Applications of cuckoo search and ant lion optimization for analyzing protein-protein interaction through regularized Markov clustering on coronavirus

A Rizki, A Bustamam* and D Sarwinda

Deparment of Mathematics, FMIPA Universitas Indonesia, Kampus UI Depok, Depok 16424, Indonesia

*alhadi@sci.ui.ac.id

Abstract. All living viruses have important structures such as protein. Proteins can interact with each other forming large networks of Protein-Protein Interaction (PPI). In order to facilitate the study of these PPI networks, there needs to be clustering analysis of the PPI. In this research, we use PPI network datasets from SARS-CoV-2 and humans. The interactions of the PPI network will then be formed into graphs. Regularized Markov Clustering (RMCL) is used to perform graph clustering. RMCL consists of three main steps which are regularization, inflation, and pruning. The RMCL algorithm is a variant of Markov Clustering (MCL). However, the inflation parameter in RMCL must be inputted manually by the user to obtain the best results. To solve the limitations of RMCL, we developed a new method by combining each Cuckoo Search (CS) and Ant Lion Optimization (ALO) with the original RMCL algorithm. The optimizers are used to optimize the inflation parameter in RMCL. CS and ALO are a part of swarm intelligence which is inspired by the behaviour of cuckoo birds and ant lions in nature. The results show that the interactions formed from CS-RMCL vary from 1401 to 1402. It is more stable than the interactions formed from ALO-RMCL which ranges from 1408 to 3641. The difference between the best elite in each iteration of ALO-RMCL is very influential to the interaction compared to the best nest from the CS-RMCL.

1. Introduction

SARS-CoV-2 is a novel type of Coronavirus that has spread globally and is closely related to the SARS-CoV virus which previously described and identified in the 2002–2003 outbreak [1]. SARS-CoV-2 caused worldwide pandemic with the disease named COVID-19. The structure of SARS-CoV-2 is similar with other viruses which contains capsids in their body. The capsids of viruses consist of protein that is used as a protective coat which makes up an important structure of the virus [2].

A recent study found that a human protein named Angiotensin Converting Enzyme 2 (ACE2) is responsible as a host receptor for human-to-human transmission of COVID-19 because it can strongly bind to SARS-CoV-2 [3]. One of the application of bioinformatics is Protein-Protein Interaction (PPI) network. The interactions of each protein form very large networks of PPI and can be represented in a graph. A big number of large clusters and singleton clusters formed by the process of clustering is a problem occurred in clustering [4]. Therefore, clustering analysis is needed to form a better representation of protein interactions [5].

The MCL algorithm is used as clustering method in many scientific disciplines to solve general problem of graph clustering, including the applications of bioinformatics [6]. Satuluri and Parthasarathy developed Regularized Markov Clustering (RMCL) in 2009 to reduce the weakness of the MCL...
algorithm, but they still maintain the strength of the MCL algorithm [7]. RMCL also has its limitations which is that the inflation parameter must be inputted manually by the researcher. RMCL can be optimized with optimization algorithms which can automatically find the appropriate inflation parameter.

Yang and Deb developed Cuckoo Search (CS) algorithm in 2009 [8]. The parasitic behavior of cuckoo birds during breeding inspired them to develop the CS algorithm that is adapted from Swarm Intelligence [9]. Apart from cuckoo birds, antlions also inspired an optimization algorithm named Ant Lion Optimization (ALO) developed by Mirjalili in 2015 [10]. The ALO algorithm adapts the hunting behaviour of antlion’s larvae and was used in solving many applications. To resolve the weakness of RMCL, we used the CS and ALO algorithms to optimize the inflation parameter in RMCL.

2. Method
2.1. Basic Theory of Graph
Out of the many topics in mathematics, graph theory is one that can simplify a problem as a network. Graph theory is also used in other scientific disciplines, such as solving many problems in computer science. A graph \( G = (V, E) \) consist of a nonempty set \( V = (v_1, v_2, ...) \) called vertices (or nodes) and a nonempty set \( E = (e_1, e_2, ...) \) called edges where each edge has either one or two vertices associated with edge [11].

If the endpoint of an edge in the undirected graph on graph \( G \) are two vertices \( u \) and \( v \), then \( u \) and \( v \) are called adjacent to graph \( G \) [11]. A symmetric matrix \( A = [a_{ij}] \) of \( n \times n \) is the adjacency matrix of graph \( G \) where the entries of \( A \) shows the edge connection between two vertices. The connection between each vertex on matrix \( A \) is arranged as follows:

\[
a_{ij} = \begin{cases} 
1 & \text{if node } v_i \text{ and } v_j \text{ adjacent} \\
0 & \text{otherwise}
\end{cases}
\] (1)

2.2. Basic Theory of Overlapping Clustering
One of the clustering techniques that allow the elements in it to be part of several clusters is called overlapping clustering. Overlapping clustering is better to use on modelling relationship between the element beside partitional clustering [4]. Partitional clustering is clustering technique where data is parted as group without overlap. In biological science, clustering technique is many used to identify functional module network on gene expression data by grouping gene which has same expression to entry at the same group [12].

2.3. Regularized Markov Clustering
Generally, the clustering result of the MCL algorithm produces too many clustering outputs because the MCL algorithm “overfits” the graph [7]. This occurs because there is no penalty when the neighboring nodes’ columns vary extensively [4]. The RMCL algorithm was made to overcome the “overfitting” problem of MCL. The problem of output fragmentation is solved in RMCL by regularizing or smoothing the flow distribution out of the nodes concerning its neighboring nodes [7]. An operator is changed in the RMCL algorithm that is its expansion operator. The regularization operator in RMCL is

\[
\text{Regularize} = M_{\text{reg}} = M \ast Mg
\] (2)

where \( Mg \) is the canonical transition matrix. All of the steps in RMCL except the regularization step are same with the MCL algorithm [7]. The pseudo-code for the RMCL algorithm modified by Ginanjar [4] is given in Table 1.

| Table 1. RMCL algorithm |
|-------------------------|
| **Input:** Adjacency Matrix \( A \) of graph \( G \) |
| **Step 1:** Add self-loop to the graph, \( A = A + 1 \) |
| **Step 2:** Create Markov matrix \( M \) and \( Mg \) by normalizing each column of matrix \( A \), which is \( M = Mg = AD^{-1} \) with \( D \) is diagonal degree matrix of \( A \) |
Step 3: Repeat step 4 to step 6

Step 4: \( M = \text{Regularize}(M) := M \ast Mg \)

Step 5: \( M = \text{Inflate}(M, r) \)

Step 6: \( M = \text{Prune}(M) \)

\[ M_{ij} = 0 \] if \( M_{ij} < \text{minval} \)

Until Global Chaos of matrix \( M < \text{threshold e} \)

Output: Matrix clustering \( M \)

2.4. CS-RMCL

Yang and Deb developed the CS algorithm as an optimization algorithm in 2009 [8]. Cuckoo birds inspired Yang and Deb to create the CS algorithm as part of Swarm Intelligence. Cuckoo birds lay their eggs in other birds’ nest and not in their own which makes them parasitic during their breeding. The adaptability of cuckoo eggs helps last in the foreign nest and keep it from being destroyed by the nest owner because the eggs can change their shape to be similar with the nest owner’s egg. If the cuckoo birds’ eggs hatch after the nest owner eggs do, then the cuckoo birds will spend their food and cause the nest owner’s birds to die [13]. If the cuckoo birds’ eggs hatch before the nest owner’s eggs do, then the cuckoo birds will drop the eggs from the nest [2]. The CS algorithm will find the best nest to optimize the inflation parameter in the RMCL algorithm so that we named this algorithm with CS-RMCL algorithm. Cuckoo birds' behavior to search the best nest will be used to find the inflation parameter automatically. The pseudo-code for the CS-RMCL algorithm adapted from Vira [2] is given in Table 2.

**Table 2. CS-RMCL algorithm**

| Input | Adjacency Matrix \( A \) of graph \( G \) |
|-------|----------------------------------------|
| Step 1 | Add self-loop to the graph, \( A = A + 1 \) |
| Step 2 | Create Markov matrix \( M \) and \( M_g \) by normalizing each column of matrix \( A \), which is \( M = Mg = AD^{-1} \) with \( D \) is diagonal degree matrix of \( A \) |
| Step 3 | Objective function \( f(x), x = (x_1, ..., x_d)^T \) |
|       | Generate an initial population of \( n \) host nests \( x_i, i = 1, 2, ..., n \) |
|       | Count fitness of nest, example \( F_i \) |
|       | Find the best nest |
| Step 4 | Repeat step 5 until step 7 |
| Step 5 | (\( t < \) Max Generation) or (stop criterion) |
|       | Get a Cuckoo’s egg (new solution) at randomly by Levy flights to the best nest in previous generation |
|       | Count fitness of Cuckoo’s egg, example \( F_j \) |
|       | Choose a nest among \( n \) randomly |
| Step 6 | If \( (F_i > F_j) \) |
|       | Replace \( j \) as a new solution to replace the nest \( i \) |
| Step 7 | End if |
|       | A fraction (\( p_a \)) of the worse nest is abandoned and new ones built |
|       | Count fitness of new nest from fraction (\( p_a \)), example \( F_k \) |
|       | Rank the solutions and find the current best |
|       | The best nest on guard to the next generation |
|       | Get the best nest (\( r = \text{best nest} \)) |
| Step 8 | Repeat step 9 until step 11 |
Step 9: $M = \text{Regularize}(M) := M \ast M_g$
Step 10: $M = \text{Inflate}(M, r)$
Step 11: $M = \text{Prune}(M)$

\[ M_{ij} = 0 \text{ if } M_{ij} < \text{minval} \]

Until Global Chaos of matrix $M < \text{threshold e}$

Output: Matrix clustering $M$

### 2.5. ALO-RMCL
Marijalili developed the ALO algorithm as an optimization algorithm in 2015 [10]. The ALO algorithm was inspired by antlion’s larva behaviour when it forages in nature. Antlions have two main phases in their lifecycle, the larva phase and adult phase. Antions hunt to forage in their larva period and reproduce in adult period. The antlion’s larva digs conical holes in the sand. They move around a circular path and dump the sand with their large jaws [14]. The ALO algorithm will find the best elite to optimize the inflation parameter in the RMCL algorithm so that we named this algorithm with ALO-RMCL algorithm. Antlions’ behavior to find the best elite in nature will automatically find the inflation parameter. The pseudo-code for the ALO-RMCL algorithm is given in Table 3.

| Table 3. ALO-RMCL algorithm |
|-------------------------------|
| **Input:** Adjacency Matrix $A$ of graph $G$ |
| **Step 1:** Add self-loop to the graph, $A = A + 1$ |
| **Step 2:** Create Markov matrix $M$ and $M_g$ by normalizing each column of matrix $A$, which is $M = M_g = AD^{-1}$ with $D$ is diagonal degree matrix of $A$ |
| **Step 3:**
- Initialize the clusters as ants and position of antlions
- Calculate the fitness of ants and antlions
- Find the best antlion and assume it as the elite (determine optimum) |
| **Step 4:** Repeat step 5 until step 6 |
| **Step 5:** $(t < \text{Max})$ or (stop criterion)
  For each ant do
  - Select an antlion using Roulette wheel
  - Update minimum and maximum of all variables at iteration
  - Create and normalize random walk
  - Update the position of ant and antlion |
| **Step 6:** End for
  - Calculate the fitness of all ants
  - Replace an antlion with its corresponding ant if it becomes fitter $f(\text{Ant}_i) < f(\text{Antlion}_j)$
  - Update elite if $f(\text{Antlion}_j) < f(\text{elite})$

Get the best elite ($r = \text{best elite}$) |
| **Step 7:** Repeat step 8 until step 10 |
| **Step 8:** $M = \text{Regularize}(M) := M \ast M_g$ |
| **Step 9:** $M = \text{Inflate}(M, r)$ |
| **Step 10:** $M = \text{Prune}(M)$

\[ M_{ij} = 0 \text{ if } M_{ij} < \text{minval} \]

Until Global Chaos of matrix $M < \text{threshold e}$
3. Result and Discussion
This study uses data on the PPI network of SARS-CoV-2 extracted from the online database BioGRID version 3.5.187. There are 1184 proteins and 1437 interactions of SARS-CoV-2 and human protein as shown in Figure 1. Each protein is represented as a node. The purple nodes are human proteins and the yellow nodes are SARS-CoV-2 proteins. The lines connecting each protein represent the interactions between the SARS-CoV-2 and human proteins. Proteins which their interactions form like a flower show that proteins in it have a strong relation. The farther the connections from the proteins are formed, the weaker the interactions are.

![Figure 1. PPI Networks on SARS-CoV-2 and Human](image)

The simulation of CS-RMCL and ALO-RMCL is performed in RStudio version 1.3.1073 by using the R programming language. This simulation program is run on a computer with the following specifications: Intel (R) Core (TM) i7-6700 CPU @ 3.40GHz (8 CPUs), ~3.4GHz; 24 GB RAM and Windows 7 Professional 64-bit operating system.

This simulation uses an inflation parameter that is automatically found from CS and ALO with iterations as many as 100, 500, 1000, and 10000. The different iterations are used to figure out the alterations of the cluster and interactions formed by each iteration from CS-RMCL and ALO-RMCL. In this research, the clustering results from CS-RMCL and ALO-RMCL is 11 clusters in all trials. The different result is significantly shown by the interaction formed on each inflation parameter and iteration.

The results of CS-RMCL simulation with best nest = 9.785 of inflation and iteration = 100 is shown in Figure 2.
Figure 2. Output from CS-RMCL with iteration = 100

The green nodes in Figure 2 are the SARS-CoV-2 and human proteins with SARS-CoV-2 proteins as the center of the cluster and the lines connecting them are the interactions between SARS-CoV-2 and human proteins. The CS-RMCL simulation output has one overlap cluster and ten partition clusters as the output. The overlap cluster formed in the CS-RMCL output shows that the cluster has similar functions and are related to each other. The protein interactions of the CS-RMCL outputs vary from 1401 to 1402 with different best nest values and iterations. The CS-RMCL simulation is shown in Table 4.

| No. | Best Nest (Inflation Parameter) | Number of Iteration | Iteration Time (Second) | Interaction | Cluster |
|-----|---------------------------------|---------------------|-------------------------|-------------|---------|
| 1.  | 9.785                           | 100                 | 706.39                  | 1402        | 11      |
| 2.  | 13.281                          | 500                 | 6195.29                 | 1401        | 11      |
| 3.  | 8.937                           | 1000                | 13397.50                | 1402        | 11      |
| 4.  | 10.100                          | 10000               | 128958.96               | 1402        | 11      |

The result of ALO-RMCL simulation with best elite = 1.185 of inflation and iteration = 100 is shown in Figure 3. The green nodes in Figure 3 are the SARS-CoV-2 and human proteins with SARS-CoV-2 proteins as the center of the cluster and the lines connecting them are the interactions between SARS-CoV-2 and human proteins. The ALO-RMCL simulation has one overlap cluster similar to the outputs of CS-RMCL. The big overlap cluster in Figure 3 shows that the interactions of the cluster will increase where many proteins in the overlap cluster have functional relationship with other proteins in it, so that the additional interaction of ALO-RMCL is formed in its overlap cluster. The protein interactions of ALO-RMCL clustering outputs vary from 1408 to 3641 with different best elite values and iterations. If the inflation parameter is smaller, then the interaction formed is bigger although the number of iterations used is different. The ALO-RMCL simulation is shown in Table 5.
Figure 3. Output from ALO-RMCL with iteration = 100

Table 5. ALO-RMCL Result

| No. | Best Elite (Inflation Parameter) | Number of Iteration | Iteration Time (Second) | Interaction | Cluster |
|-----|---------------------------------|---------------------|-------------------------|-------------|---------|
| 1   | 1.185                           | 100                 | 1124.54                 | 3641        | 11      |
| 2   | 3.389                           | 500                 | 19213.05                | 1615        | 11      |
| 3   | 1.965                           | 1000                | 10036.9                 | 1695        | 11      |
| 4   | 5.062                           | 10000               | 128949.88               | 1408        | 11      |

4. Conclusion
The number of interactions in the CS-RMCL clustering result is almost same although it uses a different best nest value and different number of iterations. The number of interactions of the ALO-RMCL clustering result is bigger when the best elite value is smaller although it uses different number of iterations. All clustering results have one overlap cluster which increases or decreases on each trial. At the same time, the ten partition cluster does not change in each trial with ORF10, nsp2, nsp3, nsp5, nsp10, nsp11, nsp14, nsp15, nsp16, and ORF3b as the center of the cluster.

Acknowledgments
This paper is supported by the PUTI KI 2020 No. NKB-778/UN2.RST/HKP.05.00/2020 grant from DRPM Universitas Indonesia.

References
[1] Zhou Y, Hou Y, Shen J, Huang Y, Martin W, and Cheng F 2020 Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2 Cell discovery 6 1–18
[2] Bustamam A, Nurazmi V Y, and Lestari D 2018 Applications of Cuckoo Search optimization algorithm for analyzing protein-protein interaction through Markov clustering on HIV AIP Conference Proceedings 2023
[3] Uddin M, Mustafa F, Rizvi T A, Loney T, Suwaidi H A, Al-Marzouqi A, Eldin A K, Alsabeeha N, Adrian T E, Stefanini C, Nowotny N, Alsheikh-Ali A, and Senok A C 2020 SARS-CoV-2/COVID-19: Viral Genomics, Epidemiology, Vaccines, and therapeutic Interventions Viruses 12 526
[4] Ginanjar R, Bustamam A, and 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 (ICACISIS) 1 297–302
[5] Permata T S and Bustamam A 2016 Clustering protein-protein interaction network of TP53
tumor suppressor protein using Markov clustering algorithm ICACSIS 2015-2015
International Conference on Advanced Computer Science and Information Systems 221-226

[6] Lestari D, Raharjo D, Bustamam A, Abdillah B, and Widhianto W 2017 Application of clustering methods: Regularized Markov clustering (R-MCL) for analyzing dengue virus similarity AIP Conference Proceedings 1862

[7] Satuluri V M 2012 Scalable clustering of modern network Published Ph.D. Dissertation (Ohio: The Ohio State University)

[8] Yang X S and Deb S 2010 Engineering optimisation by Cuckoo Search International Journal of Mathematical Modelling Numerical Optimisation 1

[9] Zhao J, Lei X, and Wu F 2016 Identifying protein complexes in dynamic protein-protein interaction networks based on Cuckoo Search algorithm IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 1288-1295

[10] Mirjalili S 2015 The ant lion optimizer Advances in Engineering Software pp 80–98

[11] Rosen K H 2012 Discrete mathematics and its applications ed 7 (New York: McGraw-Hill Companies)

[12] Lu H, Hong Y, Street W N, Wang F, and Tong H 2012 Overlapping clustering with sparseness constraints In Data Mining Workshops (ICDMW) IEEE 12th International Conference 486–494

[13] Yang X S 2014 Cuckoo Search and Firefly algorithm (London: Springer)

[14] Majhi S and Biswal S 2018 Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer Karbala International Journal of Modern Science