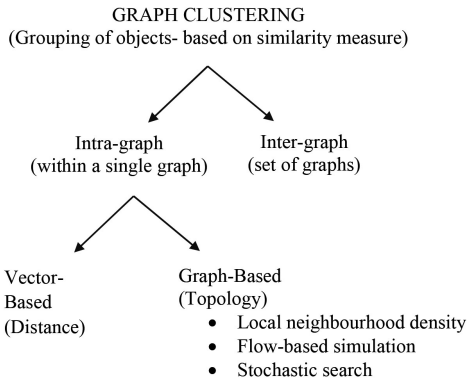

Fig. 2. Perspective of graph clustering.

based. Vector-based clustering is of purely distance-based and graph-based clustering is established on the topological characteristics of the graph.

Graph-based clustering techniques for biological application explicitly uses theoretical graph terms to cluster the data that is represented as a graph. Local neighbourhood density search, flow-based simulation, and stochastic search are some of the graph-based clustering methods.

Figure 2 shows the perspective of graph clustering. Flow simulation method based on a random walk [45], or biological knowledge can be applied to PPI networks to uncover protein clusters. The MCL [27] is a flow simulation method based on a random walk. Moreover, the MCL is one of the most successful approach to cluster proteins in PPI networks [22,46].

5. Measures for cluster validation

Clustering is an unsupervised learning technique and gives distinct clustering results on the same data with distinct parameters. Clustering coefficient is one of the metrics available for some indication of the quality of the clusters [19,25,26,47–49]. The clustering coefficient lies between 0.0 and 1.0. If the clustering coefficient tends to 1, the neighbourhood is fully connected, and graph possesses a maximal structure [50].

The dimension that represents the trend of a graph to be segregated into clusters is known as the clustering coefficient. A cluster is a subspace of vertices connected by edges.

Let v, be a vertex and let e_v be the number of edges joining the j_n neighbours of n. The clustering coefficient C_n of the vertex n, is delineated in the Eq. (5).

$$C_n = \frac{2v_n}{j_n}(j_n - 1) \tag{5}$$

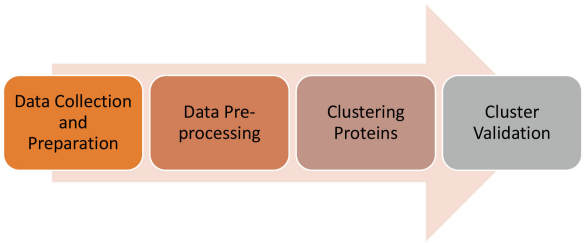

Fig. 3. The working process of the proposed model.

The clustering coefficient of C_{D_i} of a cluster D_i is the average clustering coefficients of the entire proteins contained in D_i. The clustering coefficient C_D of the clusters $D = \{D_1, \ldots, D_k\}$ is defined in the Eq. (6).

$$C_D = \frac{\sum_{i=1}^{k} C_{D_i}}{k} \tag{6}$$

For a biological network, clusters need to be validated using the domain knowledge to ensure the natural function of the objects. The most efficient way of biological validation is to test the gene ontology for the enrichment of classified clusters, and to check if proteins are functionally homogeneous.

Gene Ontology [51] is a database that consists of three categories of associations, namely, molecular function, cellular component, and biological process. The functions of proteins are labelled with a GO-term. A p-value to calibrate the biological annotation is computed with the hyper-geometric principle that is delineated in Eq. (7).

$$p\text{-}value = \frac{\begin{bmatrix} |P_{su}| \\ |S_{su}| \end{bmatrix} \begin{bmatrix} |P| - |P_{su}| \\ |S| - |S_{su}| \end{bmatrix}}{\begin{bmatrix} |P| \\ |S| \end{bmatrix}} \tag{7}$$

The proteins in the whole network and its linked GO-term are captured as the population (P) for the hyper-geometric test. The proteins in the cluster with their matching GO-term is captured as the sample (S). This is to examine whether a specific GO-term enhances the cluster. The proteins in the population that are elucidated with a specific GO-term are the successes in the population (P_{su}). The proteins in the sample cluster that are elucidated with that specific GO-term are the successes in the sample (S_{su}). The threshold of p-value is set as 0.05.

6. Methods

The proposed model for finding proteins with related biological processes from a PPI network linked with COVID-19 disease consists of four phases. Figure 3 shows the working process for the proposed model.

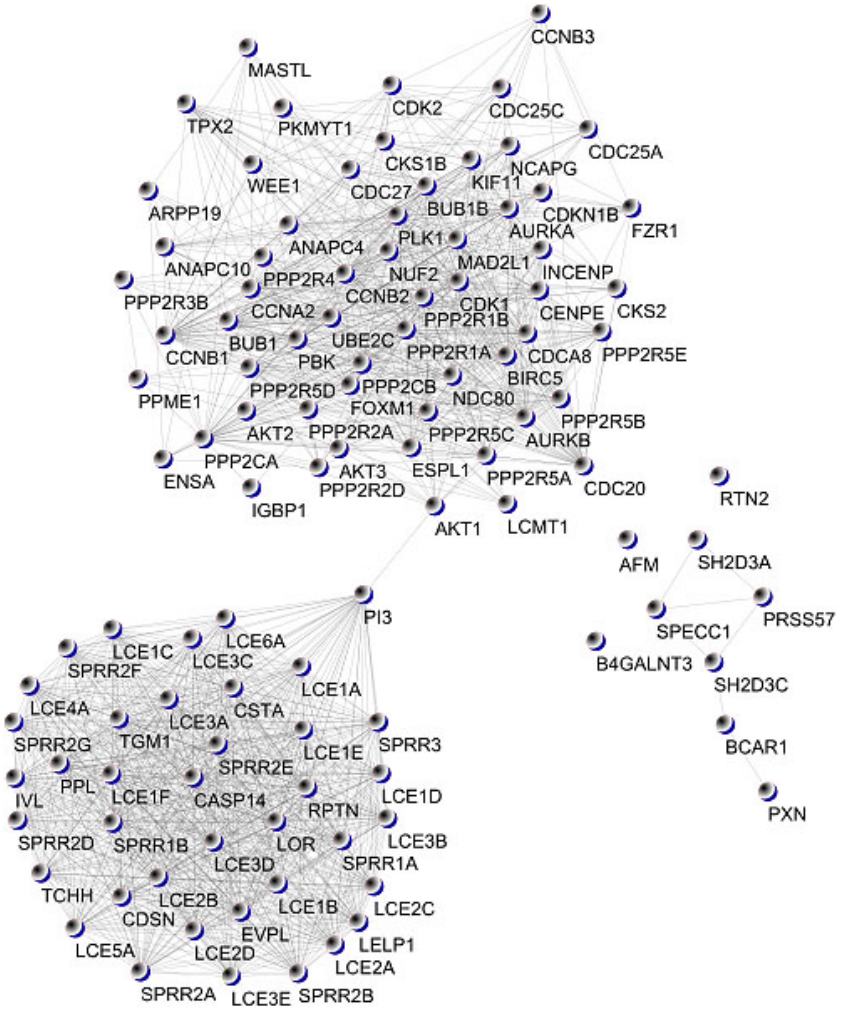

Fig. 4. COVID-19 disease – human PPI graph.

6.1. Data collection and preparation

The STRING is a database of predicted biomolecule interactions [52]. We used the database STRING to collect the PPI data. Also, the authors employed 'Cytoscape' automation to query the STRING database to retrieve network of proteins associated with COVID-19 disease [53].

The implicated COVID-19 disease proteins ("Nsp1", "Nsp2", "Nsp3", "Nsp4", "Nsp5", "Nsp6", "Nsp7", "Nsp8", "Nsp9", "Nsp10", "Nsp11", "Nsp12", "Nsp13", "Nsp14", "Nsp15", "Nsp16", "S", "Orf3a", "Orf3b", "E", "M", "Orf6", "Orf7a", "Orf7b", "Orf8", "N", "Orf9b", "Orf9c", "Orf10") are given from 'R' tool using the package 'RCy3' [9,54]. The RCy3 is an R package to link R language with Cytoscape. We restricted the query with the confidence score between 0.9 and 1.0 to get enough positive protein-protein interactions. The restricted disease protein query returned 57 proteins and 786 predicted interactions. The obtained COVID-19 disease network was then expanded with human host proteins. The authors kept the limit of host human protein interactors as 50. Here, we kept the confidence score as same as the disease protein-protein interaction network. The resultant final network contained 107 proteins and 1446 interactions.

Network visualisation of extensive protein-protein interaction data in a single frame is challenging. First, these networks tend to be large, typically consisting of hundreds of proteins with thousands of interactions between them. Cytoscape presents different layout algorithms. The authors created a force-directed spring embedded layout which uses the Kamada-Kawai algorithm to visualise the graph. The Cytoscape user inter-

	CDK2	FZR1	CDK1	CKS2	CDC20	PPP2R5E	PPP2R2A	PPP2CA	PPP2R3B
CDK2	.	.	.	.	.	.	.	.	.
FZR1	1	.	.	.	1	.	.	.	.
CDK1	1	1	.	1	1	1	1	.	.
CKS2	1	.	.	.	1	.	.	.	.
CDC20	1	.	.	.	.	1	.	.	.
PPP2R5E	.	.	.	.	.	.	.	.	.
PPP2R2A	.	.	.	.	.	.	.	.	.
PPP2CA	.	.	1	.	1	1	1	.	1
PPP2R3B	.	.	.	.	.	.	1	.	.

Fig. 5. The adjacency matrix for created PPI graph.

face allows us to drag nodes to override the automatic layout interactively. Figure 4 shows the constructed PPI graph.

6.2. Data pre-processing

In order to obtain fine clusters, the algorithm specifies some pre-processing steps. We analysed whether there are multiple edges and also removed three isolated proteins. Consequently, it was found that the resultant network to be clustered consists of 104 proteins and 1446 edges. The graph data needs to be hoarded in the framework of the adjacency matrix to create an association or transition probability matrix for the MCL algorithm. This matrix will be the current probability matrix for the calculation of the Markov matrix. Figure 5 shows the part of adjacency matrix representation for the created PPI graph.

6.3. Clustering proteins

The MCL algorithm is established with random walks using Markov chain [19,27]. The Markov chain is explained as, for a graph G, the data hoarded as a matrix M. Let $r > 0$ be a number, the matrix derived after scaling each column of M with power coefficient r is called $\Gamma_r M$, and Γ_r is called the inflation parameter with power coefficient r.

Write $\sum_{rq} M$ for the summation of all the entries in column q of M raised to the power r. Formally, $(\Gamma_r M)$ is defined by the Eq. (8).

$$\Gamma_r\left(M_{pq}\right) = M_{pq}^r \Big/ \sum_{rq}(M) \tag{8}$$

Every column q of a stochastic matrix M matches with the vertex q of the stochastic graph connected with M. The row entry p in column q matches with the probability of bustling from vertex q to vertex p. The clusters are formed when the matrix reaches a steady-state or all the values in a row become the same. The procedure for the algorithm is as follows:

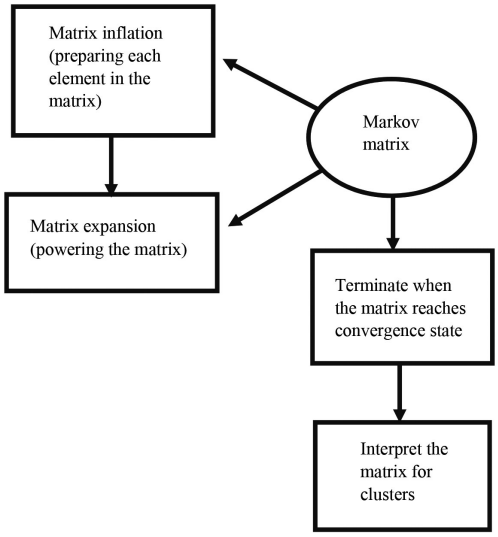

Fig. 6. The workflow of clustering proteins.

Step1: Input: Adjacency matrix of the graph
Step 2: add self-loops
Step3: normalise the matrix to create Markov matrix
Step 4: repeat steps 5 and 6 continuously to attain a convergence state
Step 5: expansion by the e^{th} power of the Markov matrix
Step 6: inflation of resulting matrix with parameter $r > 0$
Step7: interpret the steady-state matrix to discover clusters

Markov Cluster algorithm has been applied to the created COVID-19 disease – human PPI network. The workflow of clustering proteins is shown in the Fig. 6. During the clustering process, the expansion and inflation operations are performed repeatedly until a convergence state occurs. The expansion has been done by powering the matrix. The inflation has been done to prepare each element in the matrix. The overlapped cluster occurs when both clusters are graphically symmetric. It is not happening in this case.

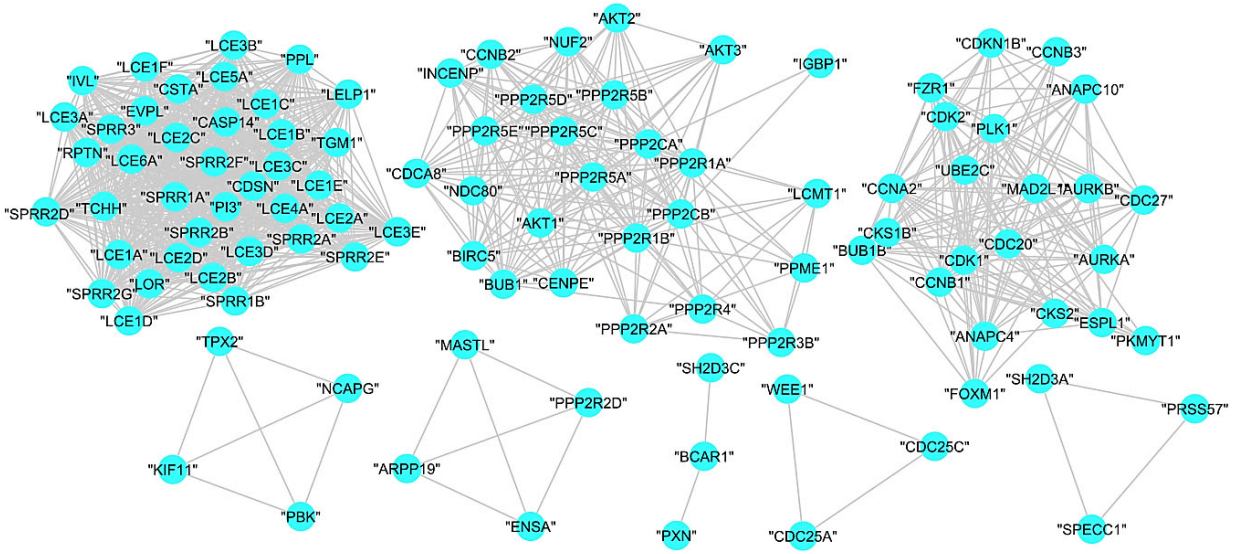

Fig. 7. Subgraphs of COVID-19 disease – human PPI graph.

The MCL returned eight clusters with an inflation parameter 1.8. Figure 7 shows the obtained subgraphs for the clustered proteins.

6.4. Cluster validation

Topological validation of the clusters has been done by scrutinising clustering coefficient. It is also termed as transitivity of the clusters. The value of the clustering coefficient deceit in intervals is 0 and1. A graph or subgraph possesses a superlative structure if its clustering coefficient is closest to 1. The connectivity of the graph is less when its clustering coefficient is 0. The clustering coefficient or transitivity of MCL is 0.82560015, which is close to 1. The clustering coefficients of resulting eight clusters are 1,0.8083624, 0.7964387, 1, 1, 0, 1, 1. This means that the proteins within the clusters are densely connected and have a functional relationship. The clusters with clustering coefficient > 0, are validated biologically by doing the functional enrichment analysis using gene ontology enrichment analysis [51].

7. Results and discussion

The COVID-19 disease – host human protein-protein interaction network with confidence level > 0.9 was downloaded from STRING database. The analysis has been done in RStudio environment on Intel Core-i5 5200U CPU, 64-bit 2.20 GHz processor and 8 GB of RAM. The R tool has over 6000 packages. The authors used igraph, mcl, and RCy3 packages for clustering and visualisation. The topological analysis of the PPI network has been done using igraph and tidyverse packages. The Table 2 shows the simple topological statistics of the created COVID-19 – human PPI network.

Table 2
Topology of COVID-19 – human PPI and protein clusters

Topology	Number of proteins
Number of proteins	107
Number of interactions	1446
Connected components	5
Isolated proteins	3
Average path length	2.591233
Network diameter	5
Density	0.2549815
Cluster coefficient/transitivity	0.8683563
Average node degree	27.02804
Avg. betweenness	70.90654

The structure and magnitude of the created PPI network are turning out to be 107 vertices and 1446 edges. The built PPI network should be further analysed to ensure that the network possesses scale-free property and small-world property. We analysed scale-free property of the constructed COVID-19 disease – human PPI network by power-law fit.

The degree distribution of biological networks approximates a power law: $DD(d) \sim p^{-d}$. This means that the probability or frequency of occurrence of a given degree in any vertex of the constructed PPI graph will be given as "p^{-d}", where p is the parameter characteristic of the PPI network and d is the numeric value of the degree. Figure 8 shows the frequency of degree distribution in the graph.

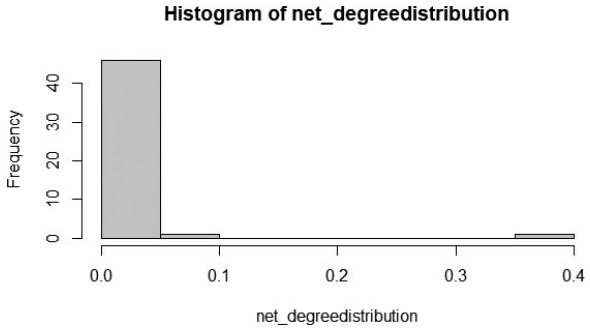

Fig. 8. Degree distribution of the PPI graph.

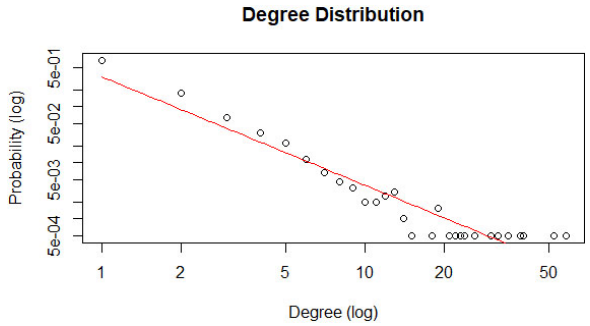

Fig. 9. Power law fit of the PPI graph.

The proportion of variability in the degree distribution is computed on logarithmic value to fit the curve linearly and computed the R-square value. A negative exponential plot for the degree distribution of a network implies that the network follows power-law fit to predict the network's scale-free property. Accordingly, Fig. 9 shows that the constructed PPI network's degree distribution graphically fitted perfectly to a negative exponential plot. Besides, the R-squared value or coefficient of determination is reported as 0.906.

In addition to graphical evaluation, we can adapt statistical evaluation also. The Kolmogorov-Smirnov test is carried mainly to evaluate scale-free networks that can be endorsed. According to this test, the power-law fit, p-value for the degree distribution of the PPI network is 0.9700409. The test result reveals that our constructed PPI network follows the scale-free property of biological networks. This strongly agrees that most vertices have a less degree and a few vertices have a higher degree.

Small-world properties of a network can be analysed with two topological properties: the average clustering coefficient or transitivity and the average path length, and its associated p-values. On the one hand, the transitivity of a vertex will be the fragment of probable edges between connected vertices that are literally represented

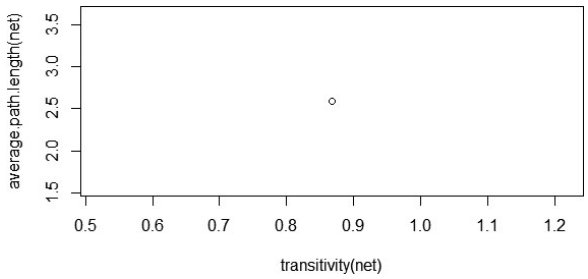

Fig. 10. Transitivity and average path length of the PPI network.

in the network. On the other hand, the transitivity of all vertices can be averaged to get the clustering coefficient of the network. It will finally become the specific topological characteristic for the small world property of the network. High clustering coefficients exhibit the property of a small-world network, whereas small clustering coefficient indents the opposite. Here, the clustering coefficient or transitivity of the PPI network is 0.8683563. Another characteristic of a small-world network is the minimum average path length. Starting from a vertex, the minimum number of jumps or the minimum number of edges passes required to reach another vertex is known as the average path length. It is also known as the shortest path between the vertices. The average path length of our PPI network is 2.591233. Figure 10 shows the transitivity and average path length of our PPI network.

The clustering coefficient and average path length of the network alone does not imply that the constructed PPI network possesses small-world property. We need to calculate the p-values of the parameters, clustering coefficients and average path length. The p-value represents the probability of clustering coefficient and average path length score associated with a random network which is higher than our PPI network. The estimation of p-values can be done by calculating the clustering coefficient and average path length of the random networks and comparing them with our constructed PPI network. Therefore, the p-value can be estimated as the frequency with which the random networks overcome our constructed PPI network's score. The accuracy of p-value depends on how large be the chosen random networks. In this example, the estimation of the p-value can be carried out with Barabasi game function in R, Sum (clustering coefficients $> 0.8683563)/1000 = 0$. The p-value associated with our PPI network's average path can be drawn out from the same number of random networks. The sum (average path lengths $> 2.591233)/1000 = 0$.

Figure 11 shows that p-values associated with both parameters, transitivity and average path length have

Table 3
Results of MCL algorithm over PPI network

Cluster	Number of proteins	Clustering coefficient	Protein name
A	39	1	"LCE1E" "PPL" "LCE1B" "LCE4A" "IVL" "LCE2D" "LCE1F" "LCE2B" "SPRR1A" "LCE2C" "LCE3B" "LCE1A" "SPRR2A" "SPRR2D" "LOR" "LCE3E" "LCE3D" "LCE6A" "LCE1C" "LCE2A" "TCHH" "SPRR2B" "SPRR2E" "SPRR2F" "SPRR3" "CASP14" "CDSN" "LCE3A" "CSTA" "SPRR2G" "LCE3C" "SPRR1B" "LELP1" "RPTN" "EVPL" "LCE5A" "PI3" "LCE1D" "TGM1"
B	26	0.8083624	"PPP2R5E" "PPP2R2A" "PPP2CA" "PPP2R3B" "BUB1" "CDCA8" "CCNB2" "PPP2R5C" "NUF2" "PPP2R4" "PPP2R1A" "CENPE" "INCENP" "AKT1" "PPP2R1B" "BIRC5" "PPP2R5D" "PPME1" "IGBP1" "NDC80" "LCMT1" "PPP2R5A" "PPP2CB" "AKT3" "AKT2" "PPP2R5B"
C	22	0.7964387	"CDK2" "FZR1" "CDK1" "CKS2" "CDC20" "ANAPC10" "CKS1B" "ANAPC4" "AURKB" "ESPL1" "CCNA2" "MAD2L1" "BUB1B" "PLK1" "CDC27" "UBE2C" "CCNB1" "PKMYT1" "FOXM1" "CCNB3" "CDKN1B" "AURKA"
D	4	1	"TPX2" "PBK" "KIF11" "NCAPG"
E	4	1	"MASTL" "PPP2R2D" "ARPP19" "ENSA"
F	3	0	"BCAR1" "SH2D3C" "PXN"
G	3	1	"CDC25C" "CDC25A" "WEE1"
H	3	1	"PRSS57" "SPECC1" "SH2D3A"

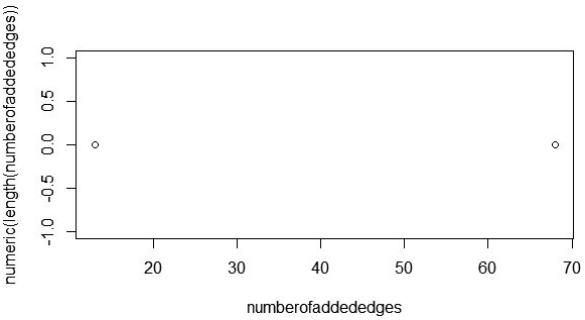

Fig. 11. Graphical representation of a random small-world network.

attained a score of 0 for thousands of random networks. The results of this evaluation strongly implicate that our constructed scale-free PPI network is also a small-world network.

A graph-based clustering technique using the MCL algorithm has been applied to the constructed COVID-19 disease – human PPI network. The MCL algorithm returned eight clusters without overlap. Among them, only seven clusters satisfied graph cluster validity through clustering coefficient between 0 and 1. Table 3 shows the results of MCL clusters.

Biological validation of the resultant clusters has been done using hypergeometric p-value test of GO-term. The gene ontology terms that are most common among the proteins that compose our eight clusters are found out. The test revealed that six clusters are involved in specific biological processes and satisfies the validity of p-value less than 0.05. The GO term finder returned the best feasible biological process associated with cluster (A): keratinization and annotated 38 proteins with p-value 3.50E-71. One protein has multiple

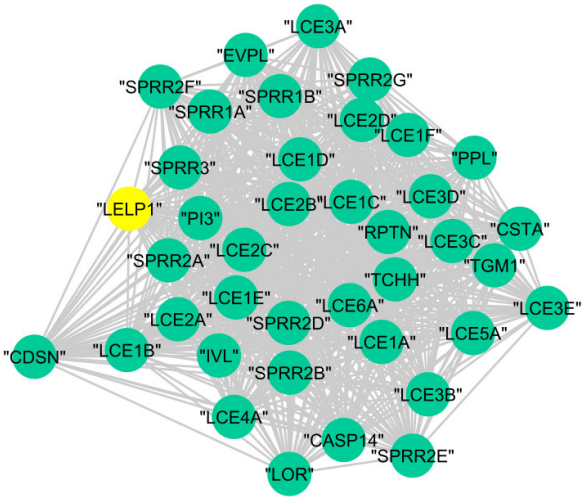

Fig. 12. The graphical representation of protein cluster (A).

mappings. The protein "LOR" has mappings to human protein identifiers "Q9UDR5" and "P23490". Figure 12 shows the graphical representation of biologically investigated protein cluster (A).

In cluster (B), 23 proteins are annotated to the regulation of cellular process with p-value 6.62E-05. There is one unmapped protein "PPP2R4". Figure 13 shows the graphical representation of biologically investigated protein cluster (B).

In cluster (C), the 22 proteins are annotated to the regulation of cell cycle with p-value 1.31E-27. Figure 14 shows the graphical representation of biologically investigated protein cluster (C).

The whole proteins of cluster (D), cluster (E), and cluster (G) are annotated to mitotic cell cycle with p-value 1.66E-06, regulation of phosphoprotein phos-

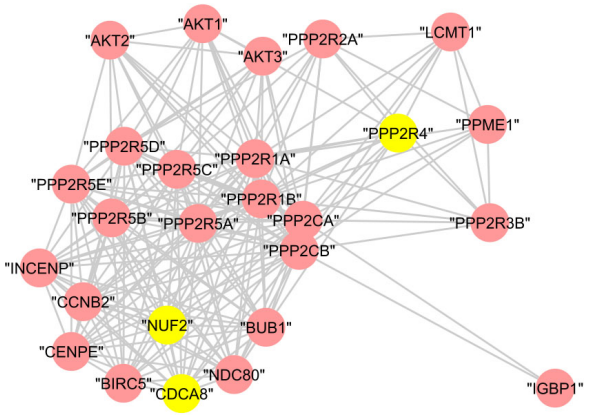

Fig. 13. The graphical representation of protein cluster (B).

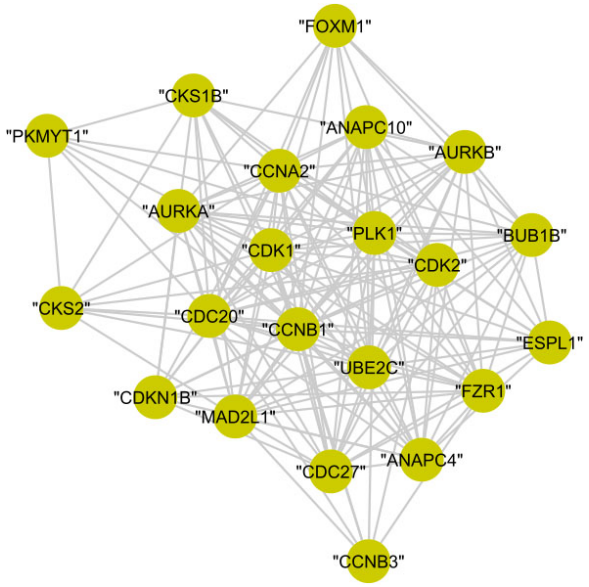

Fig. 14. The graphical representation of protein cluster (C).

phatase activity with p-value 1.15E-09, and G2/M transition of the mitotic cell cycle with p-value 3.03E-07, respectively.

The graphical representation of biologically investigated protein cluster (D), cluster (E), and cluster (G) is shown in Fig. 15. Meanwhile, cluster (F) and cluster (H) has no statistically significant results.

The authors created a COVID-19 disease – human PPI network and analysed the topological characteristics applying the MCL clustering to find protein clusters. Then the resultant clusters were analysed topologically and biologically. We relied on the evaluation measure and the clustering coefficient to validate the clusters topologically.

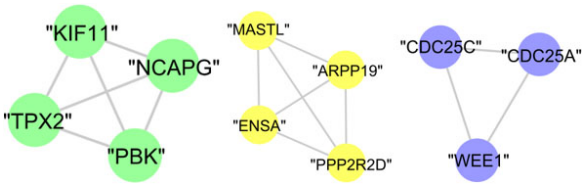

Fig. 15. The graphical representation of protein clusters (D), (E), and (G).

The clustering coefficient of the identified six protein clusters was either 1 or closest to 1. Hence, the sub-graphs possesses a maximal structure. The gene ontology term of proteins in each cluster is associated with a list of genes and results in a fine p-value < 0.05. The majority of proteins in the same cluster possess the same biological process. This implicates that each cluster is a protein complex with some biological significance. Consequently, the generated COVID-19 disease – human PPI network results in six statistically significant protein clusters with high clustering coefficient score. Figure 16 shows the statistics of proteins annotated to the gene ontology term from each cluster verses topologically significant clustered proteins. The number of proteins annotated to the gene ontology term for the biological process in each cluster, and the value of its clustering coefficient is pictured.

The topological analysis of the constructed COVID-19 disease – human PPI network provides evidence of the small world and scale-free properties of biological networks. Flow simulation-based graph clustering algorithm, MCL has been applied to the PPI network for further investigation and returned eight clusters. The clustering coefficient of MCL is equivalent to 0.8256, which implies that the clusters possess a maximal structure. The best feasible biological process associated with cluster (A) is keratinisation involving 38 proteins with p-value 3.50E-71. In total, 23 proteins of a cluster (B) are involved in the regulation of cellular process with p-value 6.62E-05, while the whole 22 proteins in a cluster (C) are involved in the regulation of cell cycle with p-value 1.31E-27. Meanwhile, the whole four proteins in the cluster (D), and cluster (E) are involved in mitotic cell cycle with p-value 1.66E-06, and regulation of phosphoprotein phosphatase activity with p-value 1.15E-09, respectively. The three proteins in the cluster (G) are reported to G2/M transition of the mitotic cell cycle with p-value 3.03E-07. However, cluster (F), and cluster (H) has no known significant biological terms. All the clusters were biologically validated using hypergeometric p-value test. Thus, six probable statistically significant protein clusters were identified.

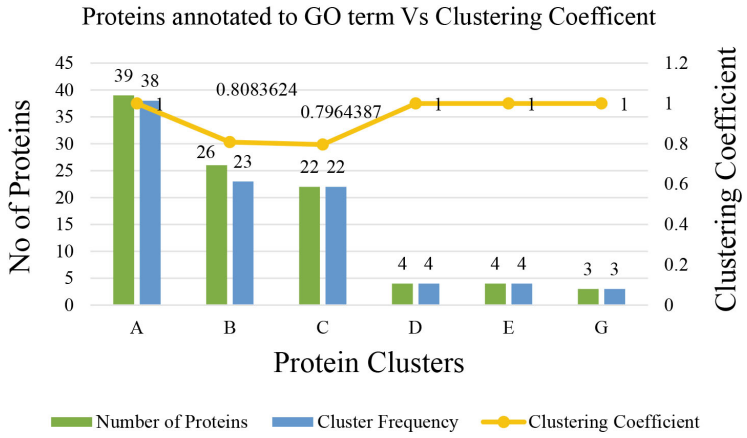

Fig. 16. Statistics of proteins annotated to the gene ontology term.

8. Possibilities, limitations, and future study

The identified protein complexes spectacle knowledge about the biological functions in the disease – human interacting proteins. This gives an advantage to the COVID-19 researchers for therapeutic purposes. Also, the computational method using topological analysis would be a piece of valuable information for the researchers in other fields too. The authors have generated a medium-size PPI network. In this method, the clusters generated depend on two input parameters: expansion and inflation. These varying parameter values generate clusters with different size and structure. For an extensive PPI network, fine-tuning to generate clusters with good structure is a time-consuming process. This model can be enhanced by applying an automatic fine-tuning of the input parameter values according to the data set.

9. Conclusion

PPI network is a core aspect of the knowledge of biological organisms. Moreover, distinct physiological movements inside the human body are responsible for these interactions. The computational method using topological characters of the network helps in analysing the interactions between COVID-19 disease proteins and human proteins. In this paper, a computational analysis of the topological characteristics of the PPI network and graph clustering has been done. The clustering coefficient of the resultant clusters provides us with the information that the protein clusters possess a maximal structure. The presented model revealed that COVID-19 disease – human PPI contains groups of densely connected proteins involved in the same biological processes. Furthermore, these dense sub-graphs have a maximal structure with high clustering coefficients. The model finds valuable information about the biological process of the protein groups, which would be helpful for therapeutic researchers to understand the dynamics of COVID-19 disease. Consequently, it is expected that this study will provide a relevant contribution to the researchers in the field of biomedicine. In this model, the size and structure of resultant clusters depend on the varying tuning parameters, expansion, and inflation. For larger network, this is a time-consuming process. Our future work focuses on designing a more effective model by applying an optimisation algorithm to tune the parameters automatically in order to get the clusters from a large PPI network. Moreover, we will focus on incorporating additional biological information and different algorithms to discover protein complexes.

Conflict of interest

The authors have declared no conflict of interest.

Compliance with ethics requirements

The article does not contain any studies with human or animal subjects.

References

[1] World Health Organization. Novel-coronavirus. 2019 [cited 2020 April 4]. Available from: www.who.int/emergencies/diseases/.

[2] Yang P, Wang X. COVID-19: a new challenge for human beings. Cell Mol Immunol. 2020. doi: 10.1038/s41423-020-0407-x.

[3] Cascella M, Rajnik M, Cuomo A, Dulebohn SC, Di Napoli R. Features, Evaluation and Treatment Coronavirus (COVID-19). In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing. 2020 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK554776/.

[4] Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of COVID-19 disease. medRxiv 2020.03.09.20033357. doi: 10.1101/2020.03.09.20033357.

[5] http://www.ecie.com.ar/images/paginas/COVID-19/4MMWR-Severe_Outcomes_Among_Patients_with_Coronavirus_Disease_2019_COVID-19-United_States_February_12-March_16_2020.pdf.

[6] Baud D, Qi X, Nielsen-Saines K, Musso D, Pomar L, Favre G. Real estimates of mortality following COVID-19 infection. The Lancet Infectious Diseases. 12 March 2020. doi: 10.1016/S1473-3099(20)30195-X.

[7] Seng JJB, Yeam CT, Huang WC, Tan NC, Low LL. Pandemic related Health literacy – A Systematic Review of literature in COVID-19, SARS and MERS pandemics. medRxiv2020.05.07.20094227; doi: 10.1101/2020.05.07.20094227.

[8] Steffen R, Lars C, Sören M. COVID-19. Scenarios of a Superfluous Crisis. SSRN: https://ssrn.com/abstract=3564920 or doi: 10.2139/ssrn.3564920. March 2020.

[9] Gordon DE, Jang GM, Bouhaddou M, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature. 2020; 583: 459–468. doi: 10.1038/s41586-020-2286-9.

[10] https://www.britannica.com/science/coronavirus-virus-group, [cited 2020 April 4].

[11] Zinzalla G, Thurston DE. Targeting protein-protein interactions for therapeutic intervention: A challenge for the future. Future Med Chem. April 2009; 1(1). doi: 10.4155/fmc.09.12. PMID: 21426071.

[12] Klapa MI, Tsafou K, Theodoridis E, Tsakalidis A, Moschonas NK. Reconstruction of the experimentally supported human protein interactome: What can we learn. BMC Syst. Biol. October 2013; 7(96). doi: 10.1186/1752-0509-7-96.

[13] Md. R. Islam, Md. L. Ahmed, Paul BK. Topology Analysis of Protein-protein Interaction Network and Identification of Gene Ontology for Obstructive Sleep Apnea and Associated Diseases Using Bioinformatics Tools. 2019 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON) 28–30 November 2019, Dhaka, Bangladesh.

[14] Liu X, Hong Z, Liu J, Lin Y, Rodríguez-Patón A, Zou Q, Zeng X. Computational methods for identifying the critical nodes in biological networks. Briefings in Bioinformatics. March 2020; 21(2): 486–497. doi: 10.1093/bib/bbz011.

[15] Zhu F, Li F, Ling X, Liu Q, Shen B. Disease associated protein-protein interaction network reconstruction based on comprehensive influence analysis. bioRxiv. 2019. doi: 10.1101/2019.12.18.880997.

[16] Zhou Y, Hou Y, Shen J, et al. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discoy. 2020; 6(14). doi: 10.1038/s41421-020-0153-3.

[17] Chen S, Liao D, Chen C, et al. Construction and analysis of protein-protein interaction network of heroin use disorder. Sci Rep. 2019; 9(4980). doi: 10.1038/s41598-019-41552-z.

[18] Wang J, Liang J, Zheng W, Zhao X, Mu J. Protein complex detection algorithm based on multiple topological characteristics in PPI networks. Information Sciences. 2019; 489: 78–92. ISSN 0020-0255. doi: 10.1016/j.ins.2019.03.015.

[19] Rujirapipat S, McGarry K, Nelson D. Bioinformatic Analysis Using Complex Networks and Clustering Proteins Linked with Alzheimer's Disease. In: Angelov P, Gegov A, Jayne C, Shen Q. (eds) Advances in Computational Intelligence Systems. Advances in Intelligent Systems and Computing, Springer, Cham, Vol. 513, 2017.

[20] Rahiminejad S, Maurya MR, Subramaniam S. Topological and functional comparison ofcommunity detection algorithms inbiological networks. BMC Bioinformatics. 2019; 20(212). doi: 10.1186/s12859-019-2746-0.

[21] Bustamam A, Mujtahidah I, Lestari D. Applications of fruit fly optimisation algorithm for analysing protein-protein interaction through Markov clustering on HIV virus. AIP Conference Proceedings 2023, 020231. 2018. doi: 10.1063/1.5064228 Published Online: 23 October 2018

[22] Zhang Z, Liang Y. Operon prediction model based on markov clustering algorithm. INT. J. BIOAUTOMATION. 2019; 23(1): 105–116. doi: 10.7546/ijba.2019.23.1.105-116.

[23] Lei X, Yang X, Fujita H. Random walk based method to identify essential proteins by integrating network topology and biological characteristics. Knowledge-Based Systems. 2019; 167: 53–67. ISSN 0950-7051. doi: 10.1016/j.knosys.2019.01.012.

[24] Nies HW, Zakaria Z, Mohamad MS, Chan WH, Zaki N, Sinnott RO, Napis S, Chamoso P, Omatu S, Corchado JM. A review of computational methods for clustering genes with similar biological functions. Processes. 2019; 7: 550.

[25] Nascimento M, Carvalho A. A graph clustering algorithm based on a clustering coefficient for weighted graphs. Journal of the Brazilian Computer Society. 2011; 17: 19–29. doi: 10.1007/s13173-010-0027-x.

[26] Yin H, Benson AR, Leskovec J. The Local Closure Coefficient: A New Perspective on Network Clustering. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19). Association for Computing Machinery, New York, NY, USA. 2019; pp. 303–311. doi: 10.1145/3289600.3290991.

[27] Dongen VS. "Graph clustering by flow simulation," PhD thesis: University of Utrecht. 2000.

[28] Pastrello C, Kotlyar M, Jurisica I. Informed Use of Protein-Protein Interaction Data: A Focus on the Integrated Interactions Database (IID). In: Canzar S, Ringeling F. (eds) Protein-Protein Interaction Networks, Methods in Molecular Biology. Humana, New York, NY. Vol. 2074, 2020.

[29] Bajpai AK, Davuluri S, Tiwary K, Narayanan S, Oguru S, Basavaraju K, Dayalan D, Thirumurugan K, Acharya KK. How helpful are the protein-protein interaction databases and which ones. bioRxiv 566372. doi: doi: 10.1101/566372.

[30] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, von Mering C. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research. 08 January 2019; 47(D1): D607–D613. doi: 10.1093/nar/gky1131.

[31] Zahiri J, Emamjomeh A, Bagheri S, Ivazeh A, Mahdevar G, Tehrani HS, Mirzaie M, Fakheri BA, Mohammad-Noori M. Protein complex prediction: A survey. Genomics. January 2020; 112(1): 174–183.

[32] Fang J, Cai C, Chai Y, Zhou J, Huang Y, Gao L, Wang Q, Cheng F. Quantitative and systems pharmacology 4. Network-based analysis of drug pleiotropy on coronary artery disease. European Journal of Medicinal Chemistry. 1 January 2019;

161: 192–204.
[33] Jones KA, Kentala K, Beck MW, An W, Lippert AR, Lewis JC, Dickinson BC. Development of a split esterase for protein-protein interaction-dependent small-molecule activation. ACS Cent. Sci. September 24, 2019; 5(11): 1768–1776. doi: 10.1021/acscentsci.9b00567.
[34] George G, Parambath SV, Lokappa SB, Varkey J. Construction of Parkinson's disease marker-based weighted protein-protein interaction network for prioritisation of co-expressed genes. Gene. 20 May 2019; 697: 67–77.
[35] Marziyeh K, Sadegh S. Identification of the effects of the existing network properties on the performance of current community detection methods. Journal of King Saud University – Computer and Information Sciences. 2020. doi: 10.1016/j.jksuci.2020.04.007.
[36] https://www.ebi.ac.uk/training/online/course/network-analysis-protein-interaction-data-introduction/properties-ppins-scale-free-networks.
[37] Mikaela K, Evangelos K, David P-E, Georgios PA. A guide to conquer the biological network era using graph theory. Frontiers in Bioengineering and Biotechnology. 2020; 8: 34. https://www.frontiersin.org/article/10.3389/fbioe.2020.00034, DOI=10.3389/fbioe.2020.00034, ISSN=2296-4185.
[38] Dubey P, Shingare A, Inamdar V. A force directed layout algorithm for biological networks. International Journal of Computer Applications (0975–8887). June 2015; 120(No. 21).
[39] Fereydoun H, et al. Not all scale-free networks are born equal: The role of the seed graph in PPI network evolution. PLoS Computational Biology. 2007; 3(7): e118. doi: 10.1371/journal.pcbi.0030118.
[40] Thareja R. Data Structures Using C. Second Edition, Oxford University Press, New Delhi. Chapter. 13, 2014, 393–414.
[41] Kamada K, Kawai S. An algorithm for drawing general undirected graphs. Information Processing Letters. April 1989; 31(1): 7–15.
[42] Kusuma WA, et al. Clustering of protein-protein interactions (PPI) and gene ontology molecular function using Markov clustering and fuzzy K partite algorithm. 2019 IOP Conf. Ser.: Earth Environ. Sci. 299 012034.
[43] Falih I, Grozavu N, Kanawati R, Bennani Y. ANCA: Attributed Network Clustering Algorithm. In: Cherifi C, Cherifi H, Karsai M, Musolesi M. (eds) Complex Networks & Their Applications VI. COMPLEX NETWORKS 2017. Studies in Computational Intelligence. Springer, Cham. Vol. 689, 2018.
[44] Celms E, et al. Application of Graph Clustering and Visualisation Methods to Analysis of Biomolecular Data. In: Lupeikiene A, Vasilecas O, Dzemyda G. (eds) Databases and Information Systems. DB&IS 2018. Communications in Computer and Information Science. Springer, Cham. Vol. 838, 2018.
[45] Brémaud P. Random Walks on Graphs. In: Discrete Probability Models and Methods. Probability Theory and Stochastic Modelling, vol 78. Springer, Cham. Vol. 78, 2017.
[46] Azad A, Pavlopoulos GA, Ouzounis CA, Kyrpides NC, Buluç A. HipMCL: A high-performance parallel implementation of the Markov clustering algorithm for large-scale networks. Nucleic Acids Research. 6 April 2018; 46(6): e33. doi: 10.1093/nar/gkx1313.
[47] Hand J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. Bioinformatics. 21(15): 3201–3212. doi: 10.1093/bioinformatics/bti517.
[48] Zaki N, Efimov D, Berengueres J. Protein complex detection using interaction reliability assessment and weighted clustering coefficient. BMC Bioinformatics. 2013; 14: 163. doi: 10.1186/1471-2105-14-163.
[49] Jaya T, et al. Review on graph clustering and subgraph similarity based analysis of neurological disorders. International Journal of Molecular Sciences. 1 Jun. 2016; 17(6): 862. doi: 10.3390/ijms17060862.
[50] Zhang X, Dai D, Ou-Yang L, et al. Detecting overlapping protein complexes based on a generative model with functional and topological properties. BMC Bioinformatics. 2014; 15: 186. doi: 10.1186/1471-2105-15-186.
[51] http://geneontology.org/, [cited 2020 May 15].
[52] https://version-11-0.string-db.org/, [cited 2020 May 15].
[53] https://cytoscape.org/, [cited 2020 May 15].
[54] http://bioconductor.org/packages/release/bioc/html/RCy3.html, [cited 2020 May 15].