Austrian Journal of Statistics
February 2020, Volume 49, 1–18.
http://www.ajs.or.at/
doi:10.17713/ajs.v49i3.1011

Application of BiMax, POLS, and LCM-MBC to Find Bicluster on Interactions Protein between HIV-1 and Human

Alhadi Bustamam Universitas Indonesia
Titin Siswantining Universitas Indonesia
Tesdiq P. Kaloka Universitas Indonesia
Olivia Swasti Universitas Indonesia

Abstract

Biclustering, in general, is a process of clustering genes and conditions simultaneously rather than clustering them separately. The purpose of biclustering is to discover a subset from experimental data. Further, biclustering results can be analyzed from a biological perspective. Biclustering can also be used for protein-protein interaction. In protein-protein interaction, biclustering can cluster interactions based on rows and columns. In this research, we applied three biclustering algorithms based on graph approach, Binary inclusion-Maximal (BiMax), local search framework based on pairs operation (POLS), and (LCM-MBC) to clustering data of protein-protein interaction between HIV-1 and human. We change the interaction protein-protein interaction data into binary then divided into two datasets called HV positive and HV negative. Then compare the biclustering results of each dataset using heatmap and analyze them with GO terms. From dataset HV positive, BiMax found 30 biclusters, LCM-MBC 31 biclusters, and POLS 13 biclusters. From dataset HV negative, BiMax found eight biclusters, LCM-MBC 14 bicluster, and POLS 10 biclusters. Based on the results of the heatmap, all bicluster entry from BiMax is a protein that interacts, whereas biclusters entry of LCM-MBC and POLS still have proteins that do not interact. It can be concluded that BiMax algorithm is good for clustering protein-protein interaction, especially for binary data.

Keywords: protein-protein interaction, biclustering, graph, BiMax, POLS algorithm, LCM-MBC.

1. Introduction

Proteins are part of an organism, to carry out their functions properly, proteins must interact with other proteins. Therefore, knowledge about protein-protein interaction is vital for identifying and investigating biological processes in cells Lestari, I. S. Musti, and Bustamam (2018). Understanding protein-protein interaction can help health experts to prevent and even treat various chronic diseases like HIV and AIDS. Acquired Immunodeficiency Syndrome (AIDS) is caused by the Human Immunodeficiency Virus (HIV) virus that attacks or interacts with humans. HIV whose attack on the immune system and weakens the body's ability to fight infections and diseases. This virus will exploit the host-cell machine thus this virus can successfully produce offspring and at the same time also avoid the immune system

Trkola (2004).

Biclustering is the process of clustering data based on rows and columns in the data matrix. Biclustering is commonly used in biology especially for clustering gene expression data because traditional clustering methods can't find patterns that are suitable for one cluster. The result from biclustering methods give much better information than clustering methods because of the better enrichment value of the resultant clusters Singh, Nagrare, Srikanth, Kumar, and Dwith (2011). Biclustering can also be interpreted by combining 2 clustering methods. In 2017 Ardaneswari, Bustamam, and Siswantining (2017b) combines the k-means algorithm and the chang and chruch algorithm in tumor carcinoma. The latest research was conducted by Bustamam, Formalidin, and Siswantining (2018a) by combining SVD methods and hybrid clustering on microarray data. In 2018 Bustamam, Zubedi, and Siswantining (2018c) using co-similarity and agglomerative hierarchical methods, another study by Bustamam, Puspa, and Siswantining (2018b) applied co-similarity and k-means partition algorithm to microarray lymphoma data. As time goes by, biclustering is not only used for gene expression data but also protein-protein interaction data. Biclustering is a powerful analytical tool for biologists and has generated significant interest over the past few decades Ardaneswari, Bustamam, and Sarwinda (2017a). Based on the value, biclustering has some types like constant bicluster, constant row bicluster, constant column bicluster, addition pattern bicluster, and multiplication pattern bicluster. Biclustering algorithm can be divided into iterative biclustering, fuzzy biclustering and graph approach biclustering depend on the theory was used Mukhopadhyay, Maulik, and Bandyopadhyay (2010).

Some work that has been done related to this research, in 2014 Mukhopadhyay, Ray, and Maulik (2014) predicted the protein-protein interaction between HIV-1 and humans using Association Rule Mining (ARM) based on BiMax biclustering results. In 2006 Prelic, Bleuler, Zimmerman, Wille, Buhlamn, Gruissem, Hennig, Thiele, and Ziztler (2006) proposed a new biclustering method based on graph theory and used a fast divide-and-conquer approach, BiMax also capable of finding all optimal bicluster for binary data. A new biclustering algorithm called LCM-MBC, combined of the Least Common Multiple (LCM) and Modular Input Consensus Algorithm (MICA) to search bipartite subgraph introduced by Liu, Li, Wong, and Li (2007). LCM-MBC can find bipartite subgraph faster than LCM and MICA in dense data with a large number of vertices. In 2018 Wang, Cai, and M (2006) proposed the local search framework based on pairs operation (POLS) method that uses two novel pair operations called addset and dropset with the basic operation is graph theory. Algorithm in Liu *et al.* (2007) and Wang *et al.* (2006) are algorithms to find bicluster by searching for bipartite subgraph in the form of biclique graph. The difference between Liu *et al.* (2007) and Wang *et al.* (2006) is if the LCM is only searching for biclique graph while the POLS is balanced biclique graph.

Heatmap is a visualization method that can make it easier to understand bicluster results. If using biological data, then the final results need to be biologically validated. GO term is a place to confirm data gene and protein Santamaria, Theron, and Quintales (2008). Another work was done by Yang, Shen, Yuan, Zhang, and Wei (2017), combining Affinity Propagation (AP) clustering and Iterative Signature Algorithm (ISA) biclustering to analyze breast cancer. From the mixed results of the two methods, nine biclusters were obtained. Then biologically validated with GO terms and KEGG pathways from 9 biclusters. From GO terms and KEGG pathways based on their genes, the breast cancer has seven sub-types.

The purpose of this study was to compare the results of three biclustering algorithms that are BiMax, POLS, and LCM-MBC using heatmap and GO terms in data of protein-protein interaction between HIV-1 and humans. protein-protein interaction data obtained from Fu, Sanders-Beer, Katz, Maglott, Pruitt, and Ptak (2008) can be accessed from NCBI then the data is converted into binary value, 0 if there is no interaction and one if there is interaction.

2. Data

Data protein-protein interaction between HIV-1 and human obtained from Fu *et al.* (2008) consist of 23 HIV-1 proteins and 3797 human proteins. The data also consists of HIV-1 gene ID, HIV-1 protein names, human gene ID, human gene symbol, human protein name, and interaction description. We only take HIV-1 protein names, protein human gene symbol, and keywords (interaction type) because these three data types are the core information of protein-protein interaction between HIV-1 and human to make the dataset. According to data at NCBI, we divided the interaction type into three types. The first is type-1, consist of interaction from HIV-1 to human, second is type-2 consist of interaction from human to HIV-1, and third is type-3, interaction bidirectional. There were 69 interaction types in type-1, 48 interaction types in type 2, 13 interaction types in type-3.

2.1. Dataset algorithm

Data protein-protein interaction must be change into a dataset so it can be processed mathematically. In this case, data transform into a bipartite graph with HIV-1, and human is the vertices, nodes are the interaction types, and the edges are the value of each interaction, 0 if there was no interaction and 1, -1, or X if there was an interaction. Value 1 if the interaction type-1, -1 if interaction type 2, and X if interaction type-3. Then Split this dataset into two submatrices, first submatrix consist of value 0, 1, and X called HV Positive and second submatrix consist of value 0, -1, and X called HV Negative. Dataset HV Positive consist of interaction type-1 and type-3 and dataset HV Negative consist of interaction type-2 and type-3. Algorithm for making the dataset could see in Table 1.

Table 1: Dataset algorithm

Input		Data HIV-1 protein, human protein, and interaction types
	1	Make a matrix $M_{m\times n} = a_{ij} = 0$ from bipartite graph with vertices are HIV-1 protein and human protein
	2	Divided interaction type into three types (type-1, type-2, and typ-3)
	3	Change the entry matrix If interaction type == **type-1**, then $a_{ij} = 1$, else If interaction type == **type-2**, then $a_{ij} = -1$, else If interaction type == **type-3**, then $a_{ij} = X$, else
	4	Make two sub-matrices from the matrix in step 3 **HV Positive** for a matrix which $a_{ij} = 0$, 1, and X **HV Negative** for a matrix which $a_{ij} = 0$, -1, and X
	5	Change all a_{ij} =-1 and X to 1
Output		Dataset HV Positive and HV Negative

3. Method and theory

There has been much research using biclustering on biological data, especially for gene expression data. In this research, we tried to make binary data from protein-protein interaction and then apply and compare three biclustering algorithms, BiMax, POLS, and LCM-MBC. More clearly the research flowchart could see in Figure 1.

We made two comparisons from the results of 3 algorithms based on the effectiveness of the algorithm to group data. First using heatmap to determine the accuracy of the classification

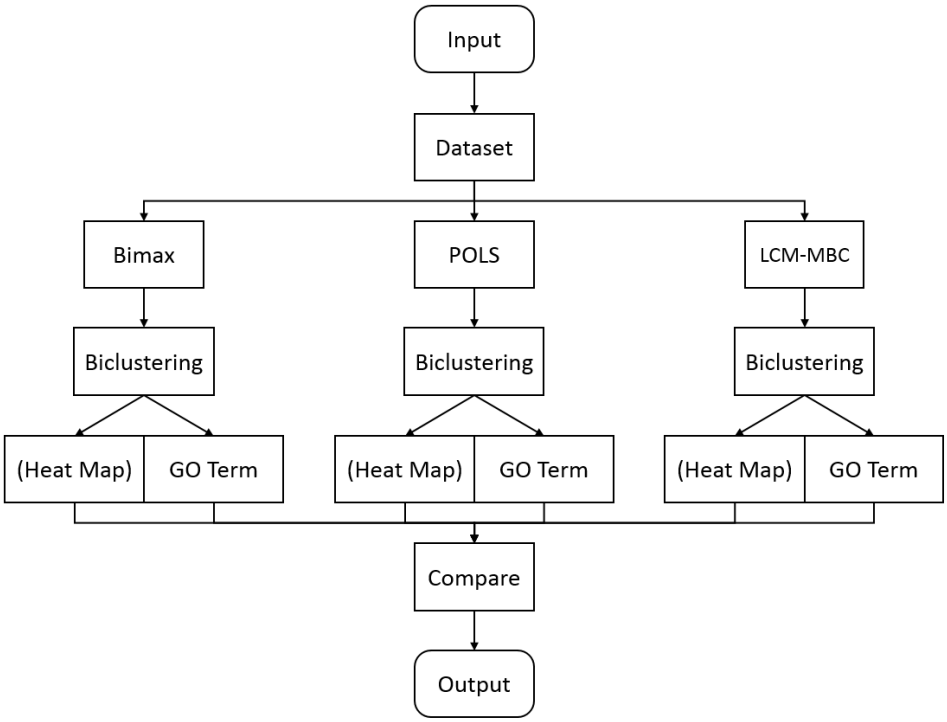

Figure 1: Flowchart researh

of each algorithm, secondly to ensure the results of biclustering can be interpreted biologically with the help of GO terms.

3.1. Graph theory

Suppose $G = G(V, E)$ is a graph, define as a pair of set (V, E) with $V = \{v_1, v_2, v_3, \ldots, v_n\}$ is a non empty set called vertex and $E = \{e_1, e_2, e_3, \ldots, e_i\}$ called edge. The number of vertex denoted by $|V|$ and the number of edge denoted by $|E|$.

Suppose $G = G(U, V, E)$ is a bipartite graph, with G can be divided into two disjoint vertex sets $U = \{u_1, u_2, u_3, \ldots, u_m\}$ and $V = \{v_1, v_2, v_3, \ldots, v_n\}$ such that everey edge connects one vertex in U and one vertex in V. The neighborhood of a vertex $u \in U$ is $N(u) = \{v \in V | (v, u) \in E\}$ and the neighborhood of a vertex $v \in V$ is $N(v) = \{u \in U | (v, u) \in E\}$. A complete bipartite graph called biclique, if $|U| = |V| = |G|$ then graph G called balanced biclique Wang *et al.* (2006).

3.2. Biclustering

Biclustering was known before 2000 but still known as some terms like direct clustering, box clustering, or simultaneous clustering row and column Cheng and Chruch (2000). Given a matrix $A(D, C)$ with D is a set of data $D = \{I_1, I_2, I_3, \ldots, I_G\}$ and C a set of condition $C = \{J_1, J_2, J_3, \ldots, J_G\}$. A bicluster is a submatrix $M(I, J) = [m_{ij}]$, $i \in I$, and $j \in J$ of matrix $A(D, C)$ where $I \subseteq G$and $J \subseteq C$.

Graph theory has been applied to detecting bicluster, some algorithm used graph theory is SAMBA, BiMax, POLS, and LCM-MBC. Graph-based algorithm changed data into a bipartite graph, with two sets of nodes is data and its conditions respectively and the edge is a value of from data and conditions. A bicluster is a subgraph of the bipartite graph Mukhopadhyay *et al.* (2010).

3.3. BiMax

Given a matrix $E^{(n\times m)}$ a binary matrix, a bicluster (G, C) corresponds to a subset $G \subseteq \{1, \ldots, n\}$ respond across a subset $C \subseteq \{1, \ldots, m\}$, that mean the pair of (G, C) is a submatrix of E. The pair $(G, C) \in 2^{\{1,\ldots,n\}} \times 2^{\{1,\ldots,m\}}$ called an inclussioin-maximal if and only if

$$\forall i \in G, j \in C : e_{ij} = 1 \tag{1}$$

$$\exists (G', C') \in 2^{\{1,\ldots,n\}} \times 2^{\{1,\ldots,m\}} \tag{2}$$

with $\forall i' \in G', j' \in C' : e_{i'j'} = 1$ and $G \subseteq G' \wedge C \subseteq C' \wedge (G', C') \neq (G, C)$

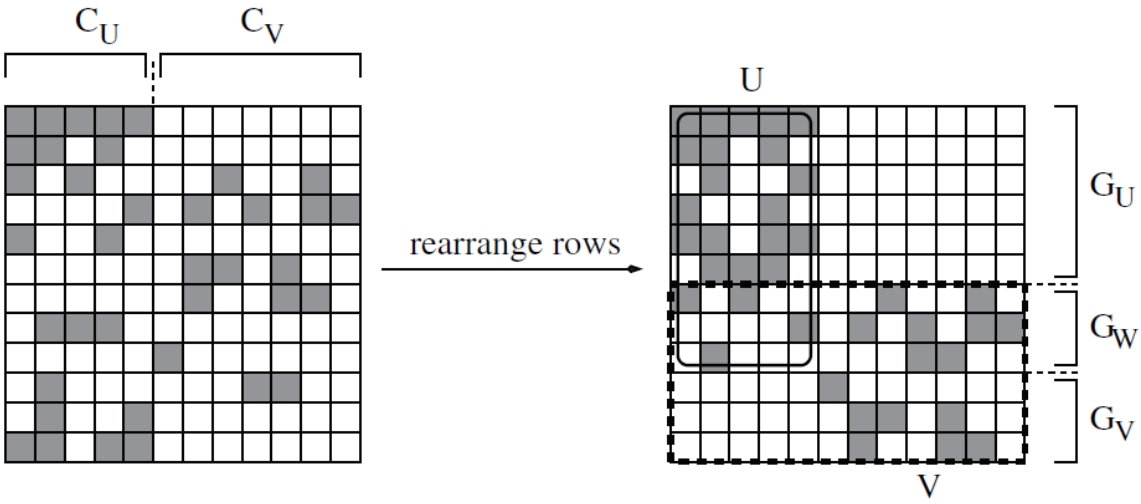

Figure 2: BiMax illustration

The idea of the BiMax algorithm (Figure 2) is first divided matrix E into two submatrices C_U and C_V which is divided according to the column. Then specify the rows corresponding to C_U, C_U and C_V, and for the last is C_V, So that there are two sub matrices U and V Mukhopadhyay *et al.* (2014)

3.4. POLS algorithm

POLS is an algorithm that has some properties such as and candidate solution, Addset, Dropset, pscore, and Add rule. Addset is a set of a dataset that will be added to the candidate solution, Dropset is a set of candidate solution, pscore is a value of each pair of edge (u, v) and add rule is a rule to select the best pairs of edges, candidate solution a place to put a temporary solution until the end of the process Wang *et al.* (2006).

Given a bipartite graph $G = (U, V, E)$ with vertex $U = \{u_1, u_2, u_3, \ldots, u_n\}$, vertex $V = \{v_1, v_2, v_3, \ldots, v_n\}$, and edge $E = \{e_1, e_2, e_3, \ldots, e_i\}$. If the solution of balanced biclique denoted by $S = (U^s, V^s, E^s)$ then addset is a set such.

$$Addset = \{(u, v) \in E | u \notin U^s, u \in N(v'); \forall v' \in V^s, v \notin V^s, v \in N(u'); \forall u' \in U^s\} \tag{3}$$

$$Dropset = \{(u, v) \in E | u \in U^S, v \in V^S\} \tag{4}$$

Pscore for pair (u, v) from the addset denoted by $pscore(u, v)$ can be find by

$$pscore(u, v) = score_{lb}(u, v) + \lfloor \frac{score_{ub}}{2} \rfloor \tag{5}$$

with $score_{lb}(u, v) = 1$ and $score_{ub}(u, v) = min\{|N(v)| - |N(v) \cap S|, |N(u)| - |N(u) \cap S|\}$ The biggest *pscore* called **add rule**. POLS algorithm could see in Table 2.

Table 2: POLS algorithm

Input		Bipartite graph $G = (U, V, E)$
	1	Initialize $S = \emptyset$, Addset, Dropset, and $S^* = S$
Repeat		
	2	Select a pair (v, u) from Addset using Add rule
	3	Put the pair to a candidate solution $S = S \cup (v, u)$
Until		
	4	$Addset = \emptyset$
	5	if $\|S\| > \|S^*\|$ then $S^* = S$ Remove pairs of vertex from Dropset
Output		Dataset HV Positive and HV Negative

The first step in POLS algorithm is initialized candidate solution, Addset, and Dropset. In step 2 and step 3 iteratively select pair from Addset using Add rule then put the pair to candidate solution. The last step is to find the best solution.

3.5. LCM-MBC algorithm

LCM-MBC is an algorithm to find a bicluster from the biclique graph that represents an enumeration tree Liu *et al.* (2007) LCM-MBC has two parameters a as the row data, and b is the column of data, where $a \leq b$ for the input (see Table 3).

Table 3: LCM-MBC algorithm

Input		Bipartite graph, a, and b
Repeat		
	1	combine the a
	2	If column on step 1 == 1 then combined the column in 1 bicluster
Until		
	3	$a < b$
Output		Biclique graph

The explanation of LCM-MBC (Table 3) easily understood by the illustration. Given the data of an interaction between four HIV-1 proteins and four human proteins (Figure 3), 0 means there is no interaction and 1 means there is an interaction.

The first step combines the first row (Gag) and the fourth row (Nev). The second step combined the column which has value one see Figure 4.

The last step is repeating step 1 and 2, from Table **??**e have four biclusters. First, a bicluster is Gag and Nev interact with HLA-A and HLA-B, the second bicluster is Gag and Nev interact with HLA-A and BECN1, the third bicluster is Gag and Nev interact with HLA-B and BECN1, and for the fourth bicluster Gag and Nev interact with HLA-A, HLA-B, and BENC1. Result for the illustration can be seen in Table 4.

	HLA-A	HLA-B	BECN1	PLA2G5
Gag	1	1	1	0
Rev	0	1	0	1
Tat	0	0	1	0
Nev	1	1	1	0

Figure 3: Data illustration LCM-MBC

	HLA-A	HLA-B	BECN1
Gag	1	1	1
Rev	0	1	0
Tat	0	0	1
Nev	1	1	1

Figure 4: Data after step 1 and step 2

4. Results and discussion

Based on chapter 2, this study used two datasets which were named HV Positive and HV Negative so that each BiMax, POLS and LCM-MBC algorithms had two results. All processes from making dataset until find bicluster results are carried out using R software version 3.5.0, specifically for BiMax algorithm we used the BICLUST package Kaiser, Santamaria, Khamiakova, Sill, Theron, Quintales, Leisch, and Troyer (2018) which is a package in R and available at https://CRAN.R-project.org/package=biclust. Each bicluster is arranged to have a minimum of five HIV-1 proteins (lower bound) so that the results are similar. Besides that, the lower bound ensures that each bicluster can be seen biologically through GO terms.

4.1. HV positive

BiMax found 30 biclusters with five HIV-1 proteins as the lower bound. Based on the number of HIV-1 and human protein, we have 23 biclusters with five HIV-1 protein and seven biclusters with six HIV-1 proteins. The maximum human protein is 13 proteins and is only found in 1 bicluster while the minimum human protein is five proteins and founded in seven biclusters (see Table 5).

Member of the five biclusters are as follows; First bicluster consists of six HIV-1 and six human protein, HIV-1 proteins are env_gp120, env_gp160, Pr55(Gag), matrix, RT, and Tat. Also, the

Table 4: The result from illustration LCM-MBC

Bicluster	HIV-1 Protein	Human Protein
1	Gag, Nev	HLA-A, HLA-B
2	Gag, Nev	HLA-A, BECN1
3	Gag, Nev	HLA-B, BECN1
4	Gag, Nev	HLA-A, HLA-B, BECN1

Table 5: Biclustering BiMax HV positive

Protein		Bicluster
HIV-1	Human	
5	5	19, 21, 22, 26
5	6	3, 6, 16, 20, 23, 27
5	7	4, 12, 18
5	8	7, 9, 24
5	9	10, 17, 29
5	11	13, 25
5	12	15
5	13	30
6	5	2, 5, 8
6	6	1, 11, 28
6	10	14

human proteins are ACTC1, ACTA2, ACTG1, ACTG2, and ICAM1. The second bicluster consists of env_gp120, env_gp160, Nef, Tat, Vpr, and Vpu are HIV-1 protein, CASP3, FOS, JUN, CD4, and TNF are human proteins (six HIV-1 and five human). The third bicluster consists of env_gp120, Nef, Tat, Vpr, and Vpu are five HIV-1 proteins, CASP3, FOS, JUN, CD4, PARP1, and TNF are human proteins. The fourth bicluster consists of five HIV-1 proteins (env_gp120, env_gp160, Nef, Tat, and Vpu) and seven human proteins (CASP3, FOS, JUN, CD4, TNF, CD3D, and CD3E). Fifth bicluster consists of env_gp120, env_gp160, Pr55(Gag), Capsid, matrix, Nef, CD63, LAMP1, HLA-A, ICAM1, and HLA-B.

The bicluster results from BiMax are checked using heatmap, the red color in Figure 5 shows the value 1. All bicluster in Figure 5 is red, meaning that each bicluster classifies interactions that have a value of 1. So, each bicluster classifies protein-proteins that interact with each other. Furthermore, the results of the bicluster are checked through the GO terms using the gene ID of each protein found in one bicluster. Table 6 shows three types of GO terms for each computer, Biological process, cellular component, and Molecular function, all results of bicluster have biological interpretations. If seen in Table 6 cluster 2 and 3 have the same cellular component, this is only natural because there are the same members in bicluster 2 and 3.

Table 6: GO terms bicluster BiMax HV positive

Bicluster	GO Term Biological Process	GO Term Cellular Component	GO Term Molecular Function
1	mesenchyme migration	Filopodium	ATP Binding
2	response to bacterium	membrane raft	macromolecular complex binding
3	hemopoiesis	membrane raft	R-SMAD binding
4	hemopoiesis	T-Cell receptor complex	protein dimerization activity
5	regulation of leukocyte mediated cytotoxicity	cell surface	TAP binding

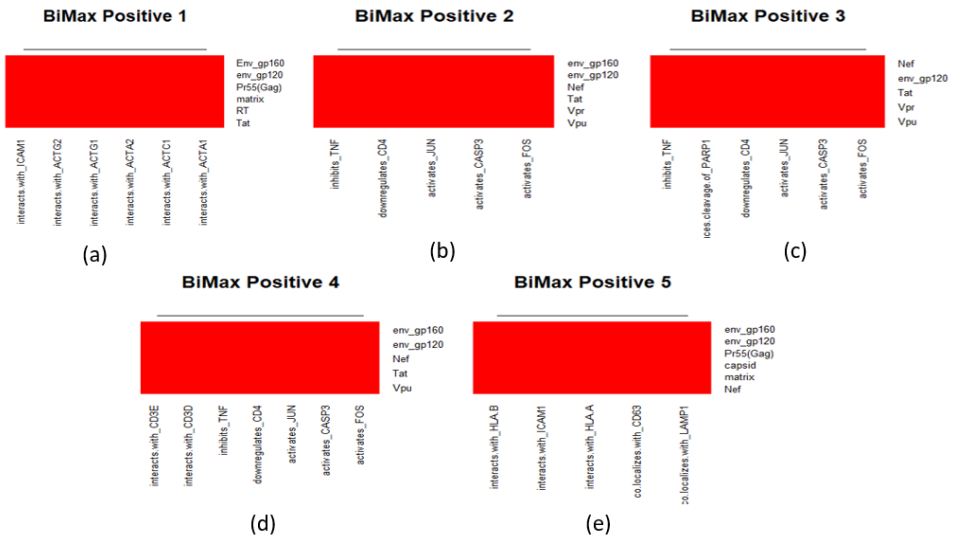

Figure 5: Heatmap 5 biclusters HV positive BiMax; (a) Bicluster 1; (b) Bicluster 2; (c) Bicluster 3; (d) Bicluster 4; (e) Bicluster 5.

Bicluster in POLS algorithm is a balanced biclique, so the number of HIV-1 protein and human protein always the same. POLS algorithm found 13 biclusters with one biggest bicluster consist of seven HIV-1 proteins and seven human proteins. There were four biclusters with five HIV-1 proteins and five human proteins (see Table 7).

HIV-1 protein in the bicluster 1-5 almost the same only, just a few different but the human proteins are different. Proteins in bicluster 1 are env_gp120, env_gp160, env_gp41, Pr55(gag), Capsid, Matrix, IFITM1, IFITM2, IFITM3, EIF2AK2, IL7, and HLA-B. Proteins in bicluster 2 are env_gp120, env_gp160, env_gp41, Pr55(gag), Capsid, Matrix, IL6, IL6R, CXCL8, XCL1, LAMP1, and HLA-DRB1. Proteins in bicluster 3 are env_gp120, env_gp160, env_gp41, Pr55(gag), Capsid, Matrix, Nucleocapsid, IL6, NFKB1, CXCL12, CD3D, CD3E, CD3G, and TNF. Proteins in bicluster 4 are env_gp120, env_gp160, env_gp41, Pr55(gag), Capsid, Matrix, CD3D, CD3E, CD3G, CD4, CD28, and TXNDC5. Proteins in bicluster 5 are env_gp120, env_gp160, env_gp41, Pr55(gag), Capsid, IFNA8, IFNB1, IL1A, IL1B, and IL10.

Table 7: Biclustering POLS algorithm HV positive

Protein		Bicluster
HIV-1	Human	
5	5	5, 6, 7, 9, 13
6	6	1, 2, 4, 8, 10, 11, 12
7	7	3

The proteins that interact in the results of the POLS algorithm bicluster can be seen in Figure 6. The red color shows the interaction with value 1 and green indicates the interaction with value 0. POLS algorithm still classify proteins that do not interact into a bicluster, because *pscore* (Equation 5) is never 0 so the POLS even clustering 0 and 1 into a bicluster. Then bicluster results validated with GO terms. The results show that even though there are proteins that do not interact, the bicluster has a biological interpretation. In Table 8, Bicluster 1 does not have a cellular component and a molecular function. Bicluster 1 does not have a cellular component and molecular function because there is no biological research on protein-protein interaction that occur in bicluster 1.

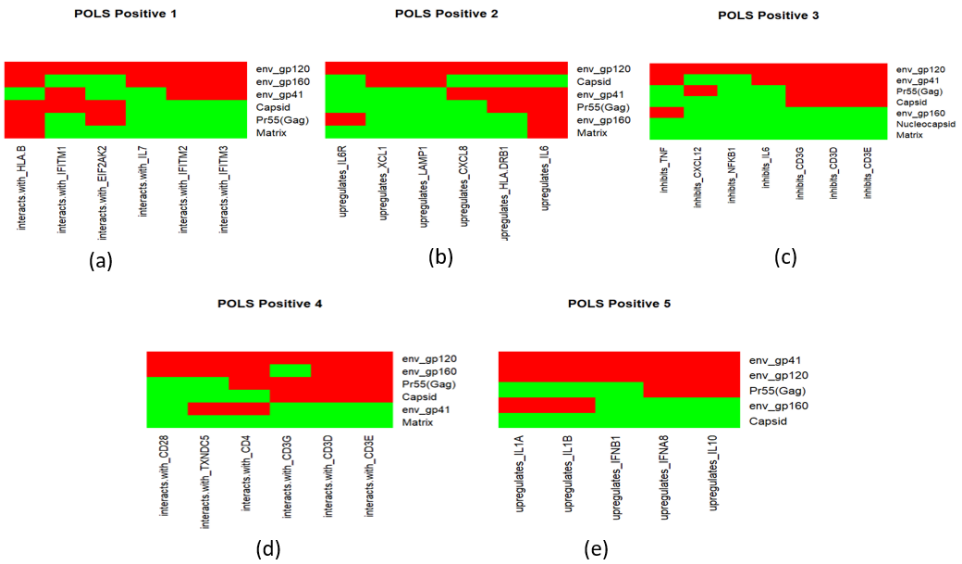

Figure 6: Heatmap 5 biclusters HV positive POLS; (a) Bicluster 1; (b) Bicluster 2; (c) Bicluster 3; (d) Bicluster 4; (e) Bicluster 5.

Table 8: GO terms POLS algorithm HV positive

Bicluster	GO Term Biological Process	GO Term Cellular Component	GO Term Molecular Function
1	response to interferonalpha	-	-
2	positive regulation of immune system process	cytokine receptor binding	extracellular region part
3	regulation of leukocyte cell-cell adhesion	protein heterodimerization activity	alpha-beta T cell receptor complex
4	T cell costimulation	signaling receptor activity	T cell receptor complex
5	lymphocyte proliferation	cytokine activity	extracellular space

LCM-MBC algorithm found 31 biclusters with five HIV-1 proteins as the lower bound, we have 29 biclusters with five HIV-1 protein and only two biclusters with six HIV-1 proteins. There is only one bicluster consist of 22 human protein and eight biclusters with five human proteins (see Table 9).

From Table 9, we can see that bicluster 1-5 have the same number of HIV-1 and human pro-tein, but the protein name is different. Bicluster 1 consists of env_gp120, env_gp160, env_gp41, Nef, Vpr, CASP3, MAPK1, MAPK3, TNF, and HLA-C. Bicluster 2 consists of env_gp120, env_gp160, Tat, Vpr, Vpu, CASP3, FOS, JUN, CD4, and TNF. Bicluster 3 consists of env_gp120, env_gp41 Pr55(Gag), Matrix, Nef, IFNG, CD63, ICAM1, IL6, and TNF. Bicluster 4 consists of env_gp160, Pr55(Gag), Matrix, Nef, Tat, ICAM1, ACTA1, IL6, ACTB, and IL6. Bicluster 5 consists of env_gp160, Nef, Tat, Vpr, Vpu, CASP3, FOS, JUN, CD4, and TNF.

Heatmap from biclusters 1-5 LCM-MBC algorithm is red, indicating that bicluster 1-5 clustering the interactions with value 1 (Figure 7). Based on the results of the heatmap, the LCM-MBC bicluster is a bicluster of HIV-1 protein and human proteins that interact with

Table 9: Biclustering PLCM-MBC algorithm HV positive

Protein		Bicluster
HIV-1	Human	
5	5	1, 2, 3, 4, 5, 6, 7, 8
5	6	9, 10, 11, 12, 13, 14, 15, 16, 17
5	7	18, 19
5	8	20, 21
5	9	22, 23, 24, 25
5	10	26
5	12	27
5	13	28
5	22	29
6	6	30
6	8	31

each other. The validation of the LCM-MBC bicluster results in Table 10 shows that all bicluster have biological interpretations. Bicluster 2 and 5 have the same Biological process, cellular component, and molecular process because HIV-1 protein in the two biclusters is the same.

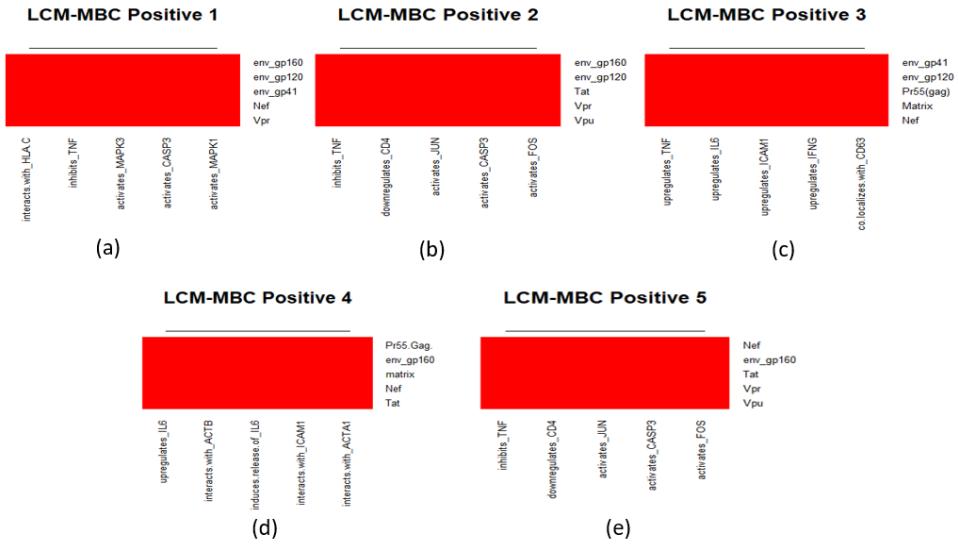

Figure 7: Heatmap 5 biclusters HV positive LCM-MBC; (a) Bicluster 1; (b) Bicluster 2; (c) Bicluster 3; (d) Bicluster 4; (e) Bicluster 5.

The heatmap from BiMax (Figure 5) and LCM-MBC (Figure 7) show that BiMax and LCM-MBC classify interactions with a value of 1 to a bicluster while POLS (Figure 6) still classifies interactions with values of 0 and 1 to a bicluster. In another way, BiMax and LCM-MBC 100% correctly classify the interactions. BiMax, POLS, and LCM-MBC are biclustering algorithms based on graph theory, bicluster from these algorithms are a biclique subgraphs. Heatmap results show that BiMax and LCM-MBC biclusters are bicliques because each edge has a value of 1 (interacting) while the POLS still has an edge that has a value of 0 (does not

Table 10: GO terms LCM-MBC algorithm HV positive

Bicluster	GO Term Biological Process	GO Term Cellular Component	GO Term Molecular Function
1	response to cytokine	membrane raft	phosphotyrosine binding
2	response to bacterium	membrane raft	macromolecular complex binding
3	positive regulation of nitric oxide biosynthetic process	cell surface	cytokine activity
4	response to metal ion	extracellular space	structural constituent of cytoskeleton
5	response to bacterium	membrane raft	macromolecular complex binding

interact). So the BiMax and LCM-MBC bicluster is biclique, while the POLS is not biclique. The POLS algorithm still has a weakness in classifying binary data, so it is necessary to use the upgrade of POLS algorithm. POLS development has been carried out by Wang *et al.* (2006) by adding several parameters.

The validation for each bicluster with GO terms (Table 6 for BiMax, Table 8 for POLS, and Table 10 LCM-MBC.) Table 6 and Table 10 explains that each BiMax and LCM-MBC bicluster has a biological interpretation, while in bicluster 1 Table 8 the POLS do not have a cell component and a molecular process, because no data or research explains the interactions of proteins found in bicluster that is. Biological explanations for each cluster can be seen through the category number of each GO term. In the 30 results of the GO terms BiMax, there is no same biological function. From 31 Go terms of the LCM-MBC bicluster results, there were four pairs of bicluster which had the same biological function. The GO terms of the POLS results, from 13 bicluster results there are still two biclusters that have no biological function. The effectiveness of bicluster results in biology can be calculated by dividing the bicluster with biological functions that are not the same as all bicluster results and multiplied by 100%. So we can conclude that BiMax, POLS, and LCM-MBC can effectively classify data by 100%, 84.62%, and 74.19% respectively.

4.2. HV negative

BiMax found eight biclusters in dataset HV negative (see Table 11). The maximum number of human protein is seven proteins, founded in three biclusters and the minimum number of human protein is five proteins founded in two biclusters.

Information for five bicluster in Table 11 are, first bicluster consists of six HIV-1 proteins (env_gp120, env_gp160, Pr55(Gag), capsid, matrix, and Nef) and five human proteins (CD63, LAMP1, HLA.A, ICAM1, and HLA-B). Second bicluster consists of five HIV-1 proteins (env_gp120, env_gp160, Pr55(Gag), capsid, and Nef) and seven human proteins (CD63, LAMP1, HLA-A, ICAM1, HLA-B, CD3D, and CD3E). Third bicluster consists of five HIV-1 proteins (env_gp120, Pr55(Gag), capsid, Nef, and Tat) and five human proteins (ICAM1, EIF2AK2, CD3D, CD3E, and CD3G). Fourth bicluster consists of sic HIV-1 proteins (env_gp120, env_gp160, Pr55(Gag), matrix, RT, and Tat) and six human proteins (ACTC1, ACTA1, ACTA2, ACTG1, ACTG2, and ICAM1). Fifth bicluster consists of five HIV-1 proteins (env_gp120, env_gp160, Pr55(Gag), matrix, and Tat) and seven human proteins (ACTC1, ACTA1, ACTA2, ACTB, ACTG1, ACTG2, and ICAM1).

Table 11: Biclustering BiMax HV negative

Protein		Bicluster
HIV-1	Human	
5	5	3
5	6	7, 8
5	7	2, 5, 6
6	52	1
6	6	4

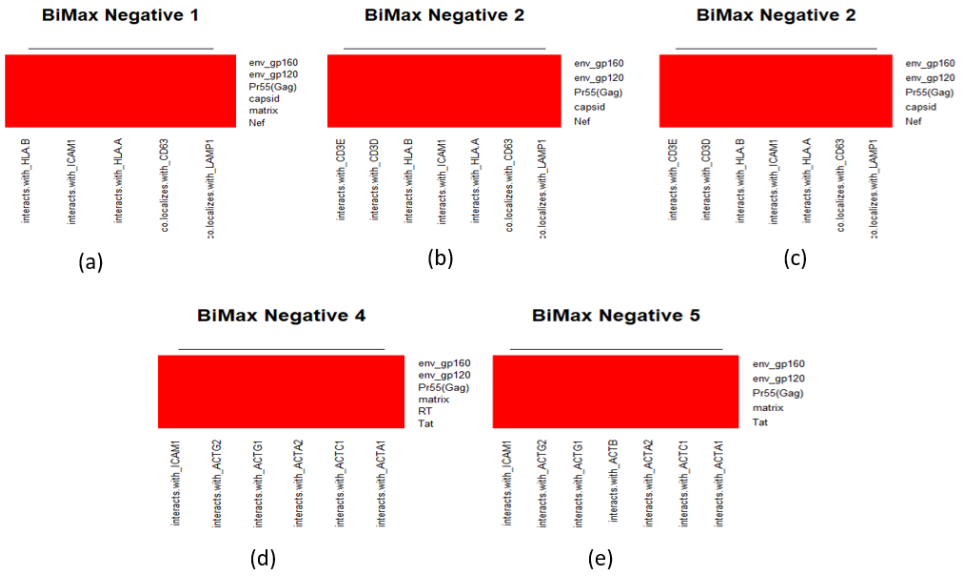

Figure 8: Heatmap 5 biclusters HV negative BiMax; (a) Bicluster 1; (b) Bicluster 2; (c) Bicluster 3; (d) Bicluster 4; (e) Bicluster 5.

Based on 8, heatmap of 5 biclusters BiMax shows a uniform color on one bicluster means BiMax classifies the interaction correctly. The red color in the heatmap implies that there is an interaction so that it can be concluded that all of the BiMax bicluster results are bicluster from the proteins that interact. The interpretation from the bicluster of BiMax in Table 12 shows that several biclusters have similarities. Bicluster 1 and 2 have the same molecular function. Bicluster 2 and 3 are the same for cellular components. Bicluster 4 and 5 have similarities in the biological process and cellular component. Based on the results of bicluster and validation, the more similar members of a bicluster with other biclusters, the more biologically the interpretation will be.

POLS algorithm found ten biclusters in dataset HV Negative, the result almost the same as dataset HV Positive. There are five biclusters consist of five proteins HIV-1 and human and one bicluster consist of seven proteins HIV-1 and human. The result for all bicluster sees Table 13.

Proteins in bicluster 1 are env_gp120, env_gp160, env_gp41, Pr55(gag), Capsid, Matrix, IFITM1, IFITM2, IFITM3, EIF2AK2, IL7, and HLA-B. Proteins in bicluster 2 are env_gp120, env_gp160, env_gp41, Pr55(gag), Capsid, Matrix, Nucleocapsid, FN1, SFTPD, CD4, TLR2, ACHE, CD59, and CD63. Proteins in bicluster 3 are env_gp120, env_gp160, env_gp41, Pr55(gag), Capsid, CD63, ADAR, LGALS3Bp, HDAC6, and IFNA1. Proteins in bicluster 4 are env_gp120, env_gp160, env_gp41, Pr55(gag), Capsid, Matrix, ACTA1, ACTA2, ACTB, ACTG1, ACTG2, and ANXA2. Pro-teins in bicluster 5 are env_gp120, env_gp160, env_gp41,

Table 12: GO terms bicluster BiMax HV negative

Bicluster	GO Term Biological Process	GO Term Cellular Component	GO Term Molecular Function
1	regulation of leukocyte mediated cytotoxicity	TAP binding	cell surface
2	regulation of leukocyt mediated cytotoxicity	T cell receptor binding	cell surface
3	regulation of immune system process	T cell receptor binding	alpha-beta T cell receptor complex
4	mesenchyme migration	ATP binding	filopodium
5	mesenchyme migration	ATP binding	blood microparticle

Table 13: Biclustering POLS algorithm HV negative

Protein		Bicluster
HIV-1	Human	
5	5	3, 6, 7, 8, 10
6	6	1, 4, 5, 9
7	7	2

Pr55(gag), Capsid, Matrix, CD3D, CD3E, CD3G, CD4, CD28, and TXNDC5.

The heatmap of the HV Negative dataset is almost the same as the HV Positive dataset. In a bicluster, there are still two interaction values, 0 and 1, green values 0 and red values 1 (See Figure 9). Then the interpretation of each cluster through GO terms can be seen in Table 14. Bicluster 1 POLS for HV Negative data is the same as HV Positive data, so bicluster 1 does not have cellular components and molecular processes.

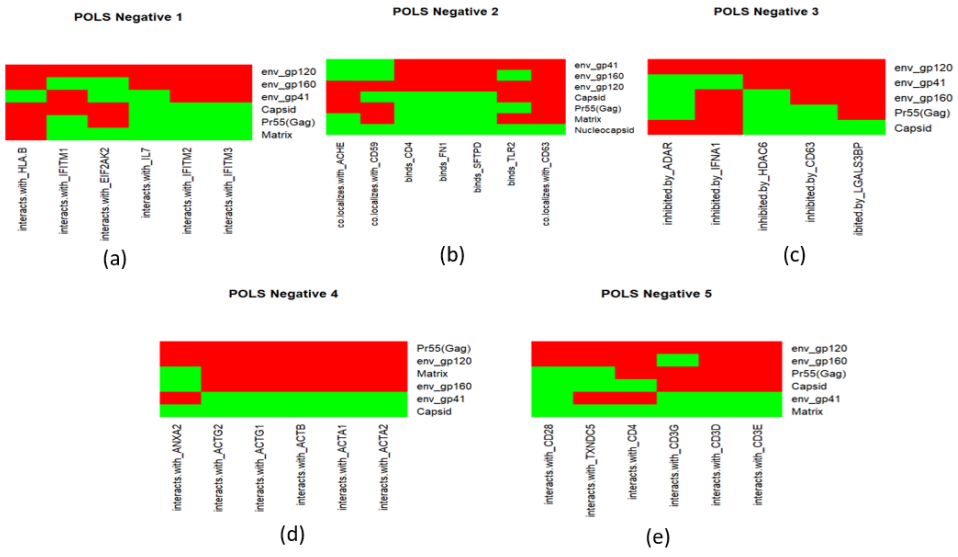

Figure 9: Heatmap 5 biclusters HV negative POLS; (a) Bicluster 1; (b) Bicluster 2; (c) Bicluster 3; (d) Bicluster 4; (e) Bicluster 5.

LCM-MBC algorithm found 14 biclusters with sized of the smallest bicluster is five HIV-1

Table 14: GO terms POLS algorithm HV negative

Bicluster	GO Term Biological Process	GO Term Cellular Component	GO Term Molecular Function
1	response to interferon-alpha	-	-
2	cell adhesion	collagen binding	cell surface
3	cellular response to organic substance	hydrolase activity, acting on carbon-nitrogen (but	platelet dense granule
4	mesenchyme migration	ATP binding	extracellular space
5	T cell costimulation	signaling receptor activity	T cell receptor complex

proteins and five human proteins and found on four biclusters. The biggest bicluster consists of six HIV-1 proteins and six human proteins and only one bicluster (see Table 15).

Table 15: Biclustering LCM-MBC algorithm HV negative

Protein		Bicluster
HIV-1	Human	
5	5	1, 2, 3, 4
5	6	5, 6, 7, 8, 9, 10
5	7	11, 12, 13
6	6	14

HIV-1 and human proteins found in bicluster 1-5 are bicluster 1 consists of env_gp120, env_gp160, Capsid, Matrix, Nef, CD63, LAMP1, HLA-A, ICAM1, and HLA-B. Bicluster 2 consists of env_gp120, Pr55(Gag), Capsid, Matrix, Nef, CD63, LAMP1, HLA-A, ICAM1, and HLA-B. Bicluster 3 consists of env_gp120, Pr55(Gag), Capsid, Nef, Tat, ICAM1, EIF2AK2, CD3D, CD3E, and CD3G. Bicluster 4 consists of env_gp120, env_gp160, Pr55(Gag), Capsid, Matrix, CD63, LAMP1, HLA-A, ICAM1, and HLA-B. Bicluster 5 consists of env_gp160, Pr55(Gag) Capsid, Matrix, Nef, CD63, LAMP1, HLA-A, ICAM1, HLA-B, and CD81.

The LCM-MBC heatmap bicluster can be seen in Figure 10. From Figure 10, all bicluster have the same color red, meaning LCM-MBC classifies proteins that interact in one bicluster. The LCM-MBC GO term can be seen in Table 16. Bicluster 1, 2, and 3 have a biological process, cellular component, and molecular processes that are the same, namely regulation of leukocyte mediated cytotoxicity, cell surface, and virus receptor activity. If we see from the results of a bicluster, then HIV-1 proteins in bicluster 1, 2, and 3 are similar, and the human protein only has one different in each bicluster which caused 3 LCM-MBC biclusters to have the same interpretation.

The results of bicluster HV Negative dataset using BiMax, POLS, and LCM-MBC are not much different from the HV Postive dataset. BiMax and LCM-MBC classify HIV-1 and human proteins into one bicluster seen in Figure 8 and Figure 10 with heatmaps have one color, while POLS still has two colors in one bicluster see in Figure 9. Based on the results of the heatmap, bicluster BiMax and LCM MBC is a biclique with vertices is a set of HIV-1 and human proteins, the edge is an interaction set that occurs between HIV-1 and humans. The POLS bicluster results are not biclique, because there are nodes that do not interact.

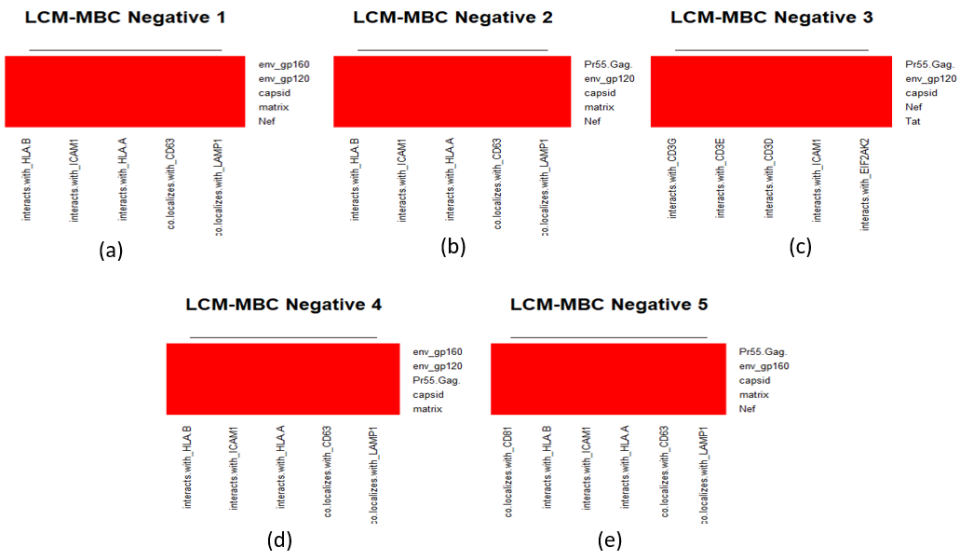

Figure 10: Heatmap 5 biclusters HV negative LCM-MBC; (a) Bicluster 1; (b) Bicluster 2; (c) Bicluster 3; (d) Bicluster 4; (e) Bicluster 5.

GO term results in Table 12, Table 14, and Table 16 for BiMax, POLS, and LCM-MBC respectively. Bicluster 1 HV POLS Negative dataset is the same as Bicluster 1 POLS HV Positive dataset, so it does not have cellular component and molecular function. BiMax has no same biological function for every bicluster results. LCM-MBC has three pairs Bicluster which has similarities. Using the same method as HV positive dataset, BiMax can classify data effectively by 100%, POLS 90%, and LCM-MBC 78.57%.

5. Conclusion

This paper has presented a comparison of several biclustering algorithms such as BiMax, POLS, and LCM-MBC to classify protein-protein interaction between HIV-1 and human. The results of the POLS algorithm show that there are still two interaction values in one bicluster, which is 0 (not interacting) and 1 (interacting). The LCM-MBC algorithm can classify the interaction data which only has 1 value (interact) in one bicluster. Every bicluster BiMax algorithm has a value of 1 (interacting). From the three bicluster results, it can be concluded that LCM-MBC and BiMax can correctly classify the protein-protein interaction data in binary form. The results of the LCM-MBC and BiMax GO terms show that the LCM-MBC bicluster is validated that several biclusters have the same function (Biological Process, Cellular Componenet, and Molecular) whereas, in the Bimax bicluster no bicluster has the same function. If some bicluster has the same biological properties, the bicluster results can be said to be precise. So from our results, we can conclude that among three biclustering algorithm, BiMax is the best algorithm for classifying binary protein-protein interaction data because BiMax can correctly classify protein-protein interaction data in binary form in mathematical and biological ways. Overall, the biclustering algorithm is also able to clustering data other than gene expression data. Future work, we can use the modified POLS algorithm, then calculate the computation time of each algorithm and a more detailed explanation of the functions of each bicluster.

Acknowledgment

We want to express our gratitude for the QQ 2019 Grant from Directorate of Research and Human Engagement Universitas Indonesia in supporting this research.

Table 16: GO terms LCM-MBC algorithm HV negative

Bicluster	GO Term Biological Process	GO Term Cellular Component	GO Term Molecular Function
1	Regulation of leukocyte mediated cytotoxicity	cell surface	Virus receptor activity
2	Regulation of leukocyte mediated cytotoxicity mediated cytotoxicity	cell surface	Virus receptor activity
3	regulation of immune system process	alpha-beta T cell receptor complex	T cell receptor binding
4	Regulation of leukocyte mediated cytotoxicity	cell surface	Virus receptor activity
5	Regulation of leukocyte mediated cytotoxicity mediated cytotoxicity	integral component of plasma membrane plasma membrane	protein complex binding

References

Ardaneswari G, Bustamam A, Sarwinda D (2017a). "Implementation of Plaid Model Biclustering Method on Microarray of Carcinoma and Adenoma Tumor Gene Expression Data." volume 893, p. 012046. doi:10.1088/1742-6596/893/1/012046.

Ardaneswari G, Bustamam A, Siswantining T (2017b). "Implementation of Parallel k-means Algorithm for Two-phase Method Biclustering in Carcinoma Tumor Gene Expression Data." volume 1825, p. 020004. doi:10.1063/1.4978973.

Bustamam A, Formalidin S, Siswantining T (2018a). "Clustering and Analyzing Microarray Data of Lymphoma Using Singular Value Decomposition (SVD) and Hybrid Clustering." volume 2023, p. 020220. doi:10.1063/1.5064217.

Bustamam A, Puspa D, Siswantining T (2018b). "Implementation of Co-similarity Measure on Microarray Data of Lymphoma Using K-means Partition Algorithm." volume 2023, p. 020222. doi:10.1063/1.5064219.

Bustamam A, Zubedi F, Siswantining T (2018c). "Implementation χâĂŞsim co-similarity and Agglomerative Hierarchical to Cluster Gene Expression Data of Lymphoma by Gene and Condition." volume 2023, p. 020221. doi:10.1063/1.5064218.

Cheng Y, Chruch G (2000). "Biclustering of Expression Data." pp. 93–103.

Fu W, Sanders-Beer S, Katz K, Maglott D, Pruitt K, Ptak R (2008). "Human Immunodeficiency Virus Type 1, Human Protein Interaction Database at NCBI." *Nucleic Acids Research*, **37**, 417–422. doi:10.1093/nar/gkn708.

Kaiser S, Santamaria R, Khamiakova T, Sill M, Theron R, Quintales L, Leisch F, Troyer ED (2018). "biclust: BiCluster Algorithms. R package version 2.0.1." URL https://CRAN.R-project.org/package=biclust.

Lestari D, I S Musti M, Bustamam A (2018). "Sequence-based Prediction of Protein-protein Interactions Using Ensemble Based Classifier Combined with Global Encoding in HIV (Human Immunodeficiency Virus)." volume 2023, p. 020230. doi:10.1063/1.5064227.

Liu G, Li H, Wong L, Li J (2007). "Maximal Biclique Subgraphs and Closed Pattern Pairs of the Adjacency Matrix: A One-to-One Correspondence and Mining Algorithms." volume 19, pp. 1625–1637. doi:10.1109/TKDE.2007.190660.

Mukhopadhyay A, Maulik U, Bandyopadhyay S (2010). "On Biclustering of Gene Expression Data." *Current Bioinformatics*, **5**(3), 204–216. doi:10.2174/157489310792006701.

Mukhopadhyay A, Ray S, Maulik U (2014). "Incorporating the Type and Direction Information in Predicting Novel Regulatory Interactions between HIV-1 and Human Proteins Using a Biclustering Approach." *BMC Bioinformatics*, pp. 15–26. doi:10.1186/1471-2105-15-26.

Prelic A, Bleuler S, Zimmerman P, Wille A, Buhlamn P, Gruissem W, Hennig L, Thiele L, Ziztler E (2006). "A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data." *Bioinformatics*, **22**(9), 1122–1129. doi:10.1093/bioinformatics/btl060.

Santamaria R, Theron R, Quintales L (2008). "A Visual Analytics Approach for Understanding Biclustering Results from Microarray Data." *BMC Bioinformatics*, **9**(47). doi:10.1186/1471-2105-9-247.

Singh A, Nagrare A, Srikanth P, Kumar D, Dwith C (2011). "A Comparison of Biclustering with Clustering Algorithms." volume 2023, pp. 1–4. doi:10.1109/PACCS.2011.5990194.

Trkola A (2004). "HIVâĂŞhost Interactions: Vital to the Virus and Key to Its Inhibition." *Current Opinion in Microbiology*, **7**(4), 407 – 411. ISSN 1369-5274. doi:https://doi.org/10.1016/j.mib.2004.06.002. URL http://www.sciencedirect.com/science/article/pii/S1369527404000682.

Wang Y, Cai S, M Y (2006). "New Heuristic Approaches for Maximum Balanced Biclique Problem." *Information Sciences*, **432**, 362–375. doi:10.1016/j.ins.2017.12.012.

Yang L, Shen Y, Yuan X, Zhang J, Wei J (2017). "Analysis of Breast Cancer Subtypes by AP-ISA Biclustering." *BMC Bioinformatics*, **18**(481). doi:10.1186/s12859-017-1926-z.

Affiliation:

Alhadi Bustamam
Department Mathematics
Faculty of Mathematics and Natural Science
Universitas Indonesia
Depok, West Java, 16424
E-mail: alhadi@sci.ui.ac.id
URL: http://staff.ui.ac.id/alhadi

http://www.ajs.or.at/
http://www.osg.or.at/
Submitted: 2019-08-21
Accepted: 2019-11-28