Indonesian Journal of Pure and Applied Mathematics
http://journal.uinjkt.ac.id/index.php/inprime
Vol. 1, No. 2 (2019), pp. 87-96
ISSN 2686-5335
doi: 10.15408/inprime.v1i2.12957

Protein Clustering in Formation of Falciparum Plasmodium using Soft Regularized-Markov Clustering Algorithm

Hafizh Amrullah[1] and M. S. Wisnubroto[2]
[1]Program Studi Matematika, Fakultas Sains dan Teknologi
Universitas Islam Negeri Syarif Hidayatullah Jakarta
[2]Program Studi Sains Data, Jurusan Sains
Institut Teknologi Sumatera
Email: syamsuddin.wisnubroto@staff.itera.ac.id

Abstract

Protein has an important role in our life. Every protein interacts with other proteins, DNA, and other molecules. It forms a very large protein interaction networks. We need clustering method to analyze it. Soft Regularized Markov Clustering (SR-MCL) algorithm is one of clustering method to reduce the weakness of Regularized Markov Clustering and Markov Clustering. In this research, SR-MCL will be applied using OpenMP. In every thread, SR-MCL is run using inflation parameter r = 2, 3, and 4. The simulation results show that, based on the fastest execution time and the smallest iteration, the parameter r = 2 produces the best cluster with 40 iterations and execution time is 613 seconds. The cluster centers obtained are 49 clusters with the largest cluster center is the XPO1 protein that interacts with 662 proteins, and 17 protein pairs that interact with each other. Therefore, the XPO1 is a very influential protein in Plasmodium Falciparum.
Keywords: SR-MCL Algorithm, Protein Interaction Network, Plasmodium Falciparum.

Abstrak

Protein memiliki peranan yang sangat penting dalam kehidupan. Setiap protein berinteraksi dengan protein-protein lain, DNA, dan molekul-molekul lainnya, sehingga terbentuklah jaringan interaksi protein yang berukuran sangat besar. Untuk memudahkan dalam menganalisisnya, diperlukan metode clustering. Algoritma Soft Regularized Markov Clustering (SR-MCL) yang merupakan pengembangan metode clustering untuk mengurangi kelemahan dari Regularized Markov Clustering dan Markov Clustering. Pada penelitian ini, SR-MCL akan diterapkan menggunakan OpenMP, yaitu setiap thread menjalankan SR-MCL dengan parameter inflasi r = 2, 3, dan 4. Hasil simulasi menunjukkan bahwa, berdasarkan waktu eksekusi tercepat dan iterasi terkecil, cluster terbaik diperoleh ketika r = 2 yang menghasilkan 40 iterasi dengan waktu eksekusi 613 detik. Pusat cluster adalah protein XPO1 yang berinteraksi dengan 662 protein dan 17 pasangan protein yang saling berinteraksi satu dengan lainnya. Oleh karena itu, protein XPO1 adalah protein yang sangat berpengaruh dalam pembentukan Plasmodium Falciparum.
Kata kunci: Algoritma SR-MCL, Jaringan Interaksi Protein, Plasmodium Falciparum.

1. INTRODUCTION

The development of technology in various branches of biology produces large-scale databases significantly such as sequencing, microarray studies, studies of gene function, genomic structure, and others [1]. In bioinformatics, the database is processed to obtain a picture of the various biological processes and work of sub-systems that exist at the molecular level, cells, tissues, even in an intact organism [2]. The key to this process is a cellular function that does not depend on DNA, RNA, proteins, or single molecules, but on interactions involving various types of molecules such as protein interactions, DNA / RNA proteins, protein metabolites, and other genetic interactions that form tissues complicated one [3]. Therefore, we need methods to simplify the shape of the protein interaction network so that it is easy to interpret that protein interaction network. One of these methods is the Markov Clustering Algorithm (MCL) [1] [4].

The Markov Clustering Algorithm (MCL) was developed by Dongen [5] to solve graph clustering problems and has been widely implemented in the field of bioinformatics such as the protein interaction networks. The advantage of this algorithm is it produces well-balanced nonhierarchical clustering [1]. Besides, the MCL method is more reliable and provides results faster than the other algorithms [6]. The problem in the clustering process is a large number of clusters and singleton clusters. The complex proteins tend to have only 1530 nodes so that it is difficult to interpret protein interactions with large clusters. While a single node on a singleton cluster does not contain enough information to identify the interactions in a network. To achieve more efficient results, many studies and simulations have been developed to improve the MCL algorithm [6]. One of them is the Regularized Markov Clustering Algorithm (R-MCL). This algorithm was developed by Satuluri et al. [3]. To improve the R-MCL boundary, Shih and Parthasarathy [7] were developed a new variation i.e. Soft Regularized Markov Clustering Algorithm (SR-MCL) to produce the overlapping clusters. The SR-MCL generates the overlap clusters by executing R-MCL repeatedly while ensuring that the resulting clusters are not always the same. To produce a different clustering on each iteration, the stochastic flow will be penalized if it flows to the node of the previous iteration.

In this paper, we apply the SR-MCL algorithm using OpenMP, i.e. run the SR-MCL in every thread, to cluster the protein that forms Plasmodium Falciparum with inflation parameter $r = 2,\ 3,$ and 4. We use the protein interaction network data on the Mushroom from theBioGRID | The Biological General Repository for Interaction Datasets in which there are Official Symbol Interactor A and Official Symbol Interactor B. The protein interaction network data were taken from the Official Symbol Interactor A as protein A and the Official Symbol Interactor B as protein B. Furthermore, the data is cleaned by removing the proteins that have the same loop and interaction.

2. METHOD

The SR-MCL implementation on protein interaction network data of Plasmodium falciparum which was found in homosapiens starts with representation data into an undirected graph. This graph is then expressed in an adjacency matrix A of $n \times n$ [8]. The next steps are:

1. Form the protein interaction network data into the adjacency matrix (called A) and add a self-loop.
2. Normalize each column in the adjacency matrix A using the formula

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

so that the sum of each column is equal to 1.

3. Run the SR-MCL algorithm:
   3.1. Determine the inflation parameter (r) and penalty ratio (β). We use $r = 2, 3, 4$ and $\beta = 1.25$.
   3.2. The expansion process on matrix M using the formula

$$Regularized = M_{reg} \stackrel{\text{def}}{=} M * M_G.$$

   3.3. The inflation proses on matrix M using:

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

   This process will strengthen the strong flows and weaken the weaker flows [1].

   3.4. The cutting process is cutting entries that have the *minval* less than or equal to 0.05 [1].
   3.5. Normalize the result of the cutting process by equalizing the columns of the matrix M_{pru}^2.
4. Global chaos is the rate of the value change of the matrix M that is generated at each iteration in SR-MCL compared to the previous iteration [1] with steps:
   4.1. Find the maximum value for every column on matrix M_{pru}^2.
   4.2. Squares the entries in each column in the matrix M_{pru}^2. Sum every column in matrix M_{pru}^2.
   4.3. Calculate the chaos value i.e. calculate the value of the difference between the maximum value per column from the normalization of the cutting process and the sum of columns from the squared of the entries of the normalization of the cutting process.
   4.4. The SR-MCL algorithm is continued until the global chaos is less than the threshold (i.e. 10^3). This process called converge and there were no significant changes in the new cluster. The global chaos also determines the best algorithm for the next iteration, by taking all the chaos values except 0 and taking the minimum value from all existing global chaos.
5. Interpret the cluster: matrix M^2 provides information on how many clusters are formed, the elements of the cluster, the location of the cluster center, and the number of cluster centers in a cluster. The clusters can be analyzed via how many rows that have nonzero elements. The number of clusters determined by counting the number of rows that have nonzero elements that are linearly independence and the diagonal element of the matrix M^2 is the center of the cluster. A cluster can have more than one cluster center if there is more than one row that has the same nonzero elements.

3. RESULT AND DISCUSSION

3.1. Data Description

The protein interaction network data of Plasmodium falciparum was obtained from the BioGRID from the website https://thebiogrid.org/. Plasmodium is a genus belonging to a group of parasitic protozoa. At present more than 200 species of this genus have been identified, of which around 10 species infect humans. The most deadly species is Plasmodium falciparum, which can cause health complications and death in humans. Acute infection by this species if left untreated can be life-

threatening, whereas chronic infection can cause severe anemia. Plasmodium falciparum requires two organisms to undergo its life cycle, i.e. the mosquito vector and the vertebrate host. Extensive studies have been developed on Plasmodium falciparum because these protozoa cause malaria which is a very deadly disease to humans. The life cycle of these protozoa are very complex and changes during transmission. Plasmodium is located in the salivary glands of female anopheles mosquitoes in the form of sporozoites. There are 68 species of Anopheles mosquitoes to transmit malaria. Plasmodium falciparum is the most dangerous type than Plasmodium vivax, Plasmodium malariae and Plasmodium ovale. Therefore, Plasmodium falciparum is one of the most studied.

The protein interaction network of Plasmodium falciparum is stored in the .txt file format. In this file, there are Official Symbol Interactor A as bait (protein A) and Official Symbol Interactor B as hit (protein B). Using Cytoscape software, these networks are represented as a graph (Figure 1).

There are 2402 network interactions with 939 proteins of Plasmodium falciparum in the .csv file format. The example of protein interaction networks can be seen in Figure 2. This .csv file format will be used as input in SR-MCL algorithm, which will then be formed into the adjacency matrix M_G.

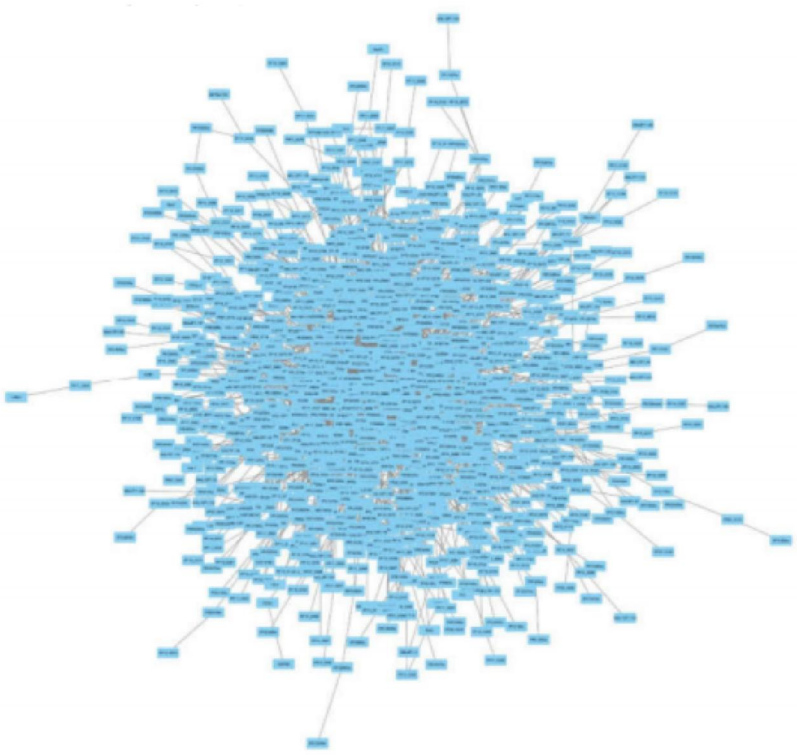

Figure 1. The protein networks interactions of Plasmodium falciparum.

3.2. The SR-MCL Implementation

We use the inflation parameter $r = 2$, 3, and 4 to analyze the characteristics of the cluster formed and find the best r. one characteristic will be analyzed is the number of clusters, the number and location of cluster centers, and the elements in each cluster. Using $r = 2$, we get 49 clusters, with the largest cluster center is the XPO1 protein that interacts with 662 proteins, and 17 protein pairs that interact with each other. The results can be seen in Figure 3. Using $r = 3$, we get 49 clusters and 31

protein pairs interact with each other. The results can be seen in Figure 4. Using $r = 4$, we get 45 clusters and 37 protein pairs interact with each other (Figure 5). Table 1 presents the number of iterations and execution times of the SM-MCL algorithm for $r = 2, 3, 4$.

It can be seen from Table 1, using the same threshold limit i.e. 10^3, the greater r the more iterations so the execution time is longer. The simulation using r = 2 produces the smallest number of clusters than the other simulations (using $r = 3$ and 4). Therefore, the best cluster is formed when we were used r = 2 with 40 iterations and execution time is 613 seconds. The protein which was very influential in the formation of Plasmodium falciparum is XPO1 (Figure 6).

Tabel 1. The number of iterations and execution times of the SM-MCL algorithm for $r = 2, 3, 4$.

r parameter	The number of iteration	Execution time (in a second)
2	40	613
3	80	1165
4	78	1139

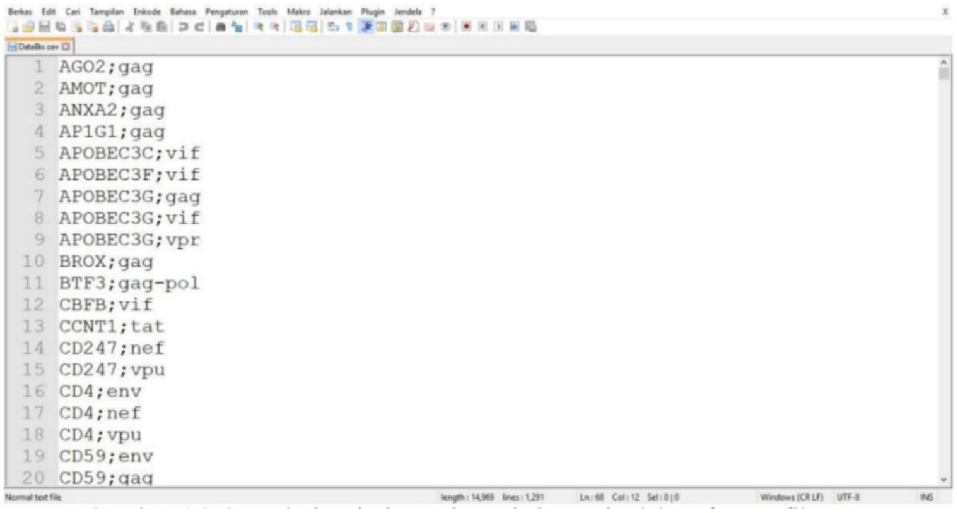

Figure 2. The example of protein interaction networks in .csv format.

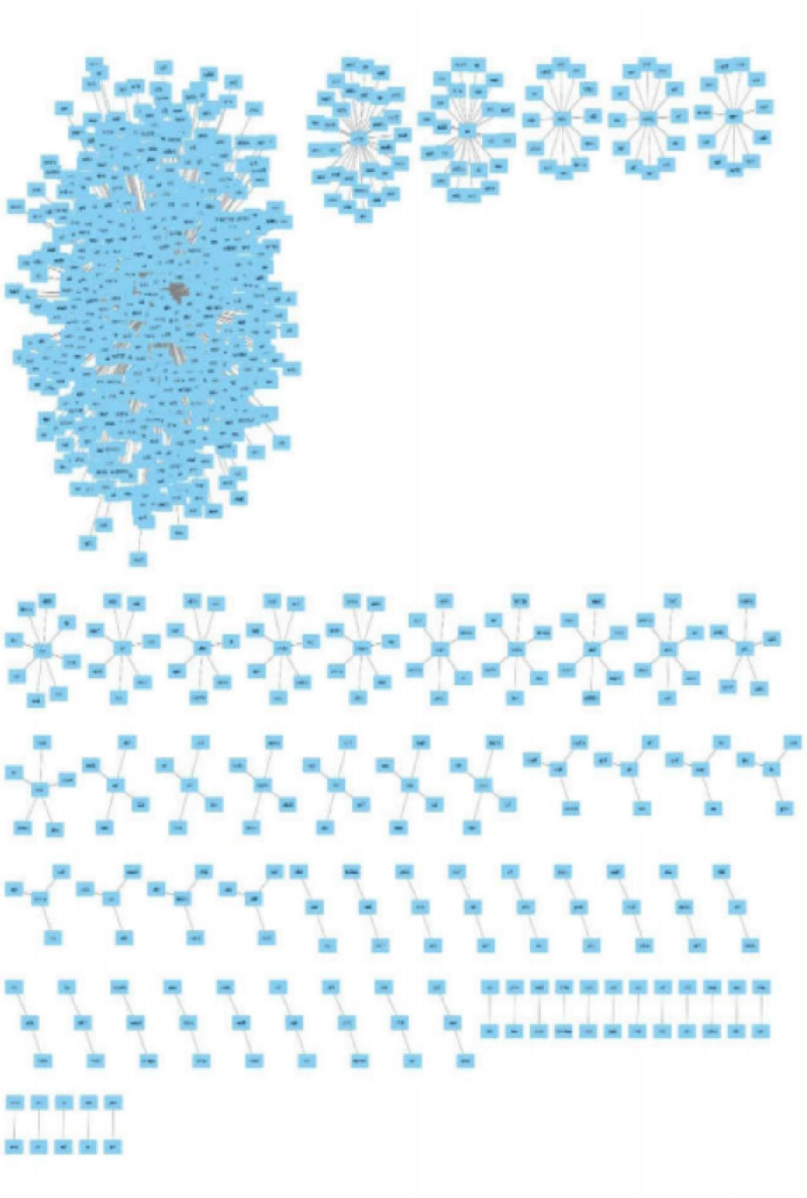

Figure 3. The protein cluster using $r = 2$.

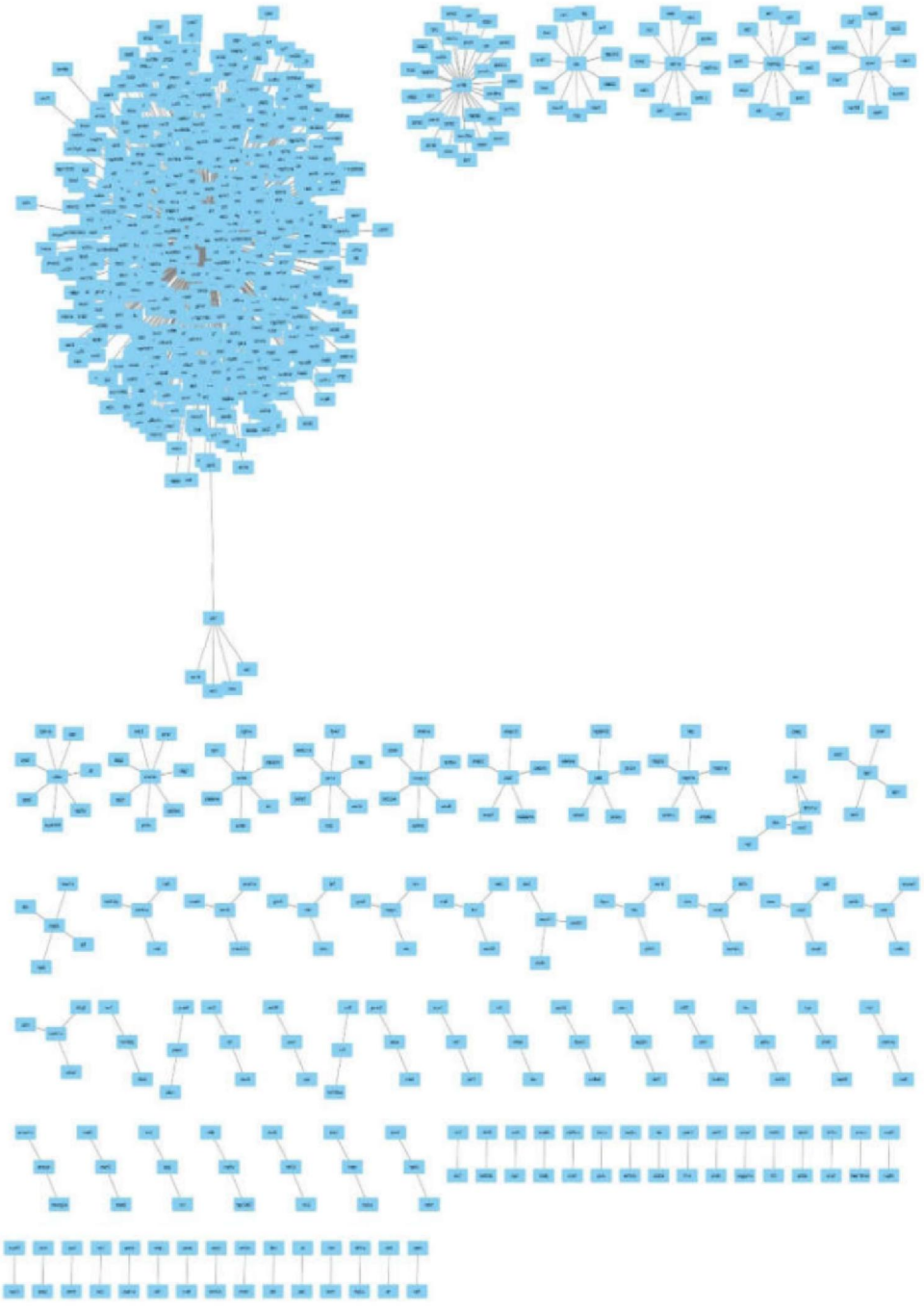

Figure 4. The protein cluster using $r = 3$.

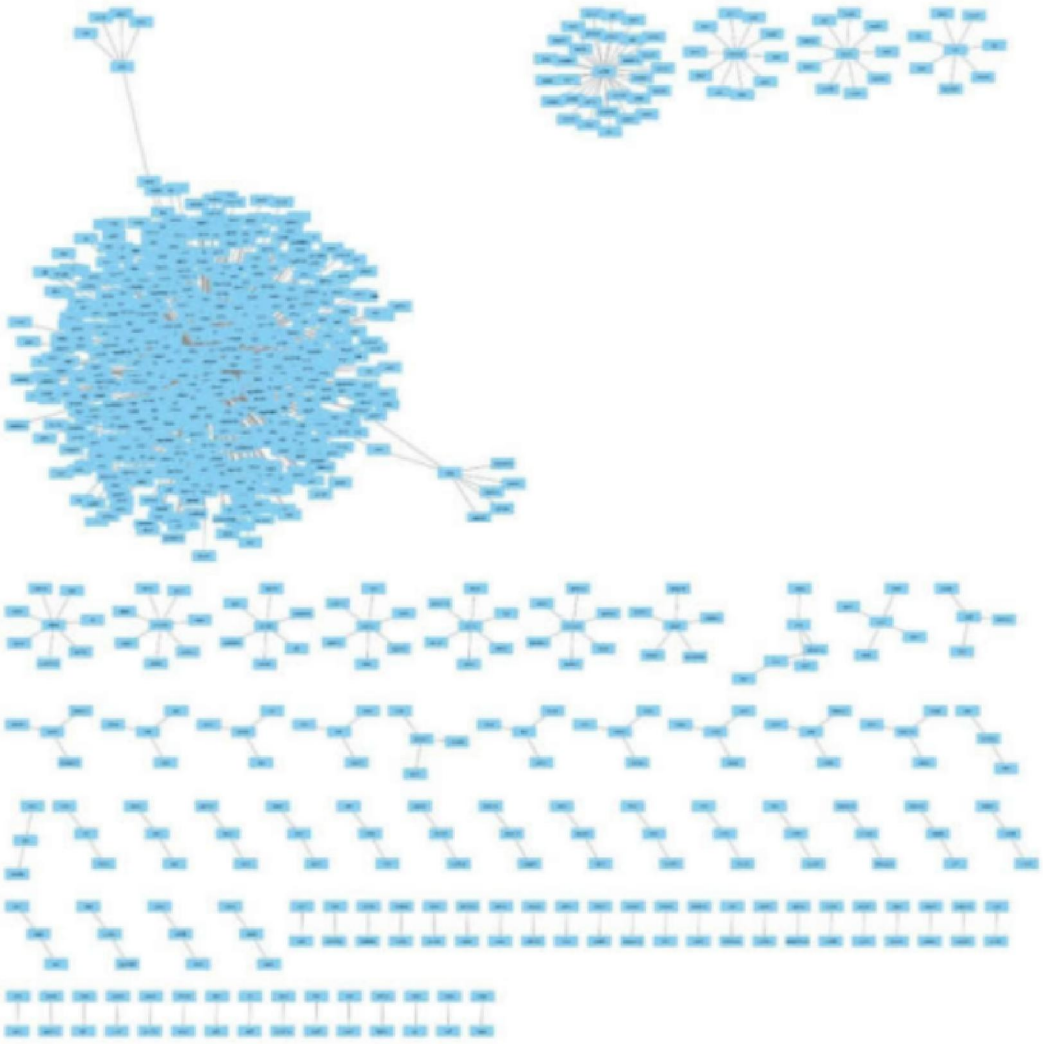

Figure 5. The protein cluster using $r = 4$.

XPO1 (Exportin 1) is a gene that is regulated from cells to encode proteins that mediate leucine-rich nuclear export signals (NES) - independent protein carriers. Proteins specifically inhibit nuclear exports of Rev and UnRNAs. It is involved in controlling several cellular processes by controlling the localization of cyclin B, MPAK, and MAPKAP kinase 2. This protein also regulates NFAT and AP-1. The function of XPO1 is to mediate nuclear exports of cellular protein (cargo) with a signal of leucine-rich nuclear exports (NES) and RNA. In the nucleus, in conjunction with RANBP3, it binds cooperatively with SEN to the target protein and the RAN GTPase in the form of active GTP-bound (Ran-GTP). This complex docking complex nuclear pore (NPC) is mediated through binding to nucleoporin. After the transit of the nuclear export complex to the cytoplasm, the demolition of the Ran-GTP complex and hydrolysis to Ran-GDP (induced by RANBP1 and RANGAP1, respectively) causes the release of cargo from export receptors. The directionality of nuclear exports is thought to be given by the asymmetrical distribution of GTP- and GDP-bound Ran forms between the cytoplasm and the nucleus. Involved in the transportation of U3 snoRNA from the dying C body to the nucleoli.

Binds to the U3 snoRNA final precursor with a TMG lid. Some viruses, including HIV-1, HTLV-1 and influenza A, use it to export unconnected or unconnected RNA out of the nucleus. Interact with, and mediate the export of the Rev-1 and HTLV-1 Rex nuclear proteins. Involved in HTLV-1 Rex multimerization.

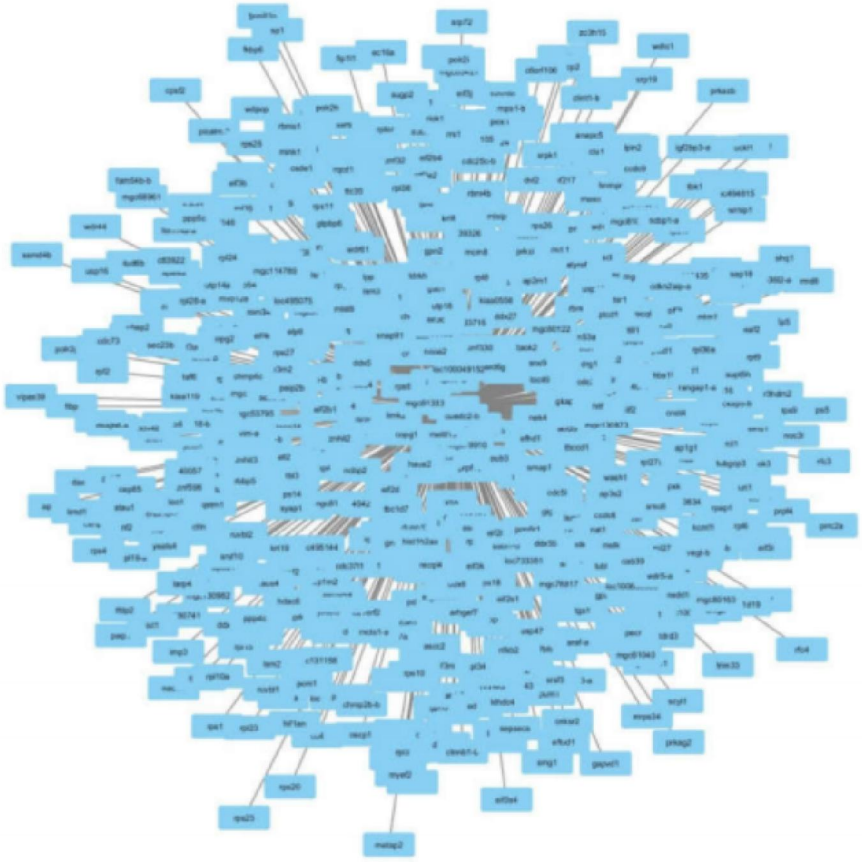

Figure 6. Protein with XPO1 as a center of the cluster.

4. CONCLUSION

To analyze the characteristics of clusters in protein interaction networks using the SR-MCL algorithm, we use the inflation parameters r = 2, 3, and 4. With a threshold 103, the simulation results show that, based on the fastest execution time and the smallest iteration, the parameter r = 2 produces the best cluster is 40 iterations and execution time is 613 seconds. The cluster centers obtained were 49 clusters, with the largest cluster center is the XPO1 protein that interacts with 662 proteins, and 17 protein pairs that interact with each other. From this, it appears that XPO1 is a very influential protein in Plasmodium Falciparum. This protein specifically inhibits nuclear exports of Rev and U snRNAs. It is involved in controlling several cellular processes by controlling the localization of cyclin B, MPAK, and MAPKAP kinase 2. This protein also regulates NFAT and AP-1. Some viruses i.e. HIV-1, HTLV-1 and influenza A use it to export unconnected or unconnected RNA out of the nucleus. Interact with, and mediate the export of the Rev-1 and HTLV-1 Rex nuclear proteins. Involved in Rex's HTLV-1 multimerization.

REFERENCES

[1] A. Bustamam, M. Wisnubroto and D. Lestari, "Analysis of protein-protein interaction network using Markov clustering with pigeon-inspired optimization algorithm in HIV (human immunodeficiency virus)," in *AIP Conference Proceedings*, 2018.

[2] R. Ginanjar, A. Bustamam and H. Tasman, "Implementation of regularized Markov clustering algorithm on protein interaction networks of schizophrenia's risk factor candidate genes," in *International Conference on Advanced Computer Science and Information System (ICACSIS), pp. 297-302*, 2016.

[3] V. Satuluri, S. Parthasarathy and D. Ucar, "Markov clustering of protein interaction networks with improved balance and scalability," in *The 1st ACM International Conference on Bioinformatics and Computational Biology, BCB 2010, pp. 247-256*, New York, 2010.

[4] A. Zhang, Protein interaction networks: computational analysis, Cambridge University Press, 2009.

[5] C. Lin, Y. Cho, W. Hwang, P. Pei and Zhan, "Clustering methods in protein-protein interaction network," *Knowledge Discovery in Bioinformatics: techniques, methods and application,* pp. 1-35, 2007.

[6] S. Dongen, Graph Clustering by Flow Simulation, Ph.D. Thesis: University of Utrecht, 2000.

[7] Y. Shih and S. Parthasarathy, "Identifying functional modules in interaction networks through overlapping Marcov clustering," *Bioinformatics,* vol. 15:28, no. 18, pp. i473-i479, 2012.

[8] V. Satuluri, Scalable clustering of modern networks, Doctoral Dissertation: The Ohio State University, 2013.

[9] K. Rosen, Discrete Mathematics and Its Applications, New York: Mac Graw-Hill, inc, 2012.

[10] S. Varia, Regularized Markov Clustering in MPI and Map Reduce, Doctoral dissertation: The Ohio State University, 2013.