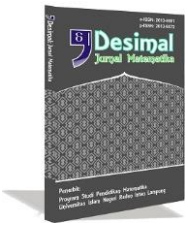
Desimal: Jurnal Matematika Vol 3 No 3 (2020) 191-200

Contents lists available at DJM

DESIMAL: JURNAL MATEMATIKA

p-ISSN: 2613-9073 (print), e-ISSN: 2613-9081 (online), DOI 10.24042/djm

http://ejournal.radenintan.ac.id/index.php/desimal/index



The Implementation of Regularized Markov Clustering with Pigeon Inspired Optimization Algorithm in Analyzing the SARS-CoV-2 (COVID-19) Protein Interaction Network

M. Syamsuddin Wisnubroto*, Marsudi Siburian, Febri Dwi Irawati

Institut Teknologi Sumatera, Indonesia

ARTICLE INFO

Article History
Received : 06-07-2020
Revised : 04-08-2020
Accepted : 15-08-2020
Published : 20-09-2020

Keywords:
Regularized Markov Clustering, Pigeon Inspired Optimization, Protein-Protein Interaction, SARS-CoV-2 (COVID-19)

*Correspondence: E-mail:
syamsuddin.wisnubroto@sd.itera.ac.id

Doi:
10.24042/djm.v3i3.6822

ABSTRACT

Proteins interact with other proteins, DNA, and other molecules, forming large-scale protein interaction networks and for easy analysis, clustering methods are needed. Regularized Markov clustering algorithm is an improvement of MCL where operations on expansion are replaced by new operations that update the flow distributions of each node. But to reduce the weaknesses of the RMCL optimization, Pigeon Inspired Optimization Algorithm (PIO) is used to replace the inflation parameters. The simulation results of IPC SARS-Cov-2 (COVID-19) inflation parameters $((X_i)) = [1,2]$ *get the result of 42 proteins as the center of the cluster and 8 protein pairs interacting with each other. Proteins of COVID-19 that interact with 20 or more proteins are ORF8, NSP13, NSP7, M, N, ORF9C, NSP8, and NSP1. Their interactions might be used as a target for drug research.*

http://ejournal.radenintan.ac.id/index.php/desimal/index

INTRODUCTION

The first reported outbreak of coronavirus or SARS-CoV-2 (COVID-19) was from Wuhan, China, in December 2019. COVID-19 has spread to 206 countries and regions around the world so that the World Health Organization (WHO) declared it as the global health emergency (A.Khailany et al., 2020). This virus is very easy to spread in areas with dense settlements. People who are infected with COVID-19 will experience mild to severe symptoms, such as elevated body temperature along with coughing, sore throat, and headaches. The effects of this disease can get worse if it is accompanied by congenital diseases before the infection. Prevention of the spread of this disease can be done by using masks, washing hands frequently with soap, avoiding public contact (physical distancing), and quarantining those who are positively infected with this disease.

Up to now, the true working drugs or vaccines to reduce the spread of this virus have not been found. COVID-19 has a genome size that varies from 29.8 kb to 29.9 kb (A.Khailany et al., 2020). An overview of the biological process is needed by utilizing protein-protein interaction networks to find out which proteins play an important role in the COVID-19 disease process or to obtain protein target candidates for drug or vaccine development (Bustamam et al., 2018). Grouping methods and an optimization to simplify and speed up the process of grouping can be utilized in the process of collecting information on the results of the interpretation of protein interaction networks (Ginanjar et al., 2016). One of the classification and optimization methods that can be used is Regularized Markov Clustering (RMCL).

RMCL maintains the strength of the Markov Clustering (MCL) algorithm while at the same time, reducing its weakness by replacing the expansion operation in the MCL process with a new operation that updates the flow distribution for each node (V. M. Satuluri, 2012). The manual input inflation parameter by the user is a weakness of RMCL (Lei et al., 2016). To improve the weakness of the RMCL, the Pigeon Inspired Optimization Algorithm (PIO) optimization method can be used by replacing the inflation parameter of the RMCL with the pigeon initialization, which will then be called RMCL-PIO (Regularized Markov Clustering with Pigeon Inspired Optimization Algorithm).

METHOD

The implementation of the RMCL-PIO algorithm on the infection prevention and control (IPC) of SARS-CoV-2 (COVID-19) was carried out using the C++ programming language and was simulated on a computer with the specification of Intel (R) Core (TM) i7-4700 CPU @ 2.40GHz (8 CPUs) specification. ; 8 GB RAM and Windows 10 Pro 64-bit operating system. Graph visualization of the IPC COVID-19 data was performed on the Cytoscape software version 3.8.0.

Markov Clustering was developed by S.V. Dongen (2000) based on a simple paradigm in clustering graphs which is a random walk on a graph G that passes through a dense cluster and tends not to leave the cluster so that many of the nodes in the cluster are traversed. There are three operators in the MCL algorithm, namely Expand, Inflate, and Prune.

The Expand operator increases the flow between nodes and unlocks the potential of the new flow(Van Dongen, 2008), denoted by:

$$Expand(M) = M * M$$

where M = adjacency matrix A.

The Inflate operator regulates the flow of the graph by strengthening the strong flow and weakening the weak flow. However, this operator depends on the parameter r input by the user (Van Dongen, 2008), denoted by:

$$Inflate(M, r) \stackrel{\text{def}}{=} \frac{M(i,j)^r}{\sum_{k=1}^{n} M(k,j)^r}$$

where r = inflation parameter.

Because the value of each element in the definite matrix is less than or equal to 1, this operator might increase the inhomogeneity of each column (as long as$r > 1$) (V. M. Satuluri, 2012).

The Prune operator cuts or deletes elements of each column with a very small value (less than the minimum limit) then each element will be re-equalized so that the number of column values is equal to 1 (Bustamam et al., 2018).

The Regularized Markov Clustering was developed by Satuluri and Parthasaranthy (2009) because the MCL algorithm produces too many clusters (overfit graphs) caused by the clustering process using the flow matrix M from the previous iteration and there is no prevention of columns from the neighboring nodes to separate without being given a penalty, so the regularizing

or smoothing the flow distribution out of the node by paying attention to neighboring nodes was done (V. Satuluri & Parthasarathy, 2009).

The Expand operator is replaced by a new operation which updates the flow distribution of each node as the results of the multiplication of matrix M by the transition matrix M_G of the graph, denoted by:

$$regularize(M) = M * M_G$$

where M_G = transition matrix.

The Pigeon Inspired Optimization Algorithm proposed by Duan HB (2014) is inspired by the flock of pigeons that have been widely used by the military because of their excellent homing behavior. They can find their way home from a great distance using three homing devices, namely magnetic field, sun, and landmark. To idealize some characteristics of a dove, there are two utilized operators, namely a map and a compass which represent the magnetic field and the sun, while the landmark operator model is designed based on a landmark (Duan & Qiao, 2014).

In the map and compass operators, the position and speed of the dove i are denoted by X_i and V_i which will be updated for each iteration of the $d-$dimensional search space. The new position and speed of the dove i of the t-iteration can be obtained by:

$$V_i(t) = V_i(t-1).e^{-RT} + rand.\left(X_g - V_i(t-1)\right)$$

$$X_i(t) = X_i(t-1) + V_i(t)$$

where R = map and compass factor, $rand$ = random value generated from a uniform distribution [0,1], X_g = best position of pigeon which can be calculated by comparing all positions among the whole herd.

In the Landmark operator, the doves are assumed to be far from their destination and it is clear that they are unfamiliar with the landmark. All pigeons are ranked based on the fitness value and half of the pigeons $\left(\frac{N_p}{2}\right)$, denoted by:

$$N_p(t) = \frac{N_p(t-1)}{2}$$

The central pigeons and the kept pigeons will be sought in iteration t where the X_c position is the desired goal, namely:

$$X_c(t) = \frac{\sum X_i(t).fitness(X_i(t))}{N_p \sum fitness(X_i(t))}$$

where $fitness\ (X_i(t))$ is the fitness value of each pigeon in the flock.

The new position of the other pigeons can be calculated by:

$$X_i(t) = X_i(t-1) + rand.\left(X_c(t) - X_i(t-1)\right)$$

where $rand$ = random values generated from a uniform distribution of [0.1]

The minimum optimization can be found through $fitness\left(X_i(t)\right) = \frac{1}{f_{min}(x_i(t)) + \delta}$

where δ = the smallest positive number and f_{min} = minimum fitness of pigeons.

The maximum optimization can be found through $fitness\left(X_i(t)\right) = f_{max}\left(X_i(t)\right)$.

where f_{max} = maximum fitness of pigeons.

This research initialized the pigeon's initial position as the Inflation parameter to obtain the best position of the pigeon((X_i)). Next, the clustering process was carried out using the RMCL algorithm. The RMCL-PIO algorithm process repeats itself until the resulting matrix does not experience a significant level of change compared to the previous iteration, specifically when the global chaos value is less than the threshold = 10^3 (Amrullah & Wisnubroto, 2019)

Table 1. *Pseudocode* Algoritma RMCL-PIO

Input	Matrix adjacency A from .csv file format of the SARS-CoV-2 (COVID-19) IPC data Pigeon initialization/Inflation parameter [1,2], $(\max(X_i)) = [1,2]$
Process	Adding self-loop into the matrix, $A = A + I$ The normalized matrix of Markov M and M_G column of matrix A, where $M = M_G$ The process is repeated until the obtained Global Chaos of matrix M $<$ threshold e The Pigeon Inspired Optimization algorithm process $i = 1$ to N $V_i(t) = V_i(t-1).e^{-RT} + rand.\left(X_g - V_i(t-1)\right)$ $X_i(t) = X_i(t-1) + V_i(t)$ where $t = t + 1$ If $N > 1$ and $N = N/2$ $i = 1$ to N $X_i(t) = \frac{\sum X_i(t).\, fitness(X_i(t))}{N_p \sum fitness(X_i(t))}$ Finding the pigeons' central position $X_i(t) = X_i(t-1) + rand.\left(X_c(t) - X_i(t-1)\right)$ The Regularized Markov Clustering algorithm process Expansion $M := M * M_G$ Inflation $M = (M, X_i(t))$ Prune $M_{ij} = 0$ if $M_{ij} < minval$
Output	Producing M matrix cluster to be visualized using Cytoscape software.

Global chaos states that the rate of change in the resulting matrix value of each previous iteration can be obtained by finding the maximum value of the columns, squaring the columns, calculating the number of columns, and the chaos value (Amrullah & Wisnubroto, 2019). The algorithm of the RMCL-PIO method can be seen in the pseudocode.

RESULTS AND DISCUSSION

The research data utilized the protein-protein interaction network data of COVID-19 which can be accessed online on the BioGRID database | The Biological General Repository for Interaction Datasets via the website https://thebiogrid.org/. The protein interaction network of COVID-19 is stored in a .txt format which contains the Official Symbol Interactor A (protein A) and the Official Symbol Interactor B (protein B) (Rosen, 2012). The data obtained was in the form of interactions between 2 proteins stored in the .csv file format and can be visualized in graphs using Cytoscape software (Figure 1).

The obtained protein-protein interaction network data consisted of 547 proteins and 674 interactions between other coronavirus-related proteins which was saved in .csv file format. This file was used as the value to be input in the RMCL-PIO algorithm process. The proteins that have a relationship are denoted by 1 and if they have no relationship, they are denoted by 0 (Golumbic, 2004).

The matrix data obtained was converted into an adjacency matrix and added a self-loop (Golumbic, 2004), then the normalization was carried out in each adjacency matrix column so that the number of each column is equal to 1 (Rosen, 2012), denoted by:

$$M(i,j) = \frac{A(i,j)}{\sum_{k-1}^{n} A(k,j)}$$

where $A(i,j) =$ adjacency matrix.

Then, the RMCL-PIO algorithm was used to initialize the pigeon/inflation parameters with the ranges of [1,2] $((X_i)) = [1,2]$ and was repeated until the resulting matrix did not change significantly compared to the previous iteration, in other words, the global chaos value should be less than the threshold = 10^3 (Amrullah & Wisnubroto, 2019). Based

on the RMCL-PIO simulation of the SARS-CoV-2 (COVID-19) IPC data, the final result adjacency matrix had been generated and returned to the 2 protein relationships visualized using Cytoscpace Software (Figure 2).

The simulation results showed that there were 50 protein interaction groups, (Figure 2). A total of 8 groups had no central protein because they only consisted of the interaction of two proteins. Then, 6 clusters had two proteins bonded to the centers of clusters. Furthermore, there were 36 clusters that each had different cluster centers and bound to different proteins. The 36 proteins as the center of the cluster indicated that these proteins had a major influence in the formation of SARS-CoV-2 (COVID-19) (Bustamam et al., 2018).

The ORF and ACE2 genes have been widely reported to play a key role in the spread of coronavirus (Kirchdoerfer & Ward, 2019; Koyama et al., 2020; Meer et al., 1998) and the results show that ORF8, ORF9B, ORF9C, and ACE2 proteins are the cluster centers that represent the importance of these proteins in their distribution (A.Khailany et al., 2020). There are also NSP genes such as the NSP13, NSP7, NSP8, and NSP12 proteins that serve as the cluster centers and play a role in triggering the initial coronavirus infection (Wan et al., 2020).

The COVID-19 proteins present in 39 interaction clusters. Out of 50 simulated clusters, 36 of them contained COVID-19 proteins as the cluster center. There are 8 clusters where the central proteins consisted of 20 or more protein interactions, namely the cluster with the central protein of ORF8, NSP13, NSP7, M, N, ORF9C, NSP8, and NSP12. Of the eight COVID-19 proteins, there are two different or new proteins compared to the SARS-CoV, namely the ORF8 and ORF9C proteins (Wu et al., 2020).

The central protein in the largest interaction cluster, ORF8, is one of the 6 accessory proteins encoded in the COVID-19 genome (A.Khailany et al., 2020). Accessory protein is an additional protein that assists other proteins during viral replication or for other non-replicative functions. ORF8 is reported to play a role in the avoidance mechanism of the COVID-19 virus from the immune response by reducing the MHC-I expression by the infected cells (Zhang et al., 2020). This is thought to be done by mediating the post-translation process of MHC-I modification in the Endoplasmic Reticulum until it is degraded in lysosomes (Zhang et al., 2020).

Based on the clustering results, it can be seen that several interacted proteins may play a role in this process, namely OS9, FOXRED2, POFUT1, ERLEC1, ER1LB, SIL1, EDEM3, EMTC1, UGGT2, TOR1A, SDF2, NGLY1, PLD3, PLEKHF2, and NEU1. These proteins play a role in the process of protein modification or the process of protein degradation in the lysosome. ORF9c protein is an accessory protein whose function is unknown. The base sequence of this protein-coding gene overlaps with the ORF9b coding gene which has a role in the defense of the virus from the immune system. In the ORF9c cluster, several proteins have functions related to the immune response, including NLRX1, ABCC1, GHITM, NDFIP2, and F2RL1.

Also, there are several proteins involved in the process of protein degradation through proteasome pathways, energy metabolism in mitochondria, or proteins related to RE and Golgi bodies (Gordon et al., 2020).

Compounds that can inhibit ORF8 or ORF9c protein interactions can be used as drug candidates for COVID-19. This of course still requires another testing stage to see the inhibitory effect of each interaction contained in the interaction cluster.

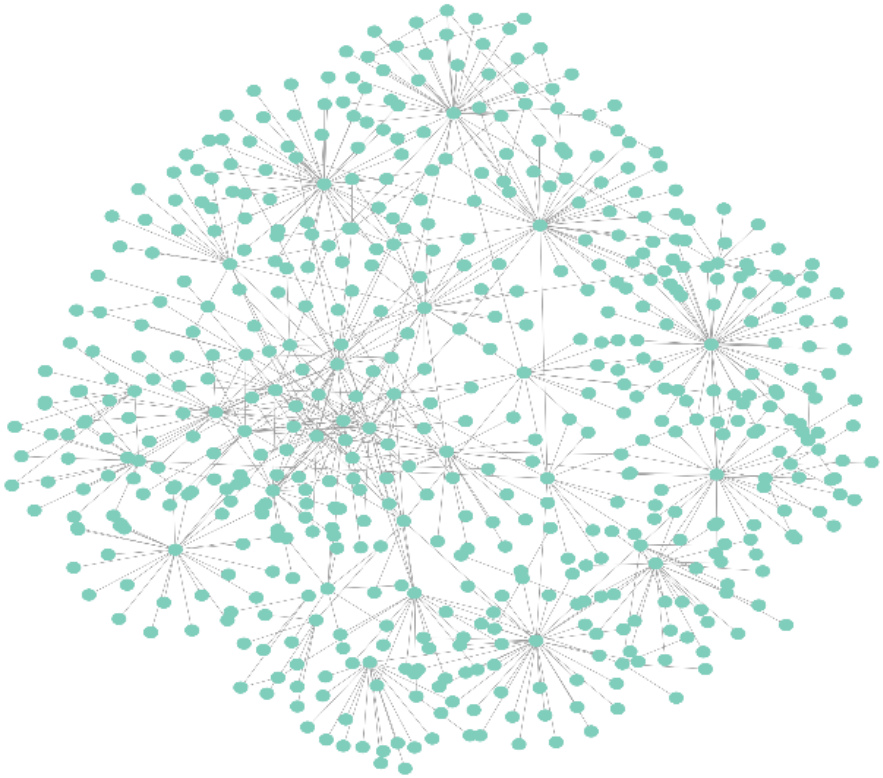
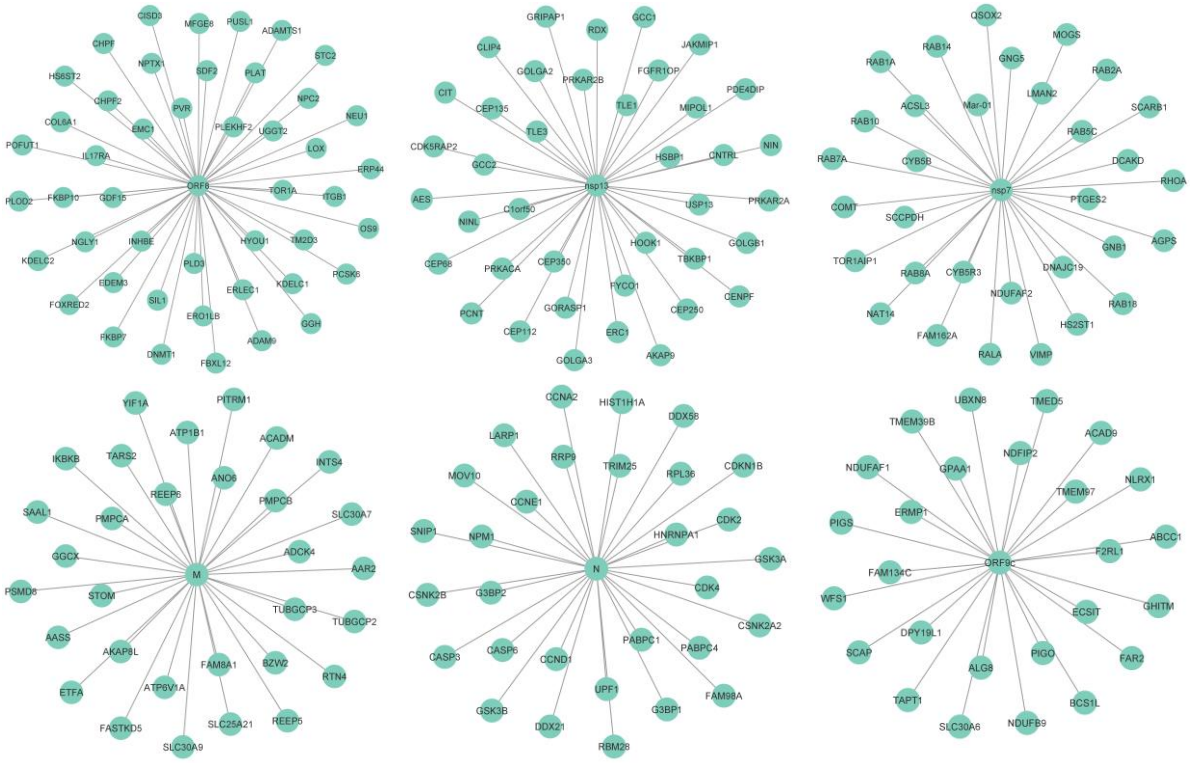
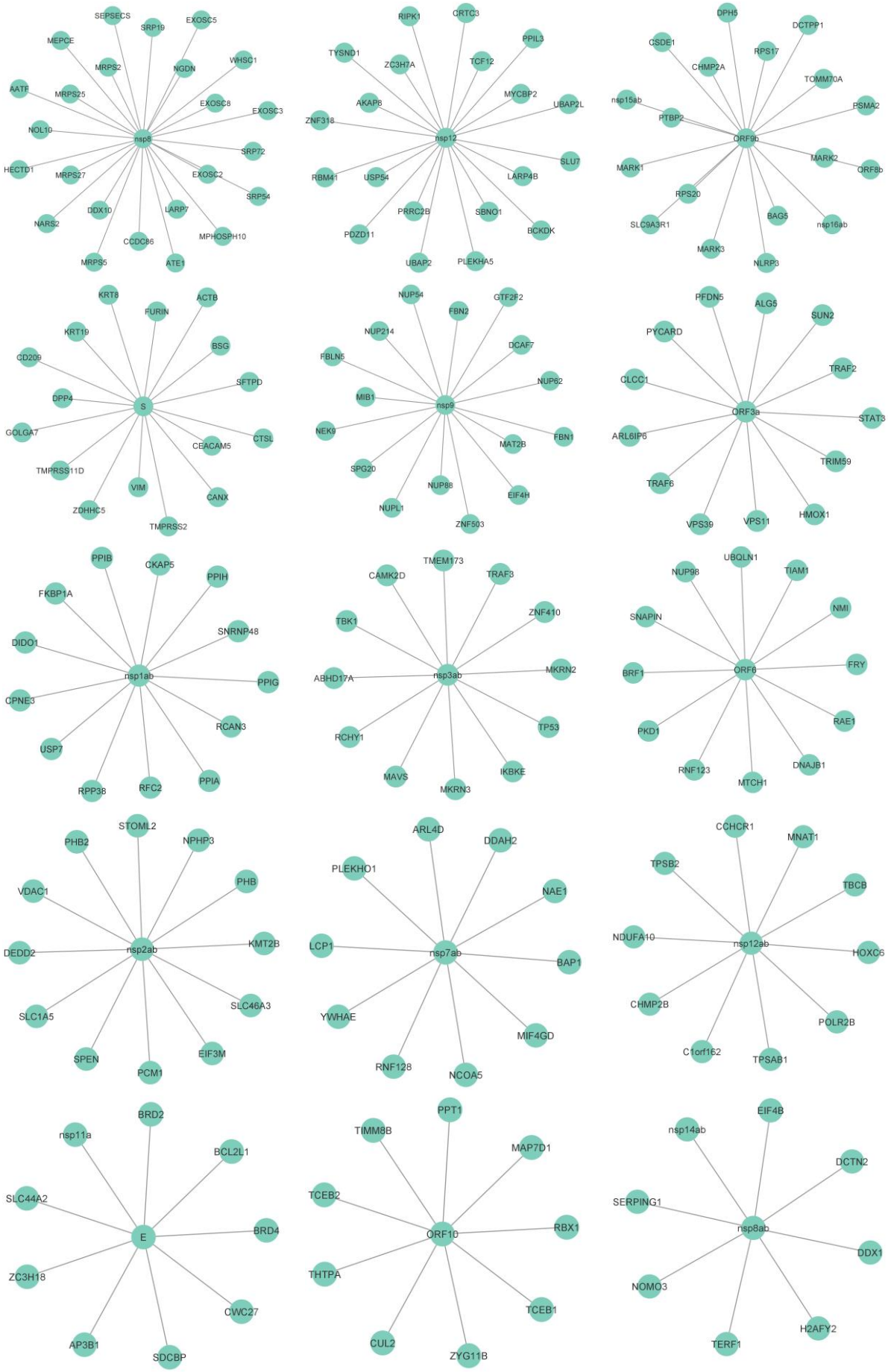
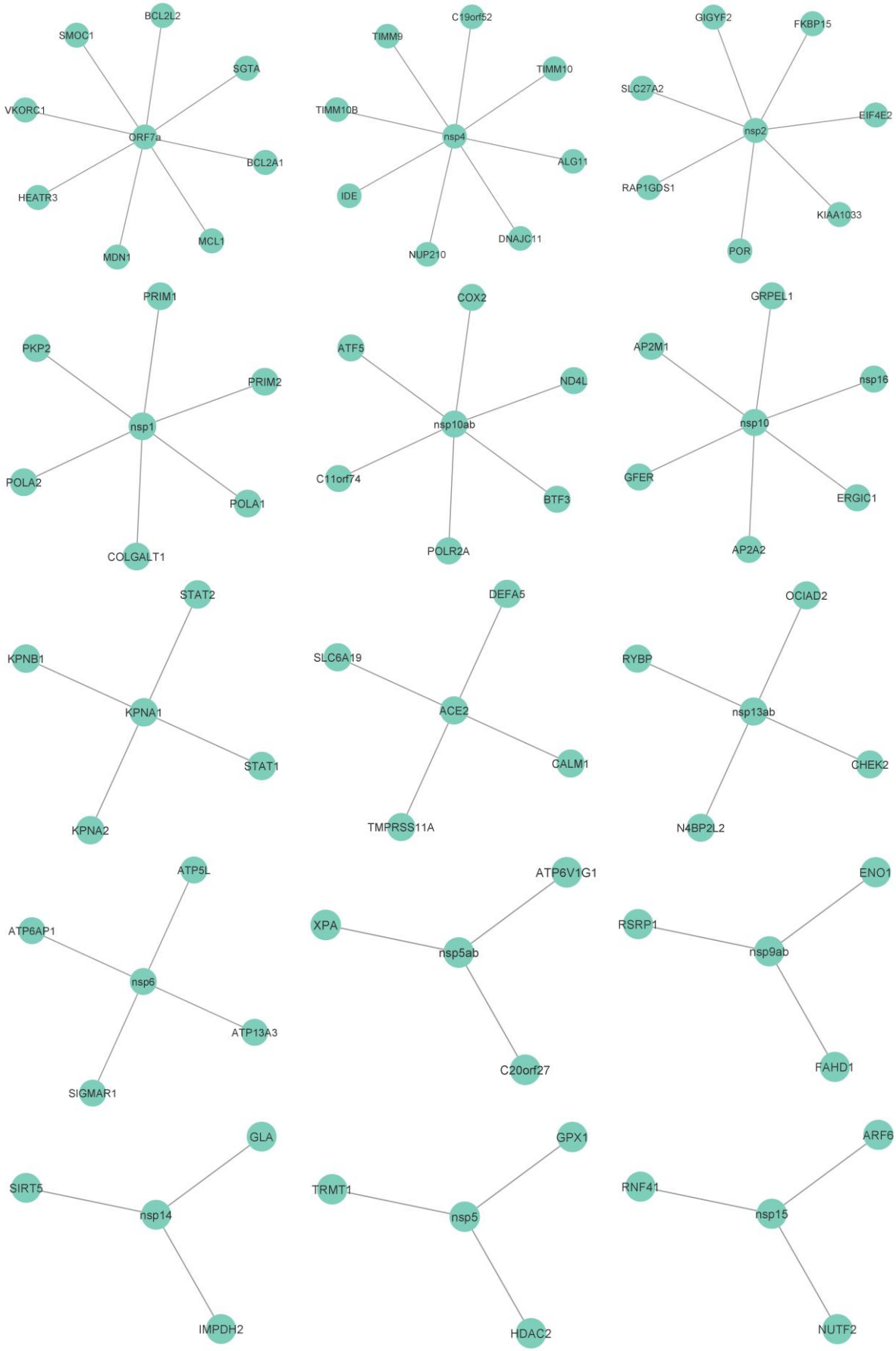


Figure 1. SARS-CoV-2 (COVID-19) IPC







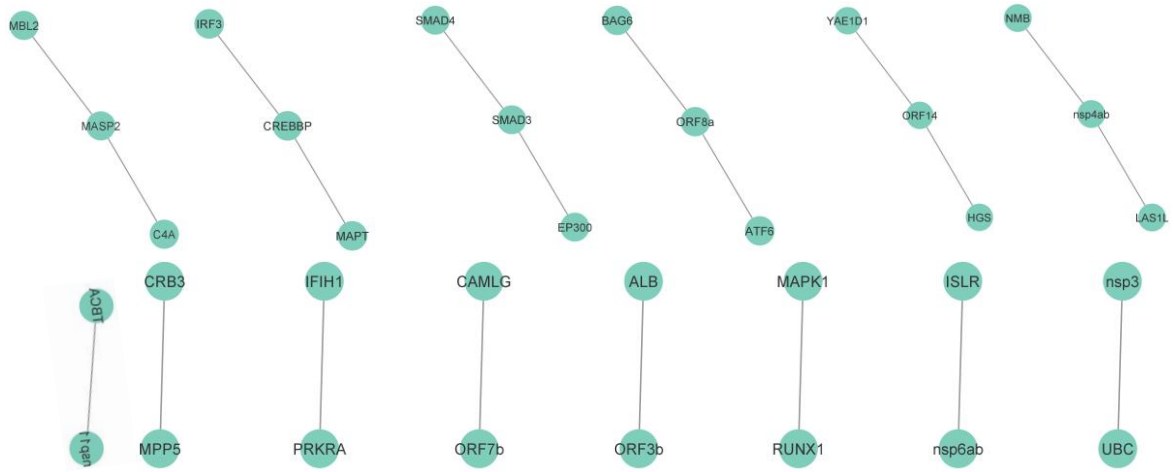


Figure 2. The Clusterization Results of RMCL-PIO

CONCLUSIONS AND SUGGESTIONS

The simulation results of the SARS-CoV-2 (COVID-19) IPC through pigeon initialization $((X_i)) = [1,2]$ obtained the best cluster results after doing 101 iterations which resulting in 42 proteins as cluster centers and 8 interacting protein pairs. The simulation took 1230 seconds to perform 101 iterations. There were 39 clusters of COVID-19 proteins that interacted with human proteins. The interaction data can be used to search for compounds that can inhibit these interactions for the benefit of developing a vaccine or drug for COVID-19.

REFERENCES

A.Khailany, R., Safdar, M., & Ozaslan, M. (2020). Genomic characterization of a novel SARS-CoV-2. *Gene Reports*, *19*, 1–6.

Amrullah, H., & Wisnubroto, S. (2019). Protein clustering in formation of falciparum plasmodium using soft regularized-markov clutering algorithm. *NPrime: Indonesian Journal of Pure and Applied Mathematics*, *1*(2), 87–96.

Bustamam, A., Wisnubroto, M. S., & Lestari, D. (2018). Analysis of protein-protein interaction network using markov clustering with pigeon-inspired optimization algorithm in HIV (human immunodeficiency virus). *AIP Conference Proceedings*, *2023*(1), 020229.

Duan, H., & Qiao, P. (2014). Pigeon-inspired optimization: A new swarm intelligence optimizer for air robot path planning. *Optimizer for Air Robot Path Planning. International Journal of Intelligent Computing and Cybernetics*, *7*(1), 24–37.

Ginanjar, R., Bustamam, A., & Tasman, H. (2016). Implementation of regularized markov clustering algorithm on protein interaction networks of schizophrenia's risk factor candidate genes. *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, *1*(6), 297–302.

Golumbic, M. C. (2004). *Algorithmic graph theory and perfect graphs* (2nd Editio). Elsevier.

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., O'Meara, M. J., Rezelj, V. V., Guo, J. Z., Swaney, D. L., Tummino, T. A., Hüttenhain, R., Kaake, R. M., Richards, A. L., Tutuncuoglu, B., Foussard, H., Batra, J., Haas, K., Modak, M., … Krogan, N. J. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, *583*, 459–468.

Kirchdoerfer, R. N., & Ward, A. B. (2019). Structure of the SARS-CoV nsp12

polymerase bound to nsp7 and nsp8 co-factors. *Nature Communications*, *10*(1), 1–9.

Koyama, T., Platt, D., & Parida, L. (2020). Variant analysis of SARS-CoV-2 genomes. *Bulletin of the World Health Organization*, *98*(7), 495–504.

Lei, X., Wang, F., Wu, F.-X., Zhang, A., & Pedrycz, W. (2016). Protein complex identification through Markov clustering with firefly algorithm on dynamic protein–protein interaction networks. *Information Sciences*, *329*, 303–316.

Meer, Y. van der, Tol, H. van, Locker, J. K., & Snijder, E. J. (1998). ORF1a-encoded replicase subunits are involved in the membrane association of the arterivirus replication complex. *Journal of Virology*, *72*(8), 6689–6698.

Rosen, K. H. (2012). *Discrete mathematics and its applications* (7Th Editio). McGraw-Hill.

Satuluri, V. M. (2012). *Calable clustering of modern networks*. The Ohio State University.

Satuluri, V., & Parthasarathy, S. (2009). Scalable graph clustering using stochastic flows: applications to community discovery. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 737–746.

Van Dongen, S. (2008). Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications*, *30*(1), 121–141.

Wan, Y., Shang, J., Graham, R., Baric, R. S., & Li, F. (2020). Receptor recognition by the novel coronavirus from wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *Journal of Virology*, *94*(7), 1–9.

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., Meng, J., Zhu, Z., Zhang, Z., Wang, J., Sheng, J., Quan, L., Xia, Z., Tan, W., Cheng, G., & Jiang, T. (2020). Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host & Microbe*, *27*, 325–328.

Zhang, Y., Zhang, J., Chen, Y., Luo, B., Yuan, Y., Huang, F., Yang, T., Yu, F., Liu, J., Liu, B., Song, Z., Chen, J., Pan, T., Zhang, X., Li, Y., Li, R., Huang, W., Xiao, F., & Zhang, H. (2020). The ORF8 protein of SARS-CoV-2 mediates immune evasion through potently downregulating MHC-I. *BioRxiv*.