Association Rule Mining to Deduce the Most Frequently Occurring Amino Acid Patterns in HIV

Kumud Pant
Department of Biotechnology
Graphic Era University
Dehradun, India

Bhasker Pant
Department of IT
Graphic Era University,
Dehradun, India

Shweta Negi
Department of Biotechnology
Graphic Era University,
Dehradun, India

ABSTRACT
HIV is one of the most dreaded diseases of the century. Throughout the world efforts are underway to develop new vaccines and design new drugs so as to combat this viral menace. In an effort to probe deeper into the functioning of these viruses we present association based rules formulation so as to decipher the most frequently occurring amino acids in these viruses.

This is a novel attempt of its kind since we are attempting to find put the most informative association rules using Apriori algorithm implemented through WEKA. The information generated can be of great use to molecular biologists and drug designers since the associated amino acids can be a very good drug targets.

Our findings suggest that L-Selenocysteine and L-Pyrrolysine are most frequently associated amino acids in the 4 classes of virulent proteins analyzed for association rules and Cyteine and Arginine show the strongest association in one of the class analyzed i.e. Gp41. Hence these can be potential drug candidates.

General Terms
Association Rule Mining, HIV, Apriori Algorithm.

Keywords
Human Immunodeficiency virus (HIV), Apriori Algorithm, Association rule mining, Amino Acids.

1. INTRODUCTION
HIV has certainly amazed scientists throughout the world with their immense capability to not only replicate enormously but also mutate at a very amazing rate. This results in production of number of mutants, hence making it difficult to develop and formulate one common and universal vaccine for this group of virus.

The number of people living with HIV rose from around 8 million in 1990 to 34 million by the end of 2010. The overall growth of the epidemic has stabilized in recent years. The annual number of new HIV infections has steadily declined and due to the significant increase in people receiving antiretroviral therapy, the number of AIDS-related deaths has also declined (1). Still the menace created by this virus has surpassed that by any other pathogenic organism.

The genome of HIV is depicted in the picture below

Fig 1: The major proteins encoded by HIV-1

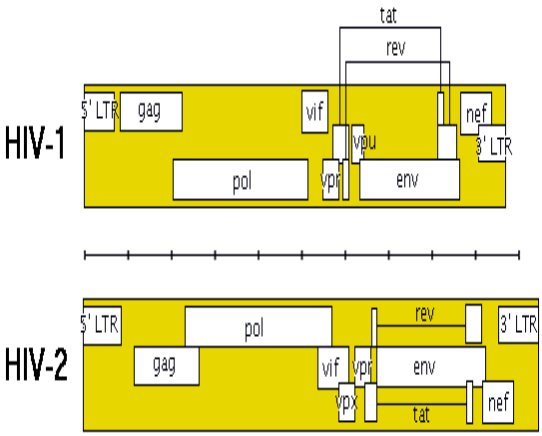

Fig 2: The arrangement of genes in HIV-1 (2, 3)

The genes in HIV's genome are as follows (3):

1) gag (coding for the viral capsid proteins)
2) pol coding for reverse transcriptase. The gag and pol together can be expressed in one long strand called "gag-pol".
3) env coding for HIV's envelope-associated proteins and the regulatory genes:
4) tat. Tat" is short for "transactivator" - it's a regulatory gene which accelerates the production of more HIV virus.
5) rev. It is one more of HIV's regulator genes. It stimulates the production of HIV proteins, but suppresses the expression of HIV's regulatory genes.
6) nef. The "negative replication factor" ("nef") gene encodes a protein which remains present in the cytoplasm of the cell, and retards HIV replication.

7) vif. The "vif" gene codes for "virion infectivity factor", a protein that increases the infectivity of the HIV particle.

8) Vpr. "Viral protein R" accelerates the production of HIV proteins

9) vpu (not present in HIV-2)

10) vpx (not present in HIV-1).

In this research paper we are trying to deduce the most frequent amino acid patterns for the above major classes of proteins in HIV so as to understand the underlying pattern of gene expression.

Here we have used Apriori algorithm and its implementation through WEKA suite of software for achieving our aim. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is free software available under the GNU General Public License (5, 6). With association rule mining we can find frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories, which in our case are repository of protein sequences present in NCBI. According to the basic concept of association rule mining, let there be an Itemset X={x1 …, xk}, ARM helps to find all the rules X→Y with min confidence and support. Support, s is the probability that a transaction contains X∪Y and confidence, c is the conditional probability that a transaction having X also contains Y. The concept of Apriori property says that all subsets of a frequent itemset must also be frequent. We generate candidate set of frequent itemsets till no more exist by pruning at each step till which completes the Apriori (4).

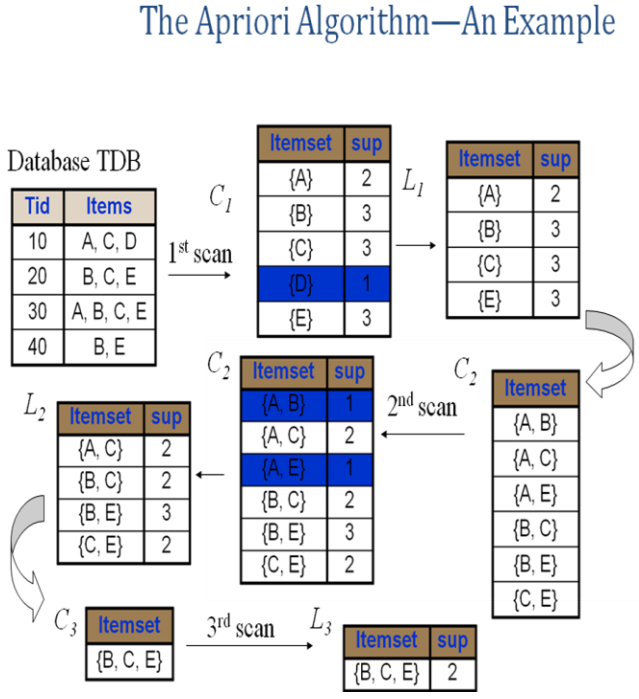

Figure1: Basic concept of Apriori with an example

We have implemented Apriori algorithm for implementation of Association Rule Mining through WEKA. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

Previously also ARM has been used by scientists to understand HIV and its gene expression. Anirban Mukhopadhyay et al proposed bisclustering approach to Association Rule Mining for predicting HIV-1 human protein interactions. The main focus has been to discover association rules that stand true for both human and HIV proteins which in turn can be used to predict new interactions (9). Similarly Dr. K.Ramesh Kumar gave concept of association rule mining from missing values to handle large dataset of HIV patient's records which can further help to understand the epidemiology of diseased cells (10). Sinu Paul and Helen Piontkivska et al used technique of association rule mining to identify cytotoxic T-lymphocyte (CTL) epitopes in different genes that co-occur in HIV genome which can be used to design complex vaccines or multi epitope vaccine and drugs (11).

Here we are attempting to understand the most common occurring amino acid/s which can define the entire virulent class of proteins with the help of association rules. The rules generated can provide us a basic understanding of amino acid patterns which in turn can aid in drug design and discovery against most frequently associated amino acids

2. MATERIALS AND METHODS

2.1 The dataset

We have taken 83 proteins of HIV belonging to these classes Gp120, Gp41, Gp160 and P17. These proteins have following functions

P17 protein matrix protein is one of the products of the HIV gag gene. It has a structural function inside mature HIV particles and helps to anchor the gp41/gp120 "spikes" to the envelope (2, 3).

The protein GP120 present on the outer envelop of the virus binds to CD4 receptor on the human cells and mediates entry of viruses in the cells.

Gp41protein is a subunit of the envelope protein complex of retroviruses, including Human immunodeficiency virus (HIV) and Simian-Human immunodeficiency virus (SHIV). This glycoprotein subunit remains non-covalently-bound to gp120, and provides the second step by which HIV enters the cell (7).

Gp160 protein is precursor protein for another important protein needed tocomplete life cycle of HIV i.e. gp41 and gp120.

2.2 The Weka mining suite

Weka is a freely available data mining suite of software with facilities to carry out many data mining tasks (6).

3. RESULTS AND DISCUSSION

3.1 Rules framed for all the classes evaluated together

With all the 23 amino acids present in above proteins of HIV at minimum support of 0.95 and confidence of 0.9 the size of large itemset L1 was 3, L2was also 3 and L3 was 1. Best rule showing the most frequently associated amino acids is

U=L 8==> O=L 83 indicating that all 83 instances belonging to 4 classes of proteins showed strong associations with these two amino acids. Here U is L-Selenocysteine and O is L-

Pyrrolysine. The amino acid composition of these 23 amino acids has been discretized into Low (L), Medium (M) and High (H) and both the above amino acids have been found to be associated in Lower range in these amino acids. Similarly strong associations have also been observed between Methionine (M) and both the above amino acids.

For Gp120 again the best association was seen in case of U an O amino acid which had 56 instances and all the instances were defined with these associations. Here Lift is the confidence divided by the proportion of all examples that are covered by the consequence. This is a measure of importance of association that is independent of Support. Leverage is the proportion of additional examples covered. Conviction (conv:) is the measure of departure from independence given by P (premise)P(!consequence)/ P (premise, !consequence).

1. U=L 56 ==> O=L 56 <conf:(1)> lift:(1) lev:(0) (0) conv:(0)

2. O=L 56 ==> U=L 56<conf:(1)> lift:(1) lev:(0) (0) conv:(0)

3. class=Gp120 56==>O=L 56<conf:(1)> lift:(1) lev:(0) (0) conv:(0)

4. O=L 56 ==> class=Gp120 56<conf:(1)> lift:(1) lev:(0) (0) conv:(0)

5. class=Gp120 56 ==> U=L 56<conf:(1)> lift:(1) lev:(0) (0) conv:(0)

6. U=L 56 ==> class=Gp120 56<conf:(1)> lift:(1) lev:(0) (0) conv:(0)

7. U=L class=Gp120 56 ==> O=L 56<conf:(1)> lift:(1) lev:(0) (0) conv:(0)

8. O=L class=Gp120 56 ==> U=L 56<conf:(1)> lift:(1) lev:(0) (0) conv:(0)

9. O=L U=L 56 ==> class=Gp120 56<conf:(1)> lift:(1) lev:(0) (0) conv:(0)

10. class=Gp120 56 ==> O=L U=L 56

For Gp41 interesting pattern was observed where 10 large item sets were generated indicating varied distribution of composition defining the class Gp41.

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 63

Size of set of large itemsets L(3): 192

Size of set of large itemsets L(4): 378

Size of set of large itemsets L(5): 504

Size of set of large itemsets L(6): 462

Size of set of large itemsets L(7): 288

Size of set of large itemsets L(8): 117

Size of set of large itemsets L(9): 28

Size of set of large itemsets L(10): 3

The best rules generated showed the associations between Cysteine (C) and Arginine (R) amino acid. The best 10 rules are indicated below

In all the rules generated above amino acid Arginine (R) was found to be associated with all the amino acids and seemed to show important associations with other amino acids.

1. C=L 16 ==> R=L 16 <conf:(1)> lift:(1) lev:(0) (0) conv:(0)

2. R=L 16 ==> C=L 16 <conf:(1)> lift:(1) lev:(0) (0) conv:(0)

3. H=L 16 ==> R=L 16 <conf:(1)> lift:(1) lev:(0) (0) conv:(0)

4. R=L 16 ==> H=L 16 <conf:(1)> lift:(1) lev:(0) (0) conv:(0)

5. L=H 16 ==> R=L 16 <conf:(1)> lift:(1) lev:(0) (0) conv:(0)

6. R=L 16 ==> L=H 16 <conf:(1)> lift:(1) lev:(0) (0) conv:(0)

7. F=L 16 ==> R=L 16 <conf:(1)> lift:(1) lev:(0) (0) conv:(0)

8. R=L 16 ==> F=L 16 <conf:(1)> lift:(1) lev:(0) (0) conv:(0)

9. P=L 16 ==> R=L 16 <conf:(1)> lift:(1) lev:(0) (0) conv:(0)

10. R=L 16 ==> P=L 16

On the basis of above findings we can say that strong association between U (L-Selenocysteine) and O (L-Pyrrolysine) found in all classes of proteins suggest a very important role played by these amino acids in functioning of the proteins. Similarly when one of the virulent protein class were analyzed to deduce the most frequently associated amino acid pattern, the number of large item set generated also showed an increase to 10 indicating variable distribution of composition and heterogeneity.

An interesting observation correlated with the findings of Long, Y showed that Arginine (R) was most frequently associated amino acid (8). In all the 10 rules deduced Arginine was found to be associated with all the amino acids.

The above problem has been analyzed using the scale of 3 ie low, medium and high. If we scale the data into further categories like very low, low ,medium ,high and very high then some more interesting rules can be framed and many of the amino acid combinations which have not been successful in defining rules for the classes can then provide the framework for association patterns in them.

4. CONCLUSION

The above rules indicate the importance of the corresponding amino acids in understanding a particular class of pathogenic proteins. The amino acids which are found to be most strongly correlated and defining the entire class of proteins (here it is is L-Selenocysteine and L-Pyrrolysine) can be very effective drug targets. The above rules generated are very stringent since high values of support and confidence are taken which eliminates the rules which are less frequent. Study of association rules can help us to understand the general trend in selection of amino acids by HIV for its pathogenicity and disease causing ability. The only drawback with the apriori is that large computation is required for mining the frequent patterns because the number of scans required is higher but refined rules come out using the correct dataset.

These association rules can be very helpful in designing new vaccine candidates for the most frequently associated amino acid which can aim at entire class of proteins.

With different rules framed for different classes frequent pattern differences can be understood between them which can help to distinguish proteins among different species but also in the same specie. The rules can be expanded to include other virulent classes of not only hiv but also classes Such association studies can be performed with freely available proteomic and genomic data of any organism.

5. REFERENCES

[1] Universal Access to HIV Prevention, Treatment, Care and Support: From Countries to Regions to the High Level Meeting on AIDS and Beyond, UNAIDS, 2011.

[2] HIV immunology site at Los Alamos HIV databases and compendia at http://www.hiv.lanl.gov/content/ immunology.

[3] The Molecules of HIV - A Hypertextbook written by Dan Stowell at http://www.mcld.co.uk/hiv/ .

[4] Ian, H. Witten, Eibe, F. and Mark, A. Hall. 2011. Data Mining: Practical machine learning tools and techniques. 3rd Edition. Morgan Kaufmann, San Francisco.

[5] Holmes, G., Donkin, A. and Witten, I.H. 1994. Weka: A machine learning workbench. Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia.

[6] Weka available at http://www.cs.waikato.ac.nz/ml/weka/arff.html.

[7] Kim, P. S., Malashkevich, V. N., Chan, D. C. and Chutkowski, C. T. 1998. Crystal structure of the simian immunodeficiency virus (SIV) gp41 core: conserved helical interactions underlie the broad inhibitory activity of gp41 peptides. Proc. Natl. Acad. Sci. U.S.A. 95 (16): 9134–9139.

[8] Long, Y., Meng, F., Kondo, N., Iwamoto, A. and Matsuda Z. 2011. Conserved arginine residue in the membrane-spanning domain of HIV-1 gp41 is required for efficient membrane fusion Protein. Cell. 2(5):369-76.

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

[10] Rameshkumar, K. 2012. Association Rules Mining from HIV/AIDS patients' case history database with missing values. International Journal on Data Mining and Intelligent Information Technology Applications (IJMIA), Volume2, Number1, March 2012.

[11] Sinu, P. and Piontkivska, H. 2009. Discovery of novel targets for multi-epitope vaccines: Screening of HIV-1 genomes using association rule mining, Retrovirology, 6:62.