Application of soft regularized markov clustering for analyzing protein-protein interaction in sars-cov-2 and other related coronavirus

S A Pratiwi, A Bustamam* and D Sarwinda

Department of Mathematics, Universitas Indonesia, Kampus UI Depok, Depok 16424, Indonesia

*alhadi@sci.ui.ac.id

Abstract. Covid-19 is a global disease that has already infected people in the various parts of the world with increasing cases each day. So far, there has been around 20 million cases of Covid-19 that have occurred around the world. Furthermore, a lot of research has been conducted to overcome and cure this disease. One of the studies was uses protein-protein interactions (PPI) in Sars-Cov-2 and other coronavirus to analyze the interactions on the virus which can be used to find out more about how this virus interacts with each other. In this study, we used Markov Clustering (MCL) to analyze this virus. There are many variations of Markov Clustering that have been used in various studies, one of the variations that used in this study is Soft Regularized Markov Clustering (SR-MCL). This model is used to ensure that modules on protein interactions do not overlap and can be used for better analysis. The result shows that SR-MCL can be used to determine the cluster from PPI of Sars-Cov-2 and the other related coronavirus.

1. Introduction

Coronavirus (Covid-19) is a global pandemic that has occurred since the end of 2019. This virus belongs to the order nidovirales, family coronaviridae, and subfamily of orthocoronavirinae [1]. Based on investigations on coronavirus, these virus gained the capacity to cross-species from animals to humans and infected humans [2]. This virus affects mammals, including humans, starting at a mild infectious disease, then developing into a global epidemic, similar to the "Severe Acute Respiratory Syndrome" (SARS) in 2003 in China, and also "Middle East Respiratory Syndrome" (MERS) in 2012 in Saudi Arabia and in 2015 in South Korea [3]. In the medical field, SARS is a common terminology used in coronaviruses that infects human [4].

Covid-19 pandemic has impacted various human activities and forced scientific communities to develop a lot of research around Covid-19. Researchers are trying to stop this pandemic so that world activities can return to normal as before. As a result, there are a lot of new research that has been grown and developed recently. Many countries around the world has developed new research and new system to minimize the spread of the disease, and a few countries already success, including South Korea, Taiwan, Australia, and New Zealand to handle this disease [5].

There are various kinds of research that can be carried out on Covid-19, one of them is using protein-protein interaction (PPI). PPI has an important role in collecting structural components of the cell [6]. PPI networks are analyzed to extract groups of proteins that either take part in the same biological process (induction of cell death for example) or performing similar molecular functions [7]. In this study, PPI will be sorted into several groups by clustering using SR-MCL (Soft Regularized
Markov Clustering). SR-MCL is a variation of R-MCL (Regularized Markov Clustering), this algorithm is known to cover some of the disadvantages that occur in using R-MCL. SR-MCL will be applied to form clusters in PPI which can later be used for further analysis purposes. In addition, the data used in this study is PPI of Sars-Cov-2 and other related coronavirus that can be obtained online as Biogrid 3.5.187.

2. Method
   2.1. SR-MCL (Soft Regularized Markov Clustering)
   SR-MCL was developed to overcome several problems that occurs in R-MCL (Regularized Markov Clustering). R-MCL is known to be suitable for solving clustering problems, but usually in R-MCL functional modules at the same node are combined. In addition, in R-MCL, modules with overlaps cannot be identified because of the overlapping clustering algorithms. It is noted that R-MCL cannot identify hierarchical modules [8].
   SR-MCL was created by Shih and Parthasarathy in 2012. SR-MCL was done by rerunning R-MCL while ensuring the resulting groups were not always the same. The SR-MCL algorithm can be seen in Table 1 below.

   Table 1. Soft Regularized Markov Clustering Algorithm

   | Input | Adjency matrix A from graph G, Ratio penalty $\beta$ |
   |-------|-----------------------------------------------------|
   | Step 1 | Add self-loop from graph, $A=A+I$                  |
   | Step 2 | Form the markov matrix $M$ and $M_G$ by normalizing each column of the matrix $A$, where $M=M_G=AD^{-1}$ where $D$ is the diagonal degree matrix of A |
   | Step 3 | Repeat step 4 to step 6                             |
   | Step 4: | $M =$ Regularized ($M$):= $M * M_G$                |
   | Step 5: | $M =$ Inflate ($M$, $r$, $\beta$)                  |
   | Step 6: | $M =$ Prune ($M$)                                  |
   |       | $M_{ij} = 0$ if $M_{ij} < \text{minval}$          |
   |       | Until global chaos from $M < \text{threshold e}$   |
   | Output | Clustering matrix $M$                              |

3. Result and Discussion
Research begins by looking at how proteins interact before applying SR-MCL. Cytoscape 3.6.0 was used in this study to obtain a picture of the interaction of protein on Sars-Cov-2 and other related coronavirus datasets. PPI network can be represented as a graph, with each node represent a protein and an edge connecting two nodes representing the interaction between two proteins [9]. The results of using Cytoscape to visualize interactions can be seen in Figure 1. After knowing the form of interaction on the dataset, the researcher will apply SR-MCL using R.
The algorithm is run on inflation parameter 2 and penalty ratio ($\beta$) = 1.25 [10]. The research was conducted with a different number of iterations, namely at 100, 500, 1000, and 10000 iterations. Figures 2, 3, 4, and 5 show the results of the different iterations, ranging from 100 to 10000 iterations. From these pictures, it can be seen that in each iteration, the results image is different. In Figure 2, it can be seen that there are differences in the image from before the application of SR-MCL. Figure 2 shows that the protein interactions began to spread and form new clusters.
From the iteration results above, it can be seen that, the more iterations, the more visible the exposure of protein interactions. Furthermore, using Cytoscape 3.6.0, the number of k (center of

Figure 3. Figure of cluster at 500 iterations

Figure 4. Figure of cluster at 1000 iterations
cluster) formed from the SR-MCL algorithm can be calculated. As the result, there are a total of 38 nodes of k.

Based on the iteration results using R, although the cluster distribution of the interaction proteins has been seen, the results obtained are still insufficient to say that up to 10000 iterations, the resulting cluster is the optimal cluster. It can be seen that there are still many proteins that are still assembled and are still connected to one another. Overlapping clusters that are formed on SR-MCL shows that the clusters are related with each other and may have similar functions and form functional modules. The application of SR-MCL show the dependence between clusters that allows the formation of functional modules. Clusters will be more visible if the number of iterations is continuously increased to form optimal clusters. But we stopped iteration up to 10000 iterations because of the total time spent doing iterations.

In addition, each iteration has a different total iteration time which can be seen in Table 2 below.

| Table 2: Total Iteration Time of Each Iteration |
|-----------------------------------------------|
| Iteration | Total Iteration Time (sec) |
|-----------|----------------------------|
| 100       | 2789.36                    |
| 500       | 11580.69                   |
| 1000      | 18366.11                   |
| 10000     | 203637.57                  |

4. Conclusion
Based on the results of the SR-MCL application on PPI from Sars-Cov-2 and another related coronavirus, it can be seen that the more the iterations that are used, the more the PPI spread will be seen. The use of SR-MCL is considered to be helpful for clustering a PPI. This is because SR-MCL can identify overlapping modules as in the dataset that the researchers use. However, better results will be obtained if iterations can be enlarged and the computation that used is parallel computation to speed up iterations at very large numbers. In addition, various kinds of optimizations can also be done to improve the optimal results in this clustering technique.

Acknowledgments
This work has been supported in part by PUTI 2020 research grant from University of Indonesia with contract number NKB-778/UN2.RST/HKP.05.00/2020.

References
[1] Carlos W G, Dela Cruz C S, Cao B, Pasnick S, and Jamil S 2020 Novel Wuhan (2019-nCoV) coronavirus American Journal of Respiratory and Critical Care Medicine
[2] Woo PC, Huang Y, Lau S K, and Yuen K Y 2010 Coronavirus genomics and bioinformatics analysis Viruses p 1804-1820
[3] Gralinski L E, and Menachery V D 2020 Return of the coronavirus: 2019-nCoV Viruses, 12(2)
[4] Ramadan N and Shaib H 2019 Middle East respiratory syndrome coronavirus (MERS-CoV): a review Germs p 35-42
[5] Al-Rohaimi A H and Al Otaibi F 2020 Novel SARS-CoV-2 outbreak and COVID19 disease; a systemic review on the global pandemic Genes & Diseases
[6] Qin G, and Gao L 2010 Spectral clustering for detecting protein complexes in protein–protein interaction (PPI) networks Mathematical and Computer Modelling
[7] Parthasarathy S, and Satuluri V 2009 Scalable Graph Clustering Using Stochastic Flows: Applications to Community Discovery p 743
[8] Shih Y K, and Parthasarathy S 2012 Identifying functional modules in interaction network through overlapping Markov clustering *Journal of Bioinformatics* **28** ECCB p i473-i479 (Oxford: Oxford University Press)

[9] Ginanjar R, Bustamam A, Tasman H 2017 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 ICACISIS 2016* art. no. 7872726 p 297-302.

[10] Lei X, Wang F, Wu, F, Zhang A, and Pedrycz W 2016 Protein complex identification through Markov clustering with firefly algorithm on dynamic protein-protein interaction networks *Information Science* p 303-316