Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2021, 2021:64
https://doi.org/10.28919/cmbn/5907
ISSN: 2052-2541

ANALYSIS AND PREDICTION OF PROTEIN INTERACTIONS BETWEEN HIV-1 PROTEIN AND HUMAN PROTEIN USING LCM-MBC ALGORITHM COMBINED WITH ASSOCIATION RULE MINING

TITIN SISWANTINING[1,*], ALHADI BUSTAMAM[1], OLIVIA SWASTI[2], HERLEY SHAORI AL-ASH[2]

[1]Department of Mathematics, Universitas Indonesia, Depok, Indonesia

[2]Universitas Indonesia, Depok, Indonesia

Abstract. Human Immunodeficiency Virus (HIV) is a virus that attacks the human immune system. This virus consists of 23 proteins in a single-stranded RNA. The protein interaction between HIV proteins and human proteins can impact to AIDS. The research about HIV-1 proteins and human proteins interactions leads to the insight of drug-target prediction. To analyze protein interactions carried out by the biclustering process. We divide the proteins interactions became a directed graph. The LCM-MBC algorithm is a biclustering algorithm used to analyze protein interactions—this algorithm based on directed graph theory. The results of biclustering used to predict with association rule mining. There is 45 bicluster that has five HIV proteins in one bicluster. From the bicluster obtained, 11 HIV-1 proteins are predicted to interact with 36 human proteins. If human protein interacts with HIV-1 proteins, it means that human proteins will relate according to the interaction type by HIV proteins.

Keywords: biclustering, directed graph; HIV-1; LCM-MBC algorithm; protein interactions.

2010 AMS Subject Classification: 92B05.

1. INTRODUCTION

Human Immunodeficiency Virus attacks the immune system and weakens the body's ability to fight infections and diseases. One type of this virus is the HIV-1 virus. HIV-1 is a virus

*Corresponding author

E-mail address: titin@sci.ui.ac.id

Received April 22, 2021

that causes AIDS (Acquired Immunodeficiency Syndrome). The disease attacks and damages the human immune system and can cause many cases of death worldwide. HIV-1 contains a gene that has a single strand of RNA, which includes 15 proteins, and this is very dependent on human cellular function [1]. Protein is a complex organic compound that is essential and useful in the survival of an organism. One of the tasks of proteins is to channel the substances needed by living things. Protein is a linear polymer molecule such as a chain consisting of tens to thousands of monomer units that hung like beads in a necklace [2]. By using data from the interaction of HIV-1 and human proteins or vice versa, an in-depth analysis can be carried out on developing antiviral drugs and optimizing healing for patients with this virus. Computational predictions on protein interactions can explore more integrated interactions from various sources and provide a complete picture of interactions [3]. Interactions between proteins can be grouped based on their direction into two groups, namely: one-way interactions and two-way interactions. Furthermore, one-way interactions are divide into one-way interactions from HIV-1 proteins to human proteins and one-way interactions from human proteins to HIV-1 proteins [4]. There are two methods for studying protein interactions, namely the experimental method and the computational method. Research using computational methods has been carried out at this time because it is more efficient and effective. Furthermore, the results of the computational method will be confirmed using experimental methods. Research on protein interactions for big data uses the Markov clustering algorithms carried out by [5]. To find out the interaction of proteins, we can group rows and columns simultaneously called bicluster [6]. Some of these methods will be used in this study. [7] found a new method for obtaining bicluster on a dataset called the LCM-MBC algorithm. LCM-MBC algorithm is a graph-based algorithm. This algorithm is a combination of LCM and MICA algorithms. In finding a bicluster, this method looks for the maximum subgraph of a graph, where the maximum subgraph obtained is one of the biclusters. This method is proven to be more efficient and effective in finding a bicluster compared to the LCM and MICA methods. From now on referred to as the LCM-MBC algorithm. Predictions of protein interactions will be made after bicluster obtained from a dataset. In 2019, the LCM-MBC algorithm applied to a dataset of protein interactions between HIV protein and human protein. The LCM-MBC algorithm utilized to analyze the interaction proteins [8]. In

2020, [9] applied three biclustering algorithm to find bicluster on protein interactions. It applies the result of analysis bicluster to find biological function about the protein interactions. In 2014, Mukhopadhyay and Maulik used the Association Rule Mining algorithm to predict future protein interactions. The idea of this algorithm is to find all the possibilities. It will occur and calculate the value of the level of confidence (confidence) of each possibility [10]. Research using this algorithm has been carried out by Mukhopadhyay, Maulik, and Bandyopadhyay using this algorithm in 2012-2014 in predicting protein interactions. [11] and [12] also applied association rule mining to protein interactions. Its involves the rule mining as a dataset classification progress. This research improved the bicluster analysis of protein interactions by the LCM-MBC algorithm. We divide the proteins interactions became a directed graph, i.e. interactions protein one-way, and bidirectional interaction. The Association Rule Mining algorithm implemented on the results of biclustering from the LCM-MBC algorithm. The results of the Rule Mining Association are predictions between HIV-1 protein and human protein.

2. LITERATURE STUDY

2.1. Protein and Its Interactions. Protein is a complex organic compound that is essential and useful in the survival of an organism. Protein is a linear polymer molecule such as a chain consisting of tens to thousands of monomer units that hung like beads in a necklace. This monomer consists of 20 natural amino acids [2]. In carrying out its functions, proteins often interact with other proteins or other nucleotides. In carrying out its functions, proteins often interact with other proteins or other nucleotides. Interactions that occur can be positive or can also lead to negative things. The negative occurrence that means that when a protein interacts with another protein or another nucleotide, applied interaction causes a disease or an abnormality. Today many researchers are interested in learning about protein interactions. By learning about protein interactions, it can expected to know how proteins interact with proteins or other nucleotides. Furthermore, it can be understood pathogenic mechanisms by focusing on concrete intracellular structures or optimizing drugs that have produced by targeting these drugs to new gene products [13].

2.2. Human Immunodeficiency Virus (HIV). HIV-1 contains a gene that has a single strand of RNA, which includes 15 proteins, and this is very dependent on human cellular function [1]. HIV will exploit host cells so that HIV can produce offspring and at the same time, avoid the immune system. These protein-protein interactions between HIV-1 and its host are essential at each stage of the viral life cycle [14]. Based on data taken from NCBI, there are 23 proteins in HIV-1 that interact with humans. According to [15] there are 19 HIV-1 proteins, namely: Envelope surface glycoprotein gp120, Envelope surface glycoprotein gp160 precursor, Envelope transmembrane glycoprotein gp41, gag, capsid, matrix, nucleocapsid, p1, p6 pol, integrase, retropepsin, reverse transcriptase, tat, rev, nef, vif, vpr, and vpu. So based on data on NCBI, the four newly discovered HIV-1 proteins are asp, p2, gag-pol, p51 subunit reverse transcriptase. In general, methods for studying protein interactions divided into two categories, namely the experimental method and the computational method. The experimental method has many shortcomings, such as a long time and very expensive costs, and only biologists can do this research. Therefore, current researchers use computational methods for time and cost-efficiency. The final results of research with computational methods will then be confirmed using experimental methods.

2.3. Graph Theory. A graph $G = G(V,E)$ is defined as a set of pairs (V,E), where $V = v_1, v_2, v_3, ..., v_n$ is a non-empty node set and $E = e_1, e_2, e_3, ..., e_m$ is a set of unsorted pairs of vertices called arcs. The number of vertices in G is denoted by $|V|$ and the number of arcs is denoted by $|E|$ [16].

Two points a and b in a directed graph G said to be adjacent in G if a and b are the endpoints of the arc e on G. Let $G = (VE)$ be a simple graph with $|V| = n$, then the adjacency matrix A of G is a matrix of size $n \times n$, with:

$$\begin{cases} 0; & if\, there\, is\, a\, connected\, edge. \\ 1; & others \end{cases}$$

with $i, j = 1, 2, \ldots, n$. One type of graph is a bipartite graph. Bipartite graph $G(U,V,E)$ is a graph where the node-set is a combination of two-node sets, namely the U node-set and the vertex V set which are mutually separated and each arc connects two vertices from different sets [17]. For example, $G = (U,V,E)$ is a bipartite graph divided into 2 sets of vertices $U =$

$u_1, u_2, u_3, ..., u_n$ and $V = v_1, v_2, v_3, ..., v_m$ so that each edge that connects one U vertex to one vertex V, where edge is the set $E = e_1, e_2, e_3, ..., e_i$. The neighbor of a node in $v \in V$is $N(v) = u \in U|(v,u) \in E$ [14]. Biclique in a graph G is a complete bipartite subgraph of graph G [18]. A biclique is called maximal biclique if it forms a complete bipartite subgraph [19].

2.4. Biclustering. Biclustering is the process of grouping data based on a collection of rows and columns in the data matrix [6, 20, 21]. The result of a bicluster is a group that has the same similarity. These results can be applied to protein-protein interactions to predict unknown protein interactions. In clustering, the grouping is only based on rows or columns, whereas in biclustering, the grouping is done together. The results of this biclustering can be applied to protein-protein interactions to predict unknown protein interactions. To find a maximum bicluster is like finding a maximum biclique (complete bipartite subgraph) of a graph [18].

2.5. Association Rule Mining (ARM). One algorithm used to predict is the Association Rule Mining (ARM) algorithm. This algorithm used by Mukhopadhyay, Maulik, Bandyopadhyay in 2010-2014. It was also used by Bandyopadhyay, Maulik, Holder, and Cook in 2005. In general, there are two steps in this algorithm, namely: find all frequent itemset and form the Association Rule of the frequent itemset [15]. Association rules are if-then statements to help uncover relationships between unrelated data in databases, database relations, or other information. Association rules used to find connections between objects that are often shared. Some applications of Association rules are basket data analysis, classification, cross-marketing, and clustering [22]. Suppose that $I = i_1, i_2, \ldots, i_n$ is a set of n members, X is a subset, where $X \subset I.T = (t_1, X_1), (t_2, X_2), \ldots, (t_m, X_m)$ is a set of m transaction, where t_i and $X_i; i = 1, 2, \ldots, m$ is an identical transaction associated with the member association. The cover of an X member set in T is defined as [4]: $cover(X,T) = t_i|(t_i, X_i) \int X \subset X_i$ Support of X member set in T is defined as: $support(X,T) = |cover(X,T)|$ The frequency of a set of members is defined as: $frequency(X,T) = \frac{support(X,T)}{|T|}$ Support from a set of X members is the number of transactions where all X members appear on each transaction. The frequency of a member set is the opportunity for a transaction to appear in T. A member group is said to be frequent if the support at T is greater than the minimum support. The purpose of an ARM is to find all the rules in the form $X \rightarrow Y, X \cap Y = \emptyset$. The confidence value of this ARM indicates that if each set of member X

appears on a transaction, then the set of member Y also appears on the transaction. The support of a rule defined as the percentage of transactions in T which contains X and Y, can be notated as $P(X \cup Y)$. The confidence values of rules $X \rightarrow Y$ in T defined as a percentage of transactions in T which contains X and also contains Y. This confidence value can consider a conditional opportunity $P(X|Y)$ or can be notated as [23]: $confidence(X \rightarrow Y, T) = \frac{(support(X \cup Y, T))}{support(X,T)}$

2.6. LCM-MBC Algorithm. The LCM-MBC algorithm is a biclustering algorithm for protein interaction datasets, which introduced by [7]. This algorithm can be applied to binary matrices. LCM algorithm and the MICA algorithm are the development of LCM-MBC algorithm. A graph represents LCM-MBC algorithm. Each point (vertex) symbolizes a protein and the side (edge) symbolizes the interaction. If there is no connecting side between the points, it means that there is no interaction between the proteins. The LCM-MBC algorithm combines proteins with the same type of interaction so that no more proteins can be combined. It means that the proteins are in one group (one bicluster) [7]. The LCM-MBC algorithm criteria stop when:

- The number of HIV proteins incorporated is higher than human proteins that not included in one group.
- No more HIV proteins can be combined (only proteins that have no interaction) [7].

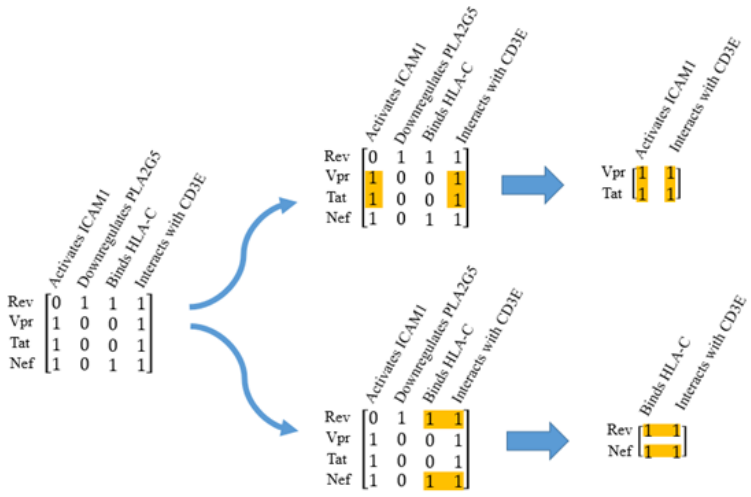

FIGURE 1. Example of LCM-MBC Algorithm.

Figure 1 is the example of LCM-MBC algorithm. There is a matrix of protein interactions. This interaction matrix has four types of HIV-1 proteins that interact with three types of human proteins. If the entry value of 1 then this means, there is an interaction between HIV-1 protein with human proteins and 0 others. The first step to get the bicluster on the interaction protein matrix above is to combine HIV proteins with HIV proteins. If we want to combine is the Vpr and Tat proteins. The next step is to look at the column in that row whose all row entries are 1. hence, the row and column containing the one entry are called a bicluster.

3. DATA

The research data was taken from the National Center for Biotechnology Information (NCBI) website. This data was developed by [24, 25, 26]. The data used are data that contains the name of HIV-1 protein, the name of human protein, and the type of interaction between HIV-1 protein and human protein. There are 23 types of HIV proteins, 3797 types of human proteins, and 130 types of interactions between human proteins and HIV proteins. Table 1 is the dataset algorithm which will be explained later.

4. METHODOLOGY

4.1. Preprocessing. After selecting the data in this research, a dataset is formed based on the type of interaction. This type of interaction divided into three classes. The first class is a class with the type of interaction from HIV-1 protein to the human protein. The second class is a class with the type of interaction from human protein to HIV-1 protein. The third class is a class with a two-way type of interaction between HIV-1 protein and human protein. The type of interaction contained in the first class is the type of interaction that has an active verb. This active verb states the type of interaction from HIV protein to the human protein. There are 60 types of first-class interactions, 47 of which are types of interactions found in the 2014-2018 time span. This type of second class interaction is a type of direct interaction from human protein to HIV protein. Passive verbs characterize this type of interaction type. There are 54 types of second-class interactions, 29 of which discovered in 2014-2018. This third class of interaction type is a two-way interaction between HIV protein and human protein. The majority of this type of interaction uses the word ”with”. There are 14 types of bilateral interactions, with 4 of them

TABLE 1. Dataset Algorithm

Input	**Bipartite Graphs of HIV-1 Protein Data, Human Protein, and Types of Interactions**
1:	Grouping interactions into three classes.
2:	Create a adjacency matrix $M_(n \times m)$ which has the entry 0 with n HIV protein vertex, m human protein vertex, and edge as the interaction.
3:	Fill the entry matrix $M_(n \times m)$ according to the type of interaction: if the interaction is on class-1, thus the entry is "1", if the interaction is on class -2, thus the entry is "-1", if the interaction is on class -3, thus the entry is "X".
4:	Divide the matrix $M_(n \times m)$ into two sub-matrix: positive protein interaction dataset (entry of the matrix is "1" and "X"). Change all the entry "X" into "1", negative protein interaction dataset (entry of the matrix is "-1" and "X"). Change all the entry "X" into "-1".
Outputs	**Positive Protein Interaction Datasets and Negative Protein Interaction Datasets.**

being new types of interactions found in 2014-2018. These three classes are then used to form a dataset with entries in the form of 1, -1, 0, and "X". Entry 1 states there is a direct interaction of the HIV-1 protein to the human protein. Entry -1 states there is a direct interaction from human protein to HIV-1 protein. The "X" entry indicates a two-way interaction between the HIV-1 protein and human protein. Entry 0 states there is no interaction between the HIV-1 protein and human protein. The formed dataset will be divided into two datasets: positive dataset and negative dataset. There are 23 columns and 7321 rows in the positive dataset and 23 columns 4478 rows in the negative dataset. The columns in the positive dataset and the negative dataset state the name of the HIV-1 protein, while the row states the type of interaction with the human protein.

4.2. LCM-MBC Biclustering Methods. The LCM-MBC algorithm used to find bicluster from a protein interaction dataset. Based on the research carried out, bicluster obtained from the following dataset (Table 2 and Table 3):

TABLE 2. The number of bicluster from positive dataset

The amount of HIV protein in 1 bicluster	The number of bicluster formed
2	454
3	313
4	139
5	31
6	2

TABLE 3. The Number of Bicluster from Negative Dataset

The amount of HIV protein in 1 bicluster	The number of bicluster formed
2	398
3	259
4	72
5	14
6	1

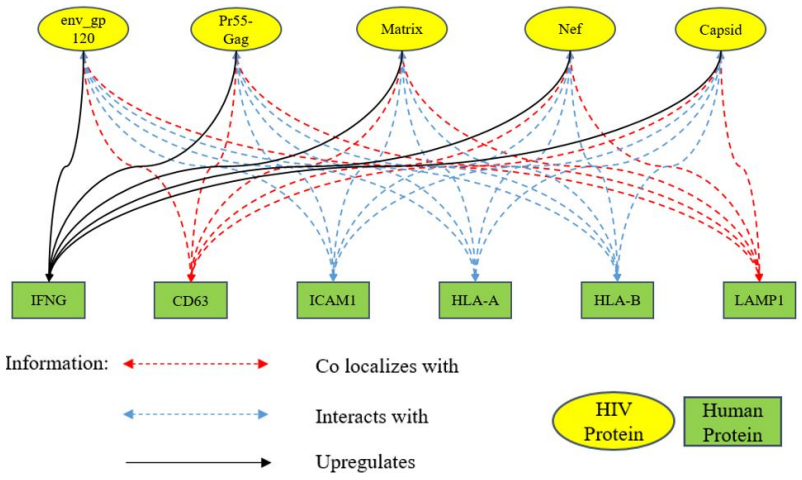

FIGURE 2. One of the Bicluster from Positive Dataset.

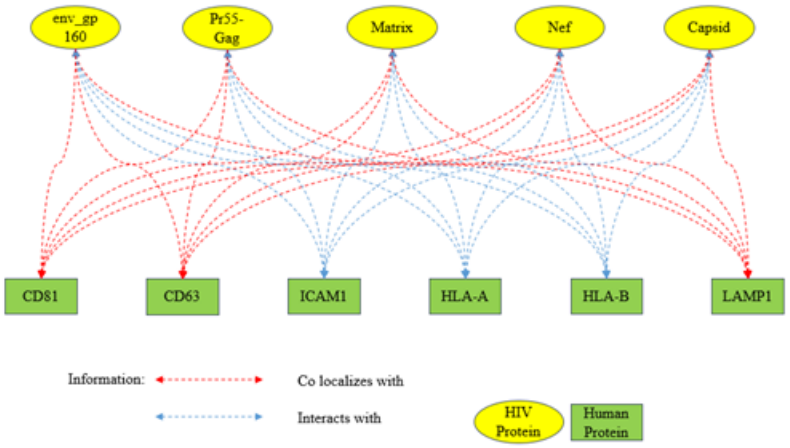

FIGURE 3. One of the Bicluster from Negative Dataset.

Based on Table 2, if the amount of HIV protein in one bicluster is 2, then there are 454 biclusters formed. The size of the bicluster formed has a minimum size of 2×2, meaning that there are 2 HIV proteins and two human proteins in one bicluster. There is 2 bicluster which has a minimum size of 6×6. Bicluster has 6 HIV proteins and a minimum of 6 human proteins.

Figure 2 is a bicluster of protein interactions in a positive dataset. Five HIV proteins interact with six human proteins. The type of interaction contained in the bicluster is a type of first-class interaction (”upregulates”) and third class type of interaction (”co-localizes with” and ”interacts with”). In the negative dataset (Table 3), if there are 2 HIV proteins in one bicluster, then there are 398 biclusters formed. If there are 3 HIV proteins in one bicluster, there are 259 biclusters. Thus if there are 6 HIV proteins in one bicluster, only one bicluster is formed. The more HIV proteins found in one bicluster, the smaller the number of biclusters formed.

5. DISCUSSION AND RESULTS

5.1. Prediction Method with the Association Rule Mining. In this paper, we divide the dataset matrix into two sub-matrices. We split that submatrix into HV matrix and VH matrix. Row in HV matrix as human’s proteins and column as HIV’s proteins. While in VH matrix, row as HIV’s protein and column as human’s proteins (Figure 4). The results of the bicluster obtained will be used to predict protein interactions that will occur. The first step taken is to form the rules of each dataset. The first type of rules are the rules of the positive dataset, and

the second type rules are the rules of the negative dataset. There are 837 rules of the first type (Table 4). Whereas for the second type of rules, there are 202 rules (Table 4). These rules (Table 4 and Table 5) have a confidence value above 0.8. If a rule has a confidence level of 1, it means that the rule results in a prediction of protein interactions that have already taken place.

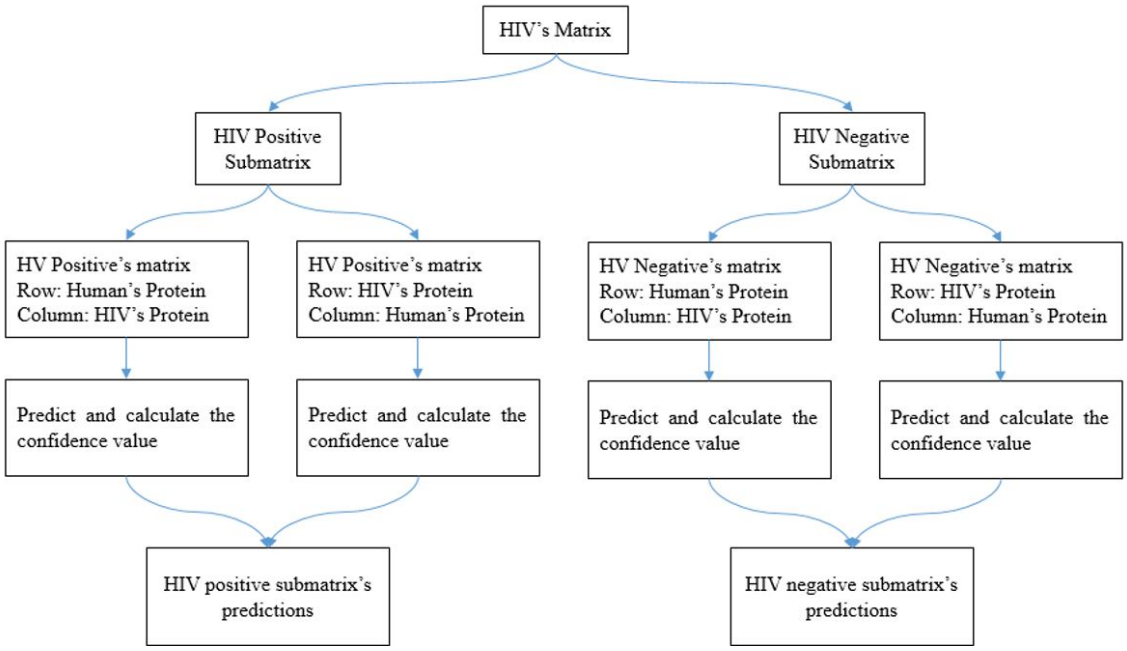

FIGURE 4. Association rule mining flowchart.

The rules that have obtained are then used to predict protein interactions between HIV-1 protein and human protein. The prediction of the first type rules will then be referred to as the first type prediction (Table 6), while the forecast of the second type rules will then referred to as the second type prediction (Table 7). There are 199 predictions of the first type (84 predictions are predictions that have already taken place, and 115 predictions are predictions that will occur). Inline 8 (Table 7), LCK (human protein) is activated (interaction type) by Tat (HIV protein-1). The CCL 5 protein (human protein) regulated by Nef and Vpr (both HIV proteins).

TABLE 4. The first type of rules.

No	Rules	Confidence value
1	$inhibits_T NF, upregulates_T NF \rightarrow upregulates_C CL2$	1
2	$activates_C ASP3, upregulates_T NF \rightarrow upregulates_C CL2$	1
3	$activates_C ASP3, upregulates_C CL2 \rightarrow inhibits_T NF$	1
4	$inhibits_T NF, upregulates_I L6 \rightarrow upregulates_C CL2$	0.833333
5	$activates_C ASP3, upregulates_I L6 \rightarrow upregulates_C CL2$	0.833333
6	$downregulates_I L2 \rightarrow activates_C ASP3$	1
7	$downregulates_I L2 \rightarrow activates_M APK1$	1
8	$activates_M APK1 \rightarrow downregulates_I L2$	0.857143
9	$activates_M APK3, upregulates_T NF \rightarrow upregulates_C CL5$	0.833333
10	$activates_M APK1, inhibits_T NF \rightarrow upregulates_C CL5$	0.833333
⋮	⋮	⋮
837	$interacts.with_A CTB \rightarrow activates_M APK1$	0.833333

TABLE 5. Second Type of Rules.

No	Rules	Confidence Value
1	$co.localizes.with_C D63, co.localizes.with_C D81 \rightarrow interacts.with_I CAM1$	1
2	$interacts.with_I CAM1, co.localizes.with_C D81 \rightarrow co.localizes.with_C D63$	1
3	$co.localizes.with_C D63, interacts.with_I CAM1 \rightarrow co.localizes.with_C D81$	0.833333
4	$interacts.with_C D3E \rightarrow interacts.with_C D3G$	0.857143
5	$interacts.with_I CAM1, interacts.with_C D3G \rightarrow interacts.with_E IF2AK2$	1
6	$co.localizes.with_C D63, interacts.with_H LA.B \rightarrow interacts.with_A CTA1$	0.833333
7	$interacts.with_I CAM1, co.localizes.with_C D63 \rightarrow interacts.with_A CTA1$	0.833333
8	$interacts.with_A CTA1 \rightarrow interacts.with_A CTA2$	0.857143
9	$co.localizes.with_C D63, interacts.with_C D3D \rightarrow interacts.with_H LA.B$	1
⋮	⋮	⋮
202	$interacts.with_I CAM1, interacts.with_C D3E \rightarrow interacts.with_E IF2AK2$	0.833333
837	$interacts.with_A CTB \rightarrow activates_M APK1$	0.833333

TABLE 6. First Type of Prediction

No	HIV-1 Proteins	Interaction Type	The Human Proteins	Confidence value
1.	$Matrix$	Interacts with	IFNG	0.833333
2.	$Matrix$	upregulates	TNF	0.875
3.	env_gp41	downregulates	CCR5	0.833333
4.	$Pr55(Gag)$	Interacts with	ICAM1	0.833333
5.	$Pr55(Gag)$	Interacts with	IFNG	0.833333
6.	env_gp160	Interacts with	ACTA1	0.833333
7.	$Pr55(Gag)$	Interacts with	CD3D	0.857143
8.	Tat	activates	LCK	0.833333
9.	Nef	Interacts with	HLA.B	0.857143
10.	$Gag-Pol$	Interacts with	CD3E	0.857143
11.	Nef	Interacts with	HLA.A	0.857143
12.	env_gp160	activates	CASP3	0.857143
13.	env_gp160	activates	FOS	0.857143
14.	env_gp41	activates	FOS	0.857143
15.	Nef	Co localizes with	CD63	1
16.	Vpu	Upregulates	CCL5	1
17.	Nef	Upregulates	CCL5	1
18.	env_gp120	Inhibits	TNF	1
19.	$Pr55(Gag)$	Co localizes with	CD63	1
20.	$Pr55(Gag)$	Co localizes with	LAMP1	1
21.	env_gp160	inhibits	TNF	1
22.	env_gp41	inhibits	TNF	1
23.	env_gp41	upregulates	CCL2	1
24.	Nef	inhibits	TNF	0.833333
25.	env_gp120	Interacts with	HLA.A	0.833333
26.	$Pr55(Gag)$	downregulates	CCR5	0.833333
⋮	⋮	⋮	⋮	⋮
199.	env_gp160	activates	FOS	0.857143

TABLE 7. Second Type of Prediction

No.	HIV-1 Proteins	Interaction Type	The Human Proteins	Confidence Value
1.	*Vpu*	interacts.with	CD3E	1
2.	*RT*	interacts.with	ICAM1	1
3.	*capsid*	interacts.with	ICAM1	1
4.	env_gp160	interacts.with	ICAM1	1
5.	reverse transcriptase p51 subunit	interacts.with	ICAM1	1
6.	matrix	interacts.with	ICAM1	1
7.	env_gp41	interacts.with	HLA.A	0.85714286
8.	*Vpu*	co.localizes.with	CD63	0.83333333
9.	env_gp41	co.localizes.with	LAMP1	0.85714286
⋮	⋮	⋮	⋮	⋮
58.	env_gp160	interacts.with	EIF2AK2	0.83333333

TABLE 8. The Final prediction of interaction between hiv-1 protein with human protein

No.	HIV-1 Proteins	Interactions Type	The Human Proteins	Confidence Value
1.	Nef	co.localizes. with	CD63	1
2.	env_gp160	inhibits	TNF	1
3.	$Pr55(Gag)$	upregulates	IL6	1
4.	$Pr55(Gag)$	activates	MAPK3	0.833333
5.	$Gag-Pol$	interacts.with	CD3D	0.857143
6.	Nef	inhibits	TNF	0.833333
7.	$Matrix$	upregulates	IFNG	0.875
8	$Matrix$	upregulates	IL10	0.833333
9	Nef	interacts.with	CD3D	0.857143
10	$Matrix$	upregulates	CCL2	0.833333
11	env_gp120	interacts.with	ICAM1	0.857143
12	env_gp160	interacts.with	CD3G	0.857143
13	env_gp120	upregulates	IL6	0.833333
14	Vpu	interacts.with	ICAM1	0.833333
15	$Capsid$	interacts.with	LCK	0.833333
16	env_gp160	upregulates	IL6	0.833333
17	Nef	interacts.with	ACTG2	0.833333
⋮	⋮	⋮	⋮	⋮
210	env_gp41	interacts.with	HLA.DRB1	70.83333333

FIGURE 5. Association rule mining flowchart.

In the second type of prediction (Table 7), there are 58 predictions (31 predictions that have occurred and 27 predictions are predictions that will happen). EIF2AK2 protein (human protein) interacts with vpu protein (HIV protein) with a confidence level of 0.8. EIF2AK2 protein also interacts with envelope glycoprotein 160 protein (HIV protein) with a confidence level of 0.83.

These two types of predictions are then combined. Combining these two predictions is referred to as the final prediction (Table 8). The type of interaction predicted in this final prediction is the type of interaction that is in the first class and third class. There are 210 predictions of protein interactions between the HIV-1 protein and human protein. Most predictions are at the confidence level with a range of 0.8010-0.85, which is 97 predictions. Predictions are at least at the level of confidence with a range of 0.8501 to 0.90, as many as 44 predictions. The level of confidence with a range of 0.9501-1.0 is the confidence value that has the second-highest forecast, which is 68 predictions.

Figure 5 is a visualization of predicted protein interactions that will occur. The HIV-1 Nef protein predicted to interact with the human protein ACTG1, ACTA2, ACTG2, and ACTC1. This type of interaction between HIV protein and human protein interacts with. TNF human

protein with upregulates interaction type predicted to interact with gp120 glycoprotein HIV envelope protein and matrix. ICAM1 human protein with interaction type interacts with, is predicted to interact with eight types of HIV proteins, namely envelope glycoprotein gp41, nef, matrix, Gag-Pol, envelope glycoprotein gp120, Retropepsin, Pr55 (Gag), and Vpu. The most predicted HIV protein will interact with human protein is Nef protein, which is as many as 27 predictions.

6. CONCLUSIONS

The LCM-MBC algorithm can produce bicluster from a dataset of protein interactions between HIV-1 protein and human protein. This algorithm provides 852 biclusters (454 biclusters from the positive dataset and 398 from the negative dataset) with at least two HIV proteins in one bicluster. If there are at least three HIV proteins in one bicluster, there are 572 biclusters. A total of 211 biclusters has four HIV proteins in one bicluster. When there is a minimum of five HIV bicluster proteins in one, 45 biclusters obtained. Forty-five biclusters have HIV protein in one bicluster of five. The results of the LCM-MBC algorithm used to predict protein interactions between HIV-1 protein and human protein. This prediction uses the association rule mining algorithm. There are 837 rules of the first type and 202 rules for the second type. The rules obtained which have the confidence value above 0.8 used to predict. There are 210 predictions of protein interactions between the HIV-1 protein and human protein. Most predictions are at the confidence level with a range of 0.8010-0.85, which is 97 predictions. Predictions are at least at the level of confidence with a range of 0.8501 to 0.90, as many as 44 predictions. The level of confidence with a range of 0.9501-1.0 is the confidence value that has the second-highest prediction, which is 68 predictions. There are 11 HIV-1 proteins which predicted to interact with 36 human proteins. The type of interaction predicted in this final prediction is the type of interaction that is in the first class and third class. If human protein connected with HIV protein according to the type of interaction, it means that human protein interacts with HIV-1 proteins. If there are no rules at the confidence value with a specific range, then no predictions are generated at the confidence value with that range. From this research, some suggestions can be made for further research, namely: Applying other prediction algorithms, so that the

algorithm can obtain the best (both in computing time and the results received). The researcher can also divide the dataset into three submatrices.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] A.D. Frankel, J.A.T. Young, HIV-1: Fifteen Proteins and an RNA, Annu. Rev. Biochem. 67 (1998), 1–25.

[2] H.S. Chan, K.A. Dill, The Protein Folding Problem, Phys. Today. 46 (1993), 24–32.

[3] T.S. Permata, A. Bustamam, Clustering protein-protein interaction network of TP53 tumor suppressor protein using Markov clustering algorithm, in: 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE, Depok, Indonesia, 2015: pp. 221–226.

[4] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, A Novel Biclustering Approach to Association Rule Mining for Predicting HIV-1–Human Protein Interactions, PLoS ONE. 7 (2012), e32289.

[5] A. Bustamam, M. S. Sehgal, N. A. Hamilton, S. Wong, M. A. Ragan, K. Burrage, An efficient parallel implementation of Markov clustering algorithm for large-scale protein-protein interaction networks that uses MPI, in: 5th IMT-GT international conference on mathematics, statistics and their applications, (2009).

[6] B. Mirkin, Mathematical Classification and Clustering, Kluwer Academic Publishers, Dordrecht, 1996.

[7] J. Li, G. Liu, H. Li, L. Wong, Maximal Biclique Subgraphs and Closed Pattern Pairs of the Adjacency Matrix: A One-to-One Correspondence and Mining Algorithms, IEEE Trans. Knowl. Data Eng. 19 (2007), 1625–1637.

[8] O. Swasti, A. Bustamam, D. Lestari, W. Mangunwardoyo, Biclustering protein interactions between HIV-1 proteins and humans proteins using LCM-MBC algorithm, in: Depok, Indonesia, 2019: p. 020015.

[9] A. Bustamam, T. Siswantining, T.P. Kaloka, O. Swasti, Application of BiMax, POLS, and LCM-MBC to Find Bicluster on Interactions Protein between HIV-1 and Human, Austrian J. Stat. 49 (2020), 1–18.

[10] S. Bandyopadhyay, U. Maulik, L.B. Holder, Advanced Methods for Knowledge Discovery from Complex Data, Springer, Dordrecht, 2005.

[11] O. Tastan, Y. Qi, J.G. Carbonell, J. Klein-Seetharaman, Prediction of Interactions Between Hiv-1 and Human Proteins by Information Integration, in: Biocomputing 2009, World Scientific, Kohala Coast, Hawaii, USA, 2008: pp. 516–527.

[12] S. Park, J.A. Reyes, D.R. Gilbert, J. Kim, S. Kim, Prediction of protein-protein interaction types using association rule based classification, BMC Bioinform. 10 (2009), 36.

[13] I.N.G. da Cruz, PPRINT: Prediction of Protein-Protein Interactions, (2014). http://hdl.handle.net/10316/35674

[14] A. Trkola, HIV–host interactions: vital to the virus and key to its inhibition, Current Opinion Microbiol. 7 (2004), 407–411.

[15] A. Mukhopadhyay, S. Ray, U. Maulik, Incorporating the type and direction information in predicting novel regulatory interactions between HIV-1 and human proteins using a biclustering approach, BMC Bioinform. 15 (2014), 26.

[16] R.B. Bapat, Graph and Matrices, Springer, London, 2006.

[17] Y. Wang, S. Cai, M. Yin, New heuristic approaches for maximum balanced biclique problem, Inform. Sci. 432 (2018), 362–375.

[18] Y. Cheng, G.M. Church, Biclustering of expression data. Proc. Int. Conf. Intel.l Syst. Mol. Biol. 8 (2000), 93–103.

[19] W.H. Haemers, Bicliques and Eigenvalues, J. Comb. Theory, Ser. B. 82 (2001), 56–66.

[20] T. Siswantining, T. Anwar, D. Sarwinda, H.S. Al-Ash, A novel centroid initialization in missing value imputation towards mixed datasets, Commun. Math. Biol. Neurosci. 2021 (2021), Article ID 11.

[21] T. Siswantining, N. Saputra, D. Sarwinda, H.S. Al-Ash, Triclustering Discovery Using the δ-Trimax Method on Microarray Gene Expression Data, Symmetry. 13 (2021), 437.

[22] J. Hipp, U. Güntzer, G. Nakhaeizadeh, Algorithms for association rule mining – a general survey and comparison, SIGKDD Explor. Newsl. 2 (2000), 58–64.

[23] T.A. Kumbhare, S.V. Chobe, An overview of association rule mining algorithm, Int. J. Computer Sci. Inform. Technol. 5 (2014), 927-930.

[24] J.W. Pinney, J.E. Dickerson, W. Fu, B.E. Sanders-Beer, R.G. Ptak, D.L. Robertson, HIV–host interactions: a map of viral perturbation of the host system, AIDS. 23 (2009), 549–554.

[25] R.G. Ptak, W. Fu, B.E. Sanders-Beer, et al. Short Communication: Cataloguing the HIV Type 1 Human Protein Interaction Network, AIDS Research and Human Retroviruses. 24 (2008), 1497–1502.

[26] W. Fu, B.E. Sanders-Beer, K.S. Katz, D.R. Maglott, K.D. Pruitt, R.G. Ptak, Human immunodeficiency virus type 1, human protein interaction database at NCBI, Nucleic Acids Res. 37 (2009), D417–D422.