Mining of Association Rules from HIV-1 Protein Data

D. Suresh Babu
Department of Computer Science & Engineering
V. R. Siddhartha Engineering College
Vijayawada, Andhra Pradesh, India

K. Suvarna Vani,
Department of Computer Science & Engineering
V. R. Siddhartha Engineering College
Vijayawada, Andhra Pradesh, India

T. D. Sravani
Department of Computer Science & Engineering
V. R. Siddhartha Engineering College
Vijayawada, Andhra Pradesh, India

Abstract

Analysis of protein structure database usually reveal different motifs associated with biological functions. Sparse matrices have been used in ab-initio methods for the problem of protein structure prediction problem. Secondary structures and contacts made by the residues are clearly visible in the matrices where helices are seen as thick bands and the beta sheets are seen as orthogonal to the diagonal. This paper explores the idea of extracting rules from sparse matrices to represent "protein motifs" information. Data mining techniques generally consider the structural and functional properties of protein secondary structure motifs along the diagonal of sparse data. Similarly we also find the number of parallel and anti-parallel motifs from lower region of the sparse data. The proposed approach to be a promising tool for detecting the hidden motifs from the HIV-1 protease data. In this paper we propose two modules one module is to generate frequent item-sets from machine learning algorithm like Apriori. Rules based on the diagonal and the off-diagonal motifs present in the matrices are extracted in order to predict HIV-1 protease data.

1. Introduction

AIDS is caused by Human Immunodeficiency Virus (HIV). HIV is a lentivirus, which is a class of retrovirus. The lentivirus means slow virus because these type of viruses take a long time to cause the disease. Most lentiviruses target cells of the immune system and thus disease is often known as immunodeficiency. There are two types of HIV: HIV-1 and HIV-2. Both are same but time taken to find the disease for HIV2 is more than HIV1. The worldwide epidemic of HIV and AIDS is caused by HIV-1 while HIV-2 is mostly restricted to West Africa. HIV-1 is more epidemic than HIV-2.

HIV-1 protease is a retroviral asperity protease which is important for the lifecycle of HIV, to cause AIDS. A protease is an enzyme that cleaves proteins to their component peptides. HIV protease cleaves newly synthesized polyproteins at the appropriate places to create the mature protein components of an infectious HIV virion. HIV viron can work by HIV protease only. Thus, mutation of HIV protease's active site or inhibition of its activity disrupts HIVs ability to replicate and infect additional cells, making HIV protease inhibition the subject of considerable pharmaceutical research.

Generally a protein has primary structure, secondary and tertiary structure.The primary structure is nothing but linear arrangement of the 20 amino acids. The secondary structure of any type of protein describes the pattern of hydrogen bonding between amino acids along the primary sequence. The common secondary structures in proteins are

- Alpha- helix
- Beta-sheet
- Turns

An Alpha-helix is a stable structure where each residue forms a hydrogen bond with another one that is four residues apart in the primary sequence. It is coiled tightly in the fashion of a spring. They appear along the diagonal in contact map.

A Beta-sheet is another type of stable structure formed by at least two beta-strands that are connected together by hydrogen bonds between the two strands. A parallel beta-sheet is a sheet where the two beta-strands have the same direction while an anti-parallel beta-sheet is one that does not. It forms a zigzag shaped protein structure called Beta-strand. They appear parallel and anti-parallel to the diagonal in contact map.

A Turn is a secondary structure that usually consists of 4-5 amino acids to connect alpha-helices or beta-sheets.

The tertiary structure is the local conformation of three-dimensional arrangement of the amino acids and they represent by the x, y and z coordinates of all the atoms of a protein or by the coordinates of the backbone atoms. The secondary structures structures are folded into tertiary structure depending on hydrophobic forces and side chain interactions, such as hydrogen bonding between amino acids.

The structure of a HIV protein reveals important information, such as the location of probable functional or interaction sites, identification of distantly related proteins, and discovery of important regions of the HIV that are involved in maintaining the structure, and so on. Hence there is a need for development of computational techniques for prediction and classification of HIV-1. Functioning of a protein in biological reactions not only depends on its amino acid sequence (primary structure) but also crucially relies on its three-dimensional configuration (tertiary structure). The classification of the tertiary structure of a protein from its amino acid sequence still remains as an unsolved issue in bioinformatics and molecular biology. Here we try to predict the tertiary structure of a protein based on the secondary structure. In this paper we extract the secondary structure features from the HIV-1 protease data and generate association rules in order to know the functioning of HIV-1 protease.

1.1. Protein Sparse Matrix

The sparse matrix of a protein is the simple representation in which we can predict the secondary structure of a protein and it is also an intermediate step to know about the 3D structure of a protein. In brief sparse matrix is a distance map which is a 2D symmetric square matrix in which it has either 0 or 1 values if the value is 1 that there is contact between two residues otherwise 0. For 1 it is defined as the distance between two residues is less than threshold value of particular distance.

1.2. Association Rule Mining

Association rule mining is one of the major techniques to detect and extract useful information from data. It aims to extract interesting correlations, frequent patterns, associations among sets of items in the transaction databases or other data repositories. It finds association rules which satisfy the predefined minimum support and confidence from a given database. Association rules are typically generated in a two-step process. First, minimum support is used to generate the set of all frequent item-sets for the data set. Frequent item-sets which satisfy the user specifying minimum support and confidence constraints. Then, in a second step, each frequent item-sets is used to generate all possible rules from it and all rules which do not satisfy the minimum confidence constraint are removed. The two basic parameters of association rule mining are support and confidence.

Support of an association rule is defined as the percentage or fraction of motifs that contain X Y to the total number of records in the database. The count for each item is increased by one every time the item is encountered in different transaction T in database D during the scanning process. Support(s) is calculated by the following

Support(XY) = Support count of XY/Total number of transaction in D

Confidence of an association rule is defined as the percentage or fraction of the number of transactions that contain XY to the total number of records that contain X, where if the percentage exceeds the threshold of confidence an interesting association rule X=>Y can be generated. Confidence is a measure of strength of the association rules.

Confidence(X|Y) = Support(XY)/Support(X)

Apriori algorithm [10] is a classic algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the

database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database. The Apriori algorithm uses a key property called the "downward-closure" of the support, which states that if an itemset passes the minimum support then all of its subset must also pass the minimum support. This means that any subset of a frequent itemset have to be frequent, where else, any superset of infrequent itemset must be infrequent.

The frequent subsets are extended one item at a time known as candidate generation, and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length k − 1. Then it prunes the candidates which have an infrequent sub pattern.

2. Related Literature

Anirban et al. [1] had extracted different association rules for the HIV 1 protease data and human proteins. They used Apriori algorithm in order to obtain the different association rules for the protein-protein interactions. Based on these rules they have predicted some new viral and human protein interactions. Kartick et al. [2] has proposed a new approach called FIST in which it combines two methods that is the bi-clustering and the association rule mining. The FIST algorithm is Frequent Itemset mining using Suffix Trees. It is a three step process: (1) Pre-processing the dataset, (2) Extracting frequent closed itemsets, (3)Finding bases for association rules and hierarchical conceptual bi-clusters. By this they identified the interactions between the human and virus proteins more efficiently. Gene et al. [3] has provided the data mining analysis of the HIV1 crystal protein structures. For each crystal protein structure they extracted the binding pockets features are being extracted using the various methods like DeepView Swiss-PdbViewer etc. From them we select a subset of features using the feature selection methods like Hybrid binary particle swarm optimization artificial neural network and random forest and unsupervised learning locally linear embedding.

Clustering is used to validate whether the selected descriptors best correlates the shape of the protein structure with its complexed ligand or not. Sandro et al. [4] has used association rules in order to obtain the specificity rules of whether the HIV1 proteins has the cleavage sites or not. Dubey et al. [5] has used various machine learning algorithms for the classification of HIV secondary structure of protein enzymes. They used algorithms such as J48, random forest, rotation forest. Here they consider the HIV1 proteins as the positive class and the HIV2 as the negative class in order to classify alpha, beta and residues of HIV reverse transcriptase, protease, ribonuclease, integrase.

Zaki [11] et al carried out mining of protein contact maps. They demonstrated how data mining is used to extract valuable information from protein. From protein sequence they discovered dense patterns using sliding window technique and used hashing for storing the results. The dense patterns are clustered using agglomerative technique. Pattern mining and clustering results can be helpful for protein structure predictions and discovering protein folding pathway.

3. Methodology

The steps for mining generation of association rules from the HIV data are:

Step 1: Preparation of the HIV dataset using the PDB files.

Step 2: Generation of the sparse matrices from the protein 3D structure.

Step 3: Feature extraction from the diagonal and off diagonal motifs.

Step 4: Mining of different association rules from the extracted motifs.

3.1. Dataset

Here the HIV-1 protease data and HIV-2 protease data has been taken from the PDB (Protein Data Bank). We considered the 50 proteins from these HIV-1 protease data and 20 proteins for HIV-2 protease data. Here these PDB files are being downloaded from [9]. This Protein Data Bank [9] is

the primary repository for experimentally determined 3D protein structures. These structures were created using crystallography methods. PDB entries contain additional information such as references, structure details and other features.

Here in these files it contains the information which describe about the protein 3d structure from these 3D structures information we generate the 2d structure of the protein. In this 2d or secondary structure we derive only the alpha helixes and the beta sheets present in the HIV data.

4. Feature Extraction

Here we extract different types of feature sets from the sparse matrices and these feature sets are applied for the apriori algorithm. For the feature set 1 and feature set 2 first we have to extract the protein sparse matrices from the PDB files. And also we have to extract the secondary structural features.

4.1. Feature Set 1: We apply the secondary structural motifs from prediction algorithm [6] which we extract 6 features namely number of helices, minimum helix length, maximum helix length, number of beta sheets, minimum beta-sheet length and maximum beta-sheet length. In Table 1 we provide sample secondary features being derived for different HIV-1 protease proteins.

4.2. Feature Set 2: We apply the Eight Neighbor Algorithm [7] for the off- diagonal matrices. Here we extract 5 features namely number of clusters, minimum cluster density, maximum cluster density, number of parallel and anti parallel beta sheets. In Table 2. we provide sample cluster features derived for different HIV-1 protease proteins.

4.3. Feature Set 3: In this set combination of feature set one and two total we get eleven features. In the Fig. 1 the first image shows the protein 3D structure and second image shows the cartoon topology of the protein structure and the last image shows the 2D representation of the protein 3D structure. In this 2D representation the red colour circle is an alpha helix along the diagonal and pink colour circles indicates that the parallel and anti-parallel beta sheets of the off diagonal motifs of the matrix for protein id is 1D4I.

5. Association Rules Generation

After extracting feature set 1, 2 and 3 we apply these feature sets to the Apriori algorithm in the WEKA [8]. Here we test the different feature datasets using the data mining toolkit WEKA (Waikato Environment for Knowledge Analysis) version 3.6.6. WEKA is an open source toolkit and it consists of collection of machine learning algorithms for solving data mining problems [8]. Here we generate frequent 1, 2, 3, 4, 5 itemsets. The frequent 3 item-sets generated for all feature sets are present in Table 3. Here we experiment on the different feature sets by setting different values for the minimum support from 0.1 to 0.4 and confidence as 1. In the Table 4 we show the best 5 rules for all the feature sets in which we consider the minimum support as 0.4 and confidence as 1.

6. Conclusion

Novel features have been obtained by mining protein sparse matrices using a simple and effective eight neighbor algorithm. It proves to be very effective in extracting continuous patterns of variable sizes from sparse matrices, unlike the sliding window algorithms in the literature[11] which give fixed size of motifs. The effectiveness of these features is to identify the maximum length of helix secondary structural motif from the HIV-1 protease data. By using this we observe the functioning of the HIV-1 protease and predict the backbone of 3D structure.

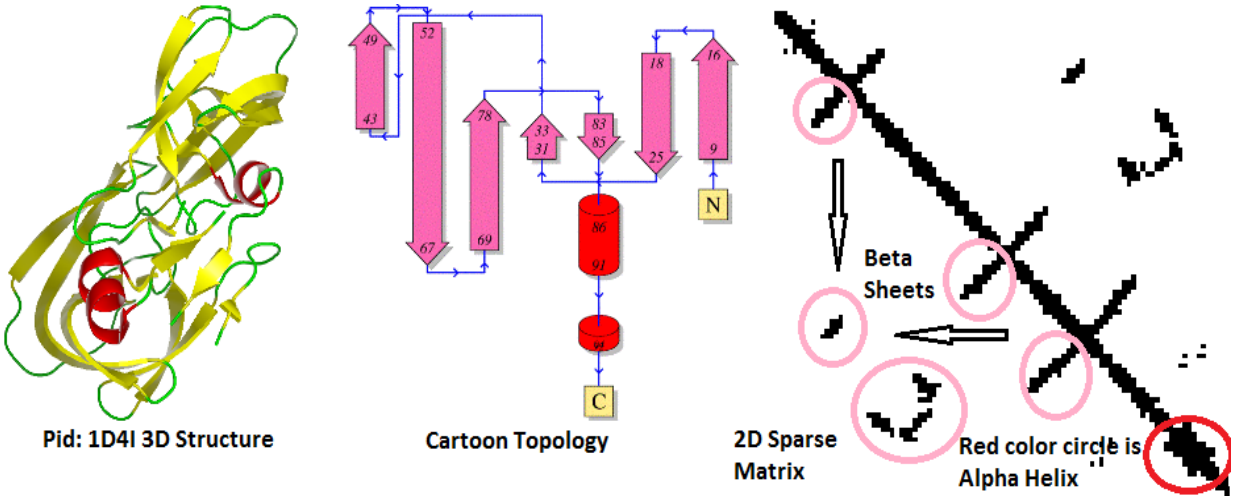

Figure 1. 1D4I: HIV-1protease protein 3D and 2D struture.

Table 1. Feature set 1: Secondary Structural Motifs Along the Diagonal Matrices

Protein id	Number of helices	Minimum helix length	Maximum helix length	Number of beta sheets	Minimum beta sheet length	Maximum beta sheet length
1AKJ	5	4	28	21	3	11
1CZQ	1	42	42	0	0	0
1D4H	1	6	6	16	3	5
1D4I	1	6	6	16	3	5
1KZK	2	3	6	11	3	5
1QNZ	2	3	3	13	3	11

Table 2. Feature set 2:Cluster Features Away from diagonal Matrices

Protein id	Number of clusters	Minimum cluster density	Maximum cluster density	Number of Parallel sheets	Number of Anti parallel sheets
1AKJ	18	41	3	2	12
1D4H	7	32	13	2	5
1KZK	8	32	4	2	5
1MFS	2	26	16	1	0
1QNZ	11	36	5	2	8

Table 3. Frequent Item Sets for Different Feature Sets

Feature set	Frequent Item sets
Feature set 1	• Number of helices=1 Min helix length=6 Max helix length=6 34 • Number of helices=1 Min helix length=6 Min beta=3 34 • Number of helices=1 Max helix length=6 Min beta=3 34 • Min helix length=6 Max helix length=6 Min beta=3 34
Feature set 2	• Number of Parallel sheets3=2 Number of Parallel sheets4=2 Number of Parallel sheets5=2 40
Feature set 3	• Min beta=3 Number of Parallel sheets3=2 Number of Parallel sheets4=2 40 • Min beta=3 Number of Parallel sheets3=2 Number of Parallel sheets5=2 40 • Min beta=3 Number of Parallel sheets4=2 Number of Parallel sheets5=2 40 • Number of Parallel sheets3=2 Number of Parallel sheets4=2 Number of Parallel sheets5=2 40

Table 4. Top 5 rules for different feature sets

Feature set	Top 5 Rules
Feature set 1	1. Maximum helix length=6 35 ==> Minimum beta sheet length=3 35 2. Minimum helix length=6 34 ==> Number of helices=1 34 3. Minimum helix length=6 34 ==> Maximum helix length=6 34 4. Minimum helix length=6 34 ==> Minimum beta sheet length=3 34 5. Minimum helix length=6 Maximum helix length=6 34 ==> Number of helices=1 34
Feature set 2	1. Number of Parallel sheets4=2 40 ==> Number of Parallel sheets3=2 40 2. Number of Parallel sheets3=2 40 ==> Number of Parallel sheets4=2 40 3. Number of Parallel sheets5=2 40 ==> Number of Parallel sheets3=2 40 4. Number of Parallel sheets3=2 40 ==> Number of Parallel sheets5=2 40 5. Number of Parallel sheets5=2 40 ==> Number of Parallel sheets4=2 40
Feature set 3	1. Number of Parallel sheets3=2 40 ==> Minimum beta sheet length=3 40 2. Number of Parallel sheets4=2 40 ==> Minimum beta sheet length=3 40 3. Number of Parallel sheets5=2 40 ==> Minimum beta sheet length=3 40 4. Number of Parallel sheets4=2 40 ==> Number of Parallel sheets3=2 40 5. Number of Parallel sheets3=2 40 ==> Number of Parallel sheets4=2 40

7. References

[1] Mukhopadhyay, A., Maulik,U., Bandyopadhyay,S., and Eils,R. , "Mining association rules from hiv-human protein interaction", In Proc.ICSMB, pages 344–348, 2010.

[2] C.Mondal K., Pasquier N., Mukhopadhyay A., daCosta PereiraC., Maulik U.and G.B.Tettamanzi A.,"Prediction of protein interactions on hiv-1 human ppi data using a novel closure-based integrated

approach", In Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms, pages164-173, 2012.

[3] Gene M. Ko, A. Srinivas Reddy, Sunil Kumar,and Rajni Garg, "Data Mining Analysis of HIV-1 Protease Crystal Structures" ACSESS proceedings 2009

[4] "Mining association rules for HIV-1 protease cleavage site Prediction ",Sandro da Silva Camargo, Paulo Martins Engel, WIM2006,pages:105-112

[5] A. Dubey, U. Chouhan, "A Computational Approach to Classify HIV Secondary Structure Of Enzymes", The Internet Journal of Medical Informatics. 2011 Volume 5 Number 2

[6] S. D. Bhavani and K.Suvarnavani, Somdatta Sinha. Mining of protein contact maps for protein fold prediction. WIREs Data Mining and Knowledge Discovery, JohnWiley & Sons, Volume 1, Pages 362-368, July/August 2011.

[7] Suvarna Vani K and S. D. Bhavani, SMOTE based Protein fold prediction classification, Advances in Computing and Information Technology 2013.

[8] http://www.cs.waikato.ac.nz/ml/weka

[9] http://www.rcsb.org

[10] Agrawal R, Srikant R (1994) Fast algorithms for mining association rules, In: Proceedings of the 20th VLDB conference, pp 487–499.

[11] Zaki M J, Nadimpally V, Bardhan D, Bystroff C: Predicting Protein Folding Mohaays. In Data Mining in Bioinformatics, Springer-Verlag London Ltd.:127–141, 2005