A MODEL FOR PREDICTING HIV-1 – HUMAN PROTEIN INTERACTIONS USING DATA MINING TECHNIQUES

Nour Moustafa[1], Ahmed Sharaf Eldin[1] and Samar Kamal Kassim[2]

[1]Faculty of Computers & Information, Helwan University, Cairo, Egypt
[2]Faculty of Medicine, Ain Shams University, Cairo, Egypt

ABSTRACT

Discovering Protein-Protein Interactions (PPI) is an area of active research in computational biology. Identifying interactions among proteins was shown to help discover new drugs and prevent many diseases. The interactions between HIV-1 proteins and Human proteins is a particular PPI problem which study might lead to the discovery of important interactions responsible for Acquired Immune Deficiency Syndrome (AIDS) . This paper presents an algorithm that applies the data mining for extracting hierarchical bi-clusters and minimal covers of PPI without losing information. The proposed algorithm is based on the frequent closed item sets framework to efficiently generate a hierarchy of conceptual clusters and non-redundant sets of association rules with supporting object lists to integrate additional information about proteins. Experimental results show that the proposed algorithm is more accurate than Apriori Algorithm and predicted new interactions didn't discover by Tastan technique.

KEYWORDS

Protein-Protein Interactions (PPI), conceptual bi-cluster, Acquired Immune Deficiency Syndrome (AIDS),

1. INTRODUCTION

1.1 Motivation

Human immunodeficiency virus-1 (HIV-1) is the main mandatory of acquired immune deficiency syndrome (AIDS) [1] and continues to cause health damage awesome lead to death [2]. The number of AIDS-related deaths was about 2.8 million in 2010 alone [3]; an estimated about 34 million people worldwide are infected with this disease [4]. Virus spreads heavily on host cells in order not to able to work successfully, at the same time avoids the immune system to carry out its functions. Protein-protein interactions (PPIs) between HIV-1 and its host are vital at every stage in the life cycle of the virus so that helps to discover medications prevent the virus from spreading in the body's cells [5]. Here, this paper proposes an algorithm that employs data integration methods on these reported interactions in annexation with a variety of different biological information sources to predict new PPIs between HIV-1 and human proteins.

1.2 Related Work

Acquired Immune Deficiency Syndrome (AIDS) is the last phase of HIV infection. At this phase, the human immune system fails to resist the virus and spread in the body's cells, and this eventually cause death. HIV is a member of the retrovirus family (lentivirus) which infects effective cells in the human immune- system. This infection is due to the interaction between proteins of both the virus and the human host in the human cells. Predicting such interactions is

an important goal of PPI research. Study and analysis of known interactions in order to discover new interactions provide information useful and help to find new drugs and explore the causes and mechanisms of this type of viral disease that causes of death. Various approaches for predicting interactions have been studied in the literature. These approaches are based on Bayesian networks [6], random forest classifiers [7], mixture of feature expert classifiers [8].Most of them have been used to find interactions within a single creature, like yeast or human (intra-species interactions). Recently, two approaches have been proposed to predict the set of interactions between HIV-1 and human host cellular proteins [9]. In particular, the proposed algorithm produces a supervised Learning framework that integrates diversified biological information to predict inter-species interactions. However, this approach solves the classification problem using association rules mining which needs positive samples of PPIs only. This paper will present the proposed algorithm to integrate bi-clustering and association rule mining to generate the frequent closed itemsets and, generate minimal non-redundant covers of association rules [7] [10].

2. METHODS

2.1 HIV-1–Human Protein Interaction Dataset

The prepared PPIs are specifically annotated according to the nature of the interactions including about 26 interaction types such as "interacts with", "binds to", "upregulates" ,"inactivates" and ‘’inhibits’’[20]. The types of interactions can be classified into direct physical interactions (e.g., "binds to"), that consist of 32% of the interactions and indirect interactions (e.g., "upregulates") that consist of 68% of the interactions .Both direct and indirect interactions are considered in this study. The indirect interactions also provide useful information of the characteristics of the interacting proteins. Also, the natures of interactions are not considered separately [20].

2.1.1 Preparing the matrix Dataset

The dataset of HIV-1 and Human proteins [21] is used to test the proposed algorithm. This dataset is a matrix of 19 columns representing the HIV-1 proteins and 1432 rows representing the human proteins. Each cell of matrix contains a 1 if this is a positive interaction between the pairs of proteins and 0 if there is a negative interaction between them.

2.1.2 Preparing the dataset in a relations database

Three tables are constructed to prepare the positive interaction between the HIV-1 and Human proteins {vprotein (id_vp , vp) , hprotein (id_hp , h_protein) , trans (id_hp , id_vp , bool)}. The total number of interactions that generated from the positive interactions is 2765 interactions.

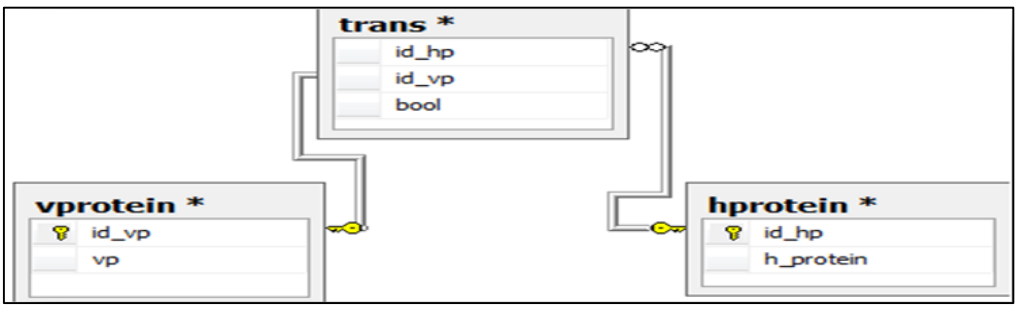

Figure 1. the relationship between HIV-1 and Human proteins

2.1.3 Preparing the dataset to apply the proposed algorithm

The proposed algorithm needs to group each viral protein with the human proteins that interact with it.

2.2 Association Rule Mining (ARM)

The use of association rule mining (ARM) to find frequent itemsets, associations, and correlations between the sets of items or objects in transaction databases, relational databases. There are applications which apply the ARM such as Basket data analysis and cross-marketing. Association analysis is the discovery of rules showing attribute–value associations that occur frequently [11]. There are two important basic measures for association rules, support and confidence. Let I={i1,i2,….,n} be a set of n items and Xmi be T={(t1,X1),(t2,X2),….(tm,Xm)} a set of m transactions where ti and Xi ,i=1,2,…m are the transactions and the associated itemset . The transactions of itemset X in T as follows to measure the Rules:

- The support of an itemset X in T is

Support(X, T) is the number of tuples containing both X and T / the total number of tuples =P(X U T).

- The confidence of an item X in T is

Confidence(X -> Y, T) is the number of tuples containing both X and T / the number of tuples containing X=P(X U T) / P (T).

The problem of ARM Finds rules that have support and confidence greater that user-specified minimum support and minimum confidence.

- Support(X, T) >= min support.
- Confidence(X -> Y, T) >= min confidence.

The ARM method consists of the following two steps [12], [13]:
1. Find all frequent itemsets in the transactions.
2. Generate strong association rules from the frequent itemsets.
The number of itemsets grows with the number of items. Apriori algorithm is used for generating the frequent itemsets [14, 15]. This is based on the method of downward closure property which declares that if even one subset of an itemset is not frequent, then cannot be frequent. It starts from all itemsets of size one, and proceeds in a recursive pattern. If any itemset is not frequent then that branch of the tree is pruned, so any possible superset of can never be frequent. Although Apriori is a common algorithm, it is complex computations and become its value is very low and its items are very large, therefore takes a large execution time. Moreover, it is necessary to ignore the redundant information in the frequent itemsets. In this context, the meaning of closed itemsets [16, 17] is necessary. An itemset is called the closest itemset if none of its adequate supersets have the same support value. It is important to find the closest itemsets in order to avoid any redundant item. Further, frequent closed itemsets are compress impersonation of frequent itemsets without loss of any information. There is no saving in Apriori algorithm to directly search for the closest itemsets. In this paper, the proposed algorithm that utilizes a biclustering method [18] for identifying the biclustering and the frequent closed itemsets directly from the generated bicluster. The PPI dataset considered in this paper is very sparse (only 2492 interacting pairs among 27208 possible protein pairs: (10.91%) and the bi-clusters is 258 from the PPI dataset. Therefore it is required to use low value for generating frequent itemsets. Moreover when the PPI matrix is arranged in a form in which the human

proteins are considered as the items, the computational time required to execute the Apriori method is large since the number of human proteins is 1432 in the dataset considered here. This is because the numbers of possible itemsets need to explore is in the order of Furthermore, it is also necessary to find the frequent closed itemsets only to avoid redundant itemsets. Therefore, it is difficult to use Apriori algorithm for this dataset. Hence, this paper have proposed a new association rule mining approach based on a biclustering method that efficiently mines the frequent closed itemsets from sparse data sets , having low value and large number of items. The following section describes the proposed algorithm.

3. THE PROPOSED ALGORITHM

The proposed algorithm is a two-phase process: (1) pre-processing the dataset (2) finding the hierarchical conceptual bi-clusters and extracting association rules (i.e., frequent closest itemsets).

3.1 Pre-processing the dataset

The input is a dataset represented as a data matrix in which rows are called objects and columns are called attributes. Each distinct value of an attribute constitutes an item. This algorithm performs one scan of the input dataset to generate the conceptual bi-clusters and the frequent closed itemsets. The example in table 1 consists of 5 objects and 4 attributes .each cell represents the interaction among a viral and a human protein that is interacted when the value is 1 and non-interacted when the value is 0.

Table 1. Example dataset.

OID	vp1	vp2	vp3	vp4
hp1	0	1	1	0
hp2	1	1	1	1
hp3	1	0	0	1
hp4	0	1	1	1
hp5	1	1	1	1

From Table1: prepares a group of each vps (virus proteins) with each positive (1) of hps (human proteins) and introduced in Table2.

Table 2. Group vps with each hps that is positive interacted.

vp1	vp2	vp3	vp4
hp2	hp1	hp1	hp2
hp3	hp2	hp2	hp3
hp5	hp4	hp4	hp4
	hp5	hp5	hp5

3.2 finding the hierarchical conceptual bi-clusters

Pseudo code of hierarchical conceptual bi-clusters and association rules:

Input: Table2, min-support value, min-confidence value.
Output: Bi-clusters (BIC), Association rules (AR).

1. Sort the object lists (human proteins) of each an item (virus protein) ascending.
2. Begin with the lowest number of object list.

3. Get each 2-objectset and compare them with all the other object lists. If they found, get all the itemsets that contain them.
4. Read each 2-objectset once from the dataset.
5. Design list that contains each itemsets with its each object lists.
6. To remove the redundant itemsets, merge all similar itemsets with their object lists.
7. To remove the redundant objectlists, merge all similar objectlists with their itemsets.
8. From previous two steps (6 and 7), construct BIC.
9. If support(R)>=min-support AND support > 1 AND Confidence>=min-confidence, map patterns in BIC into AR.
10. Decompose the AR to pairs of (itemset, Objectlist) such that each pair satisfies the previous conditions in step 9.

The lists after execute the algorithm:

Table 3. Itemsets and object lists (step 5).

Itemset	Objectlist
vp1,vp4	hp2,hp3
vp1,vp2,vp3,vp4	hp2,hp5
vp1,vp4	hp3,hp5
vp2,vp3	hp1,hp2
vp2,vp3,vp4	hp2,hp4
vp2,vp3,vp4	hp4,hp5
vp2,vp3	hp1,hp4
vp2,vp3	hp1,hp5
vp4	hp3,hp5

Tables 4. Merge the similar itemsets.

Itemset	Objectlist
vp1,vp4	hp2,hp3,hp5
vp1,vp2,vp3,vp4	hp2,hp5
vp2,vp3	hp1,hp2,hp4
vp2,vp3,vp4	hp4,hp5
vp4	hp3,hp4

4. EXPERIMENTAL RESULTS

The proposed algorithm was implemented by developing an application in Dot Net (C#). Experiments were conducted on a PC with an Intel(R) Core (TM) i3-2370M processor at 2.40 GHZ and 4 GB of RAM, running under the 32 bits Windows 7 Professional Edition operating system. The PPI dataset used for performance experiments was constructed from the HIV-1-Human Protein Protein Interaction Database of the NIAID [19]. This dataset is a matrix of 19 columns corresponding to the different HIV-1 proteins and 1432 rows corresponding to the human proteins. Each cell of the matrix contains a 1 if there is a positive interaction between the corresponding pair of proteins and a 0 if no interaction is reported. To assess the scalability of the proposed algorithm when each viral protein interacts with all the positive human protein and prevent the redundant of the same viral protein with the same positive human protein leads to appear only 2492 interacting pairs among 27208 possible protein pairs that appeared in Tastan technique [9].

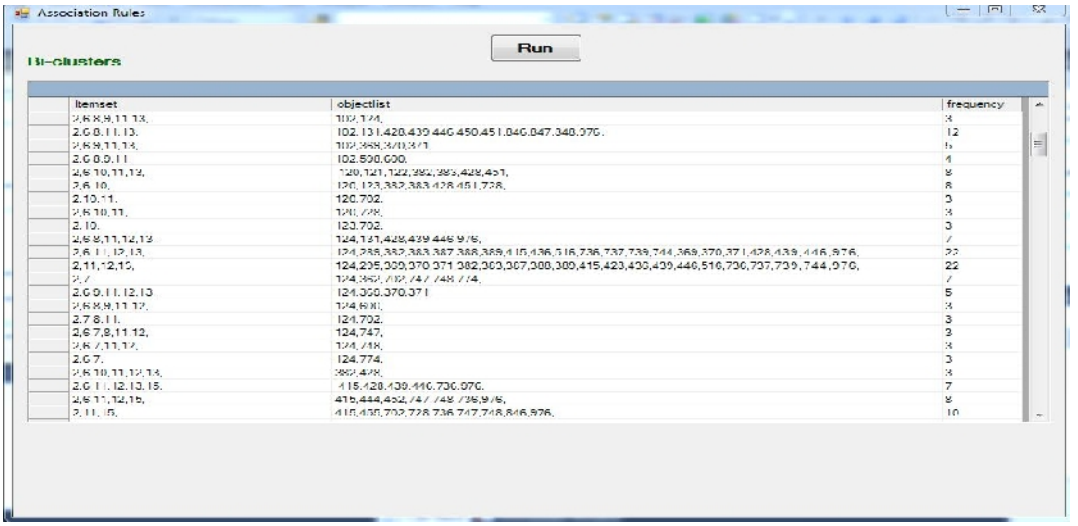

Figure 2. The itemset and the clusters of objectlist with their supports

4.1 Discussion

It is important to compare the results of the proposed algorithm to the results obtained by Tastan technique [9], which are the most comprehensive HIV-Human PPI results available to date. This comparison focus on the results generated under the consideration of minsupport = 0.1% and minconfidence = 0.1%, which are the lowest threshold values appeared and thus contain maximal information. The interaction between the pairs of the viral and human protein is 2492 interaction predicted but in Tastan technique [9] is 3372, where Tastan technique [9] used the random forest classifier that achieve a mean average precision (map) of 0.23 on this problem, the meaning that about 23 % of the predicted pairs should be expected to be true positive; So 880 predicted pairs in Tastan technique [9] are inaccurate. This represents 26% of their predicted pairs.

4.1.1 The results at minsupport = 0.1% and minconfidence = 0.1%

The predicted interaction when the minsup=0.1, minconf=0.1 is 327 predicted pairs.

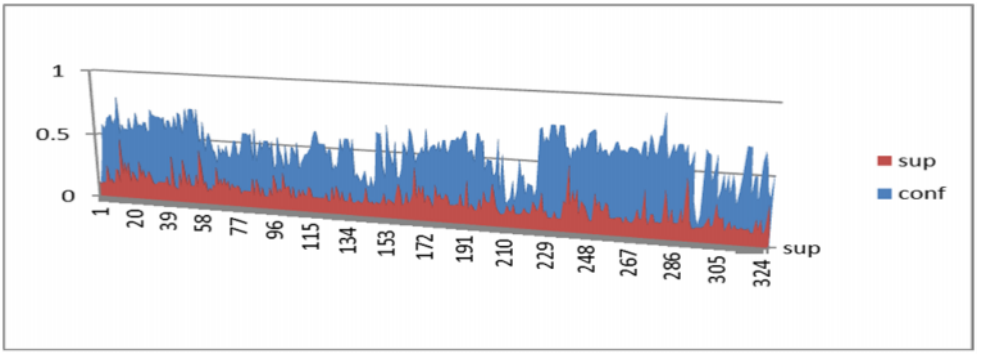

Figure 3. The representative of the support and confidence of the predicted pairs

Figure 4. the highest pairs confidence

4.1.2 The results at minsupport = .16% and minconfidence = 0.7%

The predicted interactions when the minsup=0.16, minconf=0.7 are 58 predicted pairs.

Table 5. The most predicted pairs between the viral and human proteins and compared the confidence score by Tastan technique.

viral protein	human protein	Confidence score provided by proposed method	Interaction score provided in Tastan technique
env_gp120	NFKBIA	0.72	0.00
env_gp120	ICAM1	0.77	2.25
env_gp120	IFNA1	0.79	0.00
env_gp120	IL1B	0.74	0.00
env_gp120	IL4	0.71	1.56
env_gp120	IL10	0.71	0.00
env_gp120	IL12A	0.70	0.40
env_gp120	CASP3	0.76	2.80
env_gp120	CD4	0.72	2.75
env_gp120	CD28	0.75	2.04
env_gp120	HLA-A	0.73	2.12
env_gp120	HLA-B	0.76	2.34
env_gp120	HLA-C	0.76	2.53
env_gp120	HLA-DRB1	0.70	2.43
env_gp120	CCL3	0.71	1.97
env_gp120	TNF	0.73	1.64
env_gp120	PPIA	0.85	0.71
env_gp120	PRKCB1	0.77	3.34
env_gp120	PRKCE	0.77	3.16
env_gp120	PRKCQ	0.77	2.55
Nef	IL4	0.71	0.00
Nef	IL12A	0.70	0.00
Nef	CD28	0.75	0.00
Nef	HLA-B	0.76	2.43
Nef	HLA-C	0.76	2.17
Nef	HLA-DRB1	0.70	1.85
Nef	CCL3	0.79	0.00
Nef	ACTB	0.70	2.41
Nef	PPIA	0.76	2.20
Tat	NFKB1	0.90	3.73
Tat	NFKBIA	0.79	2.67

Tat	ICAM1	0.77	1.84
Tat	IFNA1	0.76	1.67
Tat	IL1B	0.77	0.00
Tat	IL2	0.73	0.00
Tat	IL4	0.79	0.05
Tat	IL6	0.75	0.00
Tat	IL10	0.76	0.00
Tat	IL12A	0.79	0.00
Tat	CASP3	0.87	0.95
Tat	CD4	0.72	0.82
Tat	CD28	0.81	1.62
Tat	HLA-A	0.75	0.14
Tat	HLA-B	0.76	0.00
Tat	HLA-C	0.76	0.00
Tat	HLA-DRA	0.72	0.00
Tat	HLA-DRB1	0.76	0.19
Tat	CCL3	0.74	0.00
Tat	TNF	0.71	1.78
Tat	TP53	0.89	3.51
Tat	ACTB	0.77	0.25
Tat	ACTG1	0.75	0.00
Tat	PRKCA	0.72	3.48
Tat	PRKCB1	0.77	3.22
Tat	PRKCE	0.77	2.81
Tat	PRKCQ	0.77	3.87
Tat	MAPK3	0.71	3.37
Vpr	ACTG1	0.82	0.00

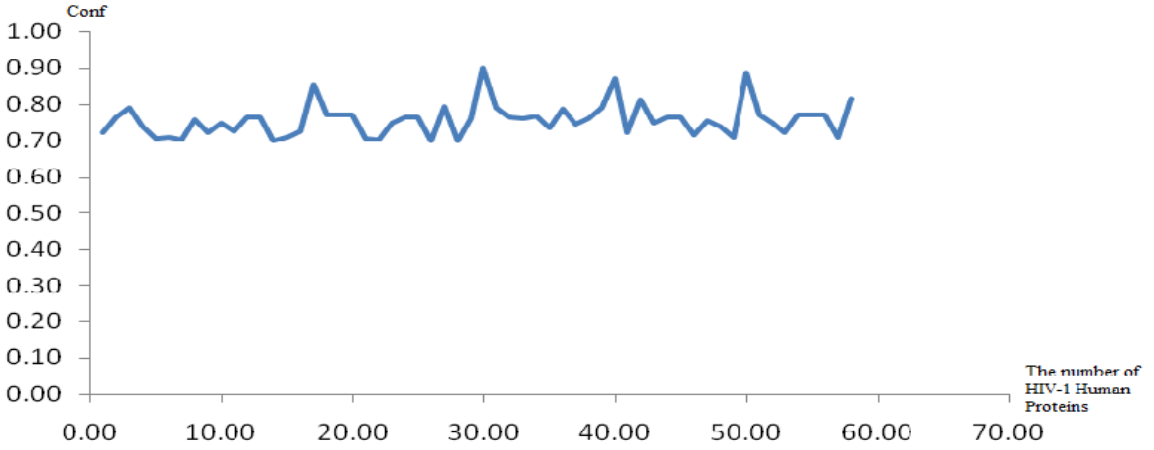

Figure 5. The confidence score provided by the proposed algorithm

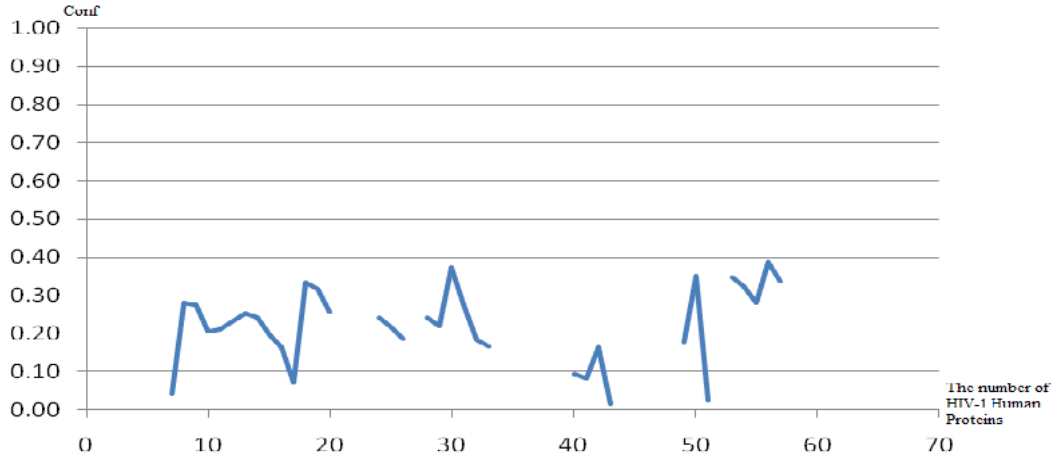

Figure 6. The interaction score provided in Tastan technique

The proposed algorithm discovered 58 predicted pairs is shown in Table 6 in the highest confidence (70%-80%) is shown in figure 5, but when the same results are compared by Tastan technique [9] is founded 40 pairs from 58 predicted pairs that have interaction score (10% - 40%) is shown in figure 6, and 18 pairs that are highlighted with red color in Table 5 are discovered by the proposed algorithm and didn't found in Tastan technique [9].

5. BIOLOGICAL RELEVANCE OF PREDICTED INTERACTIONS

To achieve the results the properties of human proteins that interact with each viral proteins must be check based on gene ontology based study. The results of these experiments are reported below.

5.1 Interactions with env_gp120

Figure 7. Gene Conversion Tool [1]

[1] http://david.abcc.ncifcrf.gov

In the case of minsup=0.1%, minconf=0.1%, the proposed algorithm found 56 human proteins interacted with the viral protein env_gp120. This paper looked into the biological relationships between these 56 human proteins by conducting a gene ontology (GO) in each of the three categories of GO based study to identify whether there are any significant non-redundant GO terms using GOTREMFINDER tool2 (see in figure 7). Then the use of the web server REVIGO3 [19] (see in figure 8) which takes as input a list of GO terms along with p-values and a GO-based semantic similarity measure, and generates an output of the set of non-redundant terms along with Dispensability values for each term. Lower value of Dispensability indicates lesser redundancy of the corresponding term [19]. The default parameter settings of REVIGO are used in the web server. Tables 6, 7 and 8 show the significant and non-redundant GO terms for certain thresholds of p-value and dispensability under Biological Process, Cellular Component and Molecular Function Categories. It is evident from the tables that the groups of human proteins that are predicted to interact with env_gp120 are biologically related and holds common biological activities. It is important to note from Table 6 that (21.67%) of these human proteins are included in biological process membrane organization. Also Table 7 shows that many of these proteins are component of coated membrane (20%). env_gp120 is established in the HIV envelope, which helps the virus to attach to and integrate with the target cell. In this process it needs to interact with many membrane proteins and the predicted human proteins seem to be largely related with membrane activities. From Table 6, it can also be found that some of these human proteins are included in the biological process of death (18.33%) and regulation of defense response to virus by virus (8.33%), which are important for raising antiviral immune response mechanism and that way lead to limit viral replication. When env_gp120 possibly interacts with these human proteins they affect their activities and as a result the immune response system may fail. This indicates that interaction of env_gp120 with these human proteins may lead to cell death. It is clear from the tables that groups of human proteins that are expected to interact with env_gp120 the biological relationships has in common with this virus.

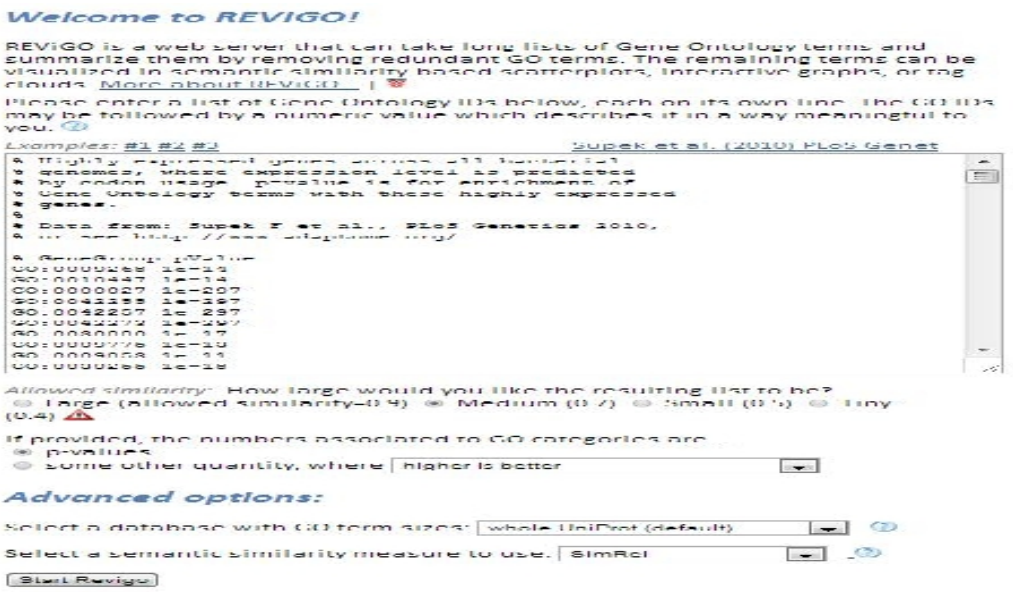

Figure 08. Revigo Tool [19]

[2] http://david.abcc.ncifcrf.gov

[3] http://revigo.irb.hr/

Table 6. Significant go terms (p-value 1E-03, dispensability 0.05) under biological process found in the human proteins that are predicted to interact with HIV-1 protein env_gp120.

No	Go-id	Term	P-value	% of proteins	Dispensability
1	GO:0007568	again	8.91E-05	10.00	0.00
2	GO:0016192	vesicle-mediated transport	8.11E-10	28.33	0.00
3	GO:0016265	death	6.87E-04	18.33	0.00
4	GO:0050690	organization of defense response to virus by virus	1.82E-08	8.33	0.00
5	GO:0030029	action filament-based process	4.63E-04	11.67	0.04
6	GO:0016044	membrane organization	3.82E-08	21.67	0.04
7	GO:006928	cell motion	7.03E-04	15	0.04

Table 7. Significant go terms (p-value 1E-08 , dispensability 0.05) under cellular component found in the human proteins that are predicted to interact with HIV-1 protein env_gp120.

No	Go-id	Term	P-value	% of proteins	Dispensability
1	GO:0030131	clarthrin adaptor complex	1.61E-19	20.00	0.00
2	GO:0031982	vesicle	3.54E-11	33.33	0.00
3	GO:0012505	endomembrane system	3.83E-09	31.67	0.00
4	GO:0048475	coated membrane	2.54E-15	20.00	0.04

Table 8. Significant go terms (p-value 1E-02 , dispensability 0.05) under molecular function found in the human proteins that are predicted to interact with HIV-1 protein env_gp120.

No	Go-id	Term	P-value	% of proteins	Dispensability
1	GO:0008565	protein transporter activity	7.43E-08	13.33	0.00
2	GO:0008454	alph-1,3 mannosylgly co-protein 4-beta-N-acetyglucosaminyltransferase activity	5.27E-05	5.00	0.00
3	GO:0042802	identical protein binding	1.35E-03	16.67	0.00

5.2 Interactions with TAT

In the case minsup = 0.1%, minconf = 0.1%, TAT interacts with 69 human proteins. By conducting GOTERMSFINDER Tool and web server REVIGO, the results are shown in Tables 9, 10 and 11. The human Proteins have several significant non-redundant terms in each of the

three categories of GO Terms. It appears from the table 9 that many of these human proteins contribute to the response to the defense, and organize positive apoptosis and respond to the virus in order to prevent it from spreading to the cells. And TAT increases in the level of transcription of dsRNA HIV, and therefore it is expected that Tat interacts with human proteins that are involved in the activities of the immune defense and to make them inactive. Moreover, TAT has been found to be a poison that causes apoptosis in uninfected T cells, and thus it helps in the spread of AIDS. Furthermore, it appears from Table 11 that the (9.09%) proteins are associated with the function of the protein molecular transport activity. This means that these proteins may have high participation in transmission between cells. Thus the human proteins predicted to interact with Tat is very intuitive because they share many biological characteristics related to HIV infections.

Table 9. Significant go terms (p-value 1E-03 , dispensability 0.15) under biological process found the human proteins that are predicted to interact with HIV-1 protein TAT.

No	Go-id	Term	P-value	% of proteins	Dispensability
1	GO:0006897	endocytosis	2.23E-05	14.55	0.00
2	GO:0006952	defense response	6.46E-08	27.27	0.00
3	GO:0070661	leukocyte proliferation	6.71E-04	7.27	0.00
4	GO:0001775	cell activation	1.20E-04	14.55	0.03
5	GO:0043065	positive regulation of apoptosis	2.47E-04	16.36	0.03
6	GO:0009615	response of virus	8.62E-04	9.09	0.10

Table 10. Significant go terms (p-value 1E-02 , dispensability 0.15) under cellular component found in the human proteins that are predicted to interact with HIV-1 protein TAT.

No	Go-id	Term	P-value	% of proteins	Dispensability
1	GO:0005829	cytosol	2.30E-10	43.64	0.00
2	GO:0031982	vesicle	5.27E-05	21.82	0.00
3	GO:0048475	coated membrane	1.13E-04	9.09	0.00
4	GO:0031226	intninsic to plasma membrane	2.47E-03	23.64	0.06

Table 11. Significant go terms (p-value 1E-02 , dispensability 0.15) under molecular function found in the human proteins that are predicted to interact with HIV-1 protein TAT.

No	Go-id	Term	P-value	% of proteins	Dispensability
1	GO:0008565	protein transporter activity	3.17E-04	9.09	0.00

2	GO:0016814	hydrolase activity ,acting on carbon-nitrogen bonds , in cyclic amidines	6.05E-03	5.45	0.00
3	GO:0051219	phosphoprotein	1.39E-04	7.27	0.00

5.3 Interactions with other HIV-1 proteins

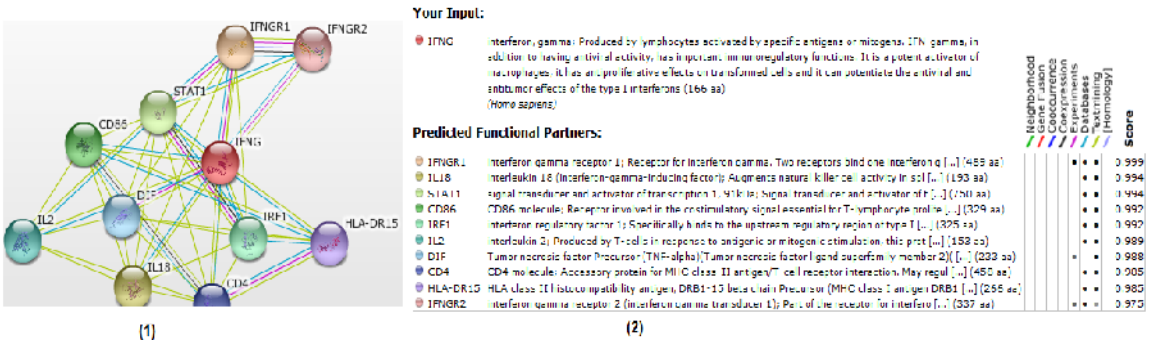

Figure 6. The association among the human protein IFNG as input (2) with all interacted human protein in (1)

In the case of minsup = 0.1%, minconf = 0.1%, the proposed algorithm predicted with 327 interactions between HIV-1 and human proteins. The HIV-1 protein capsid interacts with (IFNG, HLA-DRA and PPIA). These 4 proteins interact with each other as found in the STRING database 4 that known and predicted protein interaction by writing the protein name and choose GO to choose the Homo sapiens organism to display all the interacted human proteins with your input human protein such as in figure 6 . The HIV-1 protein Gag_Pr55 interacts with (IFNA1, IL1B, HLA-A, HLA-DRA and PPIA). These 5 proteins assist in the positive regulation of the immune system. These Proteins are found in the T-cell receptor signaling pathway and responsible for the protection of the host.

In the case of minsup = 0.1%, minconf = 0.7%, The HIV-1 protein env_gp160 is predicted to interact with 8 human partners (HLA-E, FOS, HLA-F, HLA-G, JUN, LCK, FAS, and MAPK8). These proteins have an important role in the regulatory movement of the cell, and the organization of B-cell activation/proliferation, and the organization of T-cell activation / proliferation. Thus, these proteins contribute to the implementation of the humoral immune cell-mediated and affects the viral infection such activities to control the immune cells. The HIV-1 protein env_gp141 is predicted to interact with (HLA-DRB3, HLA-DRB4, HLA-DRB5, PRKCD, PRKCG, PRKCH, PRKCI and PRKCZ) which are found to be involved in the B-cell/T-cell signaling pathway, apoptosis and programmed cell death. Moreover all of them are involved in the signaling pathway of prostate cancer. The HIV-1 protein Nef interacts with (NFATC1, FN1, FOS, IL4, IL8, IL12A, CASP3, CD3D, CD28, HLA-B, HLA-C, HLA-DRB1, HLA-E, HLA-F, JUN, HLA-G, LCK, CCL3, CCL4, CCL5, TP53, ACTB, FAS, BCL2, PPIA, PRKACA and MAPK8). Nef causes T-cell activation and it helps in the survival of infected cells in the surface of the immune function of the host. Nef interacted with above 27 human proteins involvement in apoptosis and programmed cell death. The human proteins predicted to interact

[4] http://string-db.org

with the HIV-1 proteins neucleocapsid, p6, retropepsin, Rev, Vpr, Vpu and RT also share many biological activities related to immune response, apoptosis and programmed cell death.

6. CONCLUSION AND FUTURE WORK

This paper produced the proposed algorithm for mining association rules and conceptual bi-clusters without information loss to predict with interactions between HIV-1 and human proteins. This algorithm produces a minimal non-redundant cover for association rules, from which all rules generated by Apriori can be, deduced if required, that is much smaller, and generating the bi-clusters defined by inclusion relation. The method was validated by applying it for predicting HIV-1-Human protein interactions. Besides proving faster than Apriori algorithm like mining methods, the results obtained by the proposed algorithm. In the future, we plan to apply the proposed algorithm to integrate additional information about proteins, like structural and sequential similarities, with protein-protein interactions to improve the results.

REFERENCES

[1] Agrawal,R., Mannila, H.,Srikant, R.,Toivonen,H.,and Verkamo,A.I.(1996).Fast discovery of association rules.In Advances in Knowledge Discovery and Data Mining, pages307–328.AAAI/MITPress.

[2] D.D. Ho and P.D. Bieniasz, Cell 133, 561-5 (2008).

[3] http://www.avert.org/worldwide-hiv-aids-statistics.htm ,20 Jul 2013.

[4] UNAIDS. 2007 AIDS epidemic update. (2007).

[5] A. Trkola, Curr Opin Microbiol 7, 555-9 (2004).

[6] Jansen,R.,H.Yu,D.G.,Kluger,Y.,Krogan,N.J.,Chung,S.,Emili,A.,Snyder,M.,Greenblatt,J.F.,andGe r stein,M.(2003). Abayesian networks approach for predicting .

[7] Arkin, M.R. and Wells ,J.A.(2004).Small-molecule inhibitors of protein-protein interactions : Progressing towards the dream. Nat. Rev. Drug Discov., 3:301–317.

[8] Pasquier, N., Taouil, R., Bastide, Y.,Stumme, G., and Lakhal, L.(2005).Generating acondensed representation for association rules. J. Intell. Inf. Syst., 1(24):29–60.

[9] Tastan, O., Qi, Y., Carbonell, J., and Klein-Seetharaman,J. (2009). Prediction of interactions between HIV-1and human proteins by information integration. In Proc. PSB, pages516–527.

[10] Madeira, S.C. and Oliveira, A.L. (2004) .Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Trans .Comput. Biol. Bioinform., 1:24–45.

[11] Bandyopadhyay S, Maulik U, Holder LB, Cook DJ (2005) Advanced Methods for Knowledge Discovery from Complex Data (Advanced Information and Knowledge Processing). Springer-Verlag, London.

[12] Hipp J, Güntzer U, Nakhaeizadeh G (2000) Algorithms for association rule mining – a general survey and comparison. SIGKDD Explorations 2: 58–64. doi: 10.1145/360402.360421.

[13] Goethals B (2002) Efficient Frequent Pattern Mining. Ph.D. thesis, University of Limburg, Belgium.

[14] Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data (SIGMOD'93).

[15] Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in large databases. In: Proc. 20th International Conference on Very Large Data Bases. San Francisco, CA, USA.

[16] Pasquier N, Bastide Y, Taouil R, Lakhal L (1999) Discovering frequent closed itemsets for association rules. Proc. 7th International Conference on Database Theory (ICDT-99). pp. 398–416.

[17] Zaki MJ, Hsiao CJ (2005) Efficient algorithms for mining closed itemsets and their lattice structure. IEEE Transactions on Knowledge and Data Engineering 17: 462–478.

[18] Mukhopadhyay A, Maulik U, Bandyopadhyay S (2010) On biclustering of gene expression data. Current Bioinformatics 5: 204–216.

[19] Ptak RG, Fu W, Brigitte , Dickerson JE, Pinney JW, et al. (2008) Cataloguing the HIV Type 1 Human Protein Interaction Network. AIDS Research and Human Retroviruses 24: 1497–1502.
[20] http://www.ncbi.nlm.nih.gov/projects/RefSeq/HIVInteractions/env.html , 15 jun 2013.
[21] http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/ , 14 Jan 2013.

Authors

Nour moustafa abd elhameed received his BSc.degree in information system from Faculty of Computers and Information, Helwan University, Cairo, Egypt in 2009. He is currently working as a teacher assistant at the same faculty, and a master's degree student under the supervision of prof. Ahmed Sharaf Eldin and prof. Samar Kamal Kassim . His areas of interests include Data mining, Bioinformatics, databases, and programming languages.

Prof. Ahmed Sharaf Eldin is among the first generation of the computer people in Egypt. He obtained his Ph. D. in Computer Science from Loughborough University in UK. His first degree was in Engineering from Cairo University. His postgraduate studies span over several fields of knowledge: computer science, statistics, operations research, management, and Islamic studies. He obtained the Cairo University prize in 1971. He is a co-inventor of a programmable calculator 1978. Obtained several prizes from different universities and professional bodies. He published more than 150 scientific and technical papers in international and national journals and conferences. Moreover, several technical reports were submitted in addition to many essays in newspapers and magazines. He authored three textbooks and co-authored three others. He was selected as session organizer and chairman for the Annual Pittsburgh Conference on Modeling and Simulation, University of Pittsburgh, Pittsburgh, USA for several years. He was an invited speaker in many international and national conferences in Egypt, Saudi Arabia, Qatar, and USA.He also obtained (with others) the Prince Sultan Ibn Abdelaziz International prize for water resources (fourth branch) in Dec., 2004. Dr. Sharaf Eldin was honored by his university (Helwan University) in July 2009 as the distinguished professor. He is also a member of the National Committee for Informatics in Egypt and a member of the National Committee for Quality Assurance (in Education) in Egypt. He was also a member of the steering committee for the establishment of the Egyptian E-learning University (EELU). He was also a member of the steering committee for the establishment of the Egyptian Japanese University for Science and Technology (E-JUST). He is also a member of the board of the Egyptian Computer Society and the Egyptian Software Engineering Association. He is also the general coordinator of the sector committee for Informatics in the supreme council of Egyptian Universities (SCU). He is also the academic advisor of the Equivalence Committee by SCU for the computing sector. Prof. Sharaf Eldin established the QA systems and mechanisms at Helwan University (HU). He also established the Student Assessment Systems at HU. Prof. Sharaf Eldin won several research and development projects sponsored by national and international bodies like SOLERAS (USA and SA) , Tempus (EU), KACST (SA), KSU (SA), HU (EG), HEEPF (EG and World Bank), QAAP (EG and World Bank), PCIQA (EG). He is also a peer reviewer for NAQAAE.

Samar Kamal Kassim Professor of Oncology Director Diagnostic Unit .Medical Biochemistry and Molecular Biology Department,Faculty of Medicine, Ain Shams University ,Egypt.