Available online at www.scholarsresearchlibrary.com

Scholars Research Library

J. Comput. Method. Mol. Design, 2011, 1 (2): 1-8
(http://scholarsresearchlibrary.com/archive.html)

ISSN : 2231- 3176
CODEN (USA): JCMMDA

Machine learning model for HIV1 and HIV2 enzyme secondary structure classification

Anubha Dubey* Bhaskar Pant* Usha Chouhan**

**Department of Bioinformatics, **Department of Mathematics*
MANIT, BHOPAL

ABSTRACT

The structure of a protein can reveal its function and its evolutionary history. Extracting this information requires knowledge of the structure and its relationship with other proteins. Secondary structures of protein are compact with helices and strands. Hence there is a need for development of computational techniques for prediction and classification of HIV-1and HIV-2 protein (enzymes) structures. In this paper a machine learning model has been developed for classification of alpha, beta and residues of HIV ribonuclease, HIV reverse transcriptase, protease, integrase, and these four types of HIV enzymes are present in HIV1 &HIV2 cycle. Various machine learning algorithms such as J48, Rotation Forest, and Random Forest have been used to classify alpha, beta and residues of HIV reverse transcriptase, protease, ribonuclease, integrase and model developed gives fair accuracy. The information generated from these models can be of great use in clinical applications.

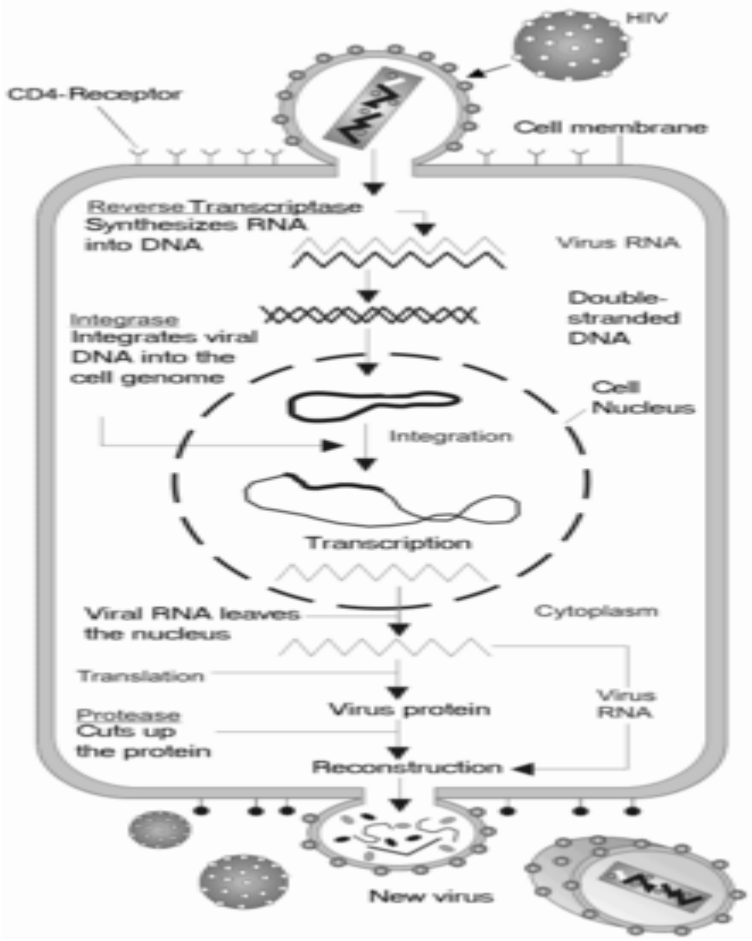
Keywords: Rotation Forest, J48, Ribonuclease, Integrase

INTRODUCTION

Human Immunodeficiency Virus (HIV) causes AIDS. HIV is of two types-HIV-1 & HIV-2. HIV is different in structure from other retroviruses. It is roughly spherical[1]with a diameter of about 120 nm, around 60 times smaller than a red blood cell, yet large for a virus [2]. It is composed of two copies of positive single-stranded RNA that codes for the virus's nine genes enclosed by a conical capsid composed of 2,000 copies of the viral protein p24 [3] The single-stranded RNA is tightly bound to nucleocapsid proteins, p7 and enzymes needed for the development of the virion such as reverse transcriptase, protease,ribonuclease, and integrase. A matrix composed of the viral protein p17 surrounds the capsid ensuring the integrity of the virion particle [4]. HIV enters macrophages and $CD4^+$ T cells by the adsorption of glycoproteins on its surface to receptors on the target cell followed by fusion of the viral envelope with the cell membrane and the release of the HIV capsid into the cell [5,6] After the viral capsid enters the cell, an enzyme called *reverse transcriptase* liberates the single-stranded (+)RNA genome from the attached viral proteins and

copies it into a complementary DNA(cDNA) molecule [7]. The process of reverse transcription is extremely error-prone, and the resulting mutations may cause drug resistance or allow the virus to evade the body's immune system. The reverse transcriptase also has ribonuclease activity that degrades the viral RNA during the synthesis of cDNA, as well as DNA-dependent DNA polymerase activity that creates a sense DNA from the *antisense* cDNA [8]. Together, the cDNA and its complement form a double-stranded viral DNA that is then transported into the cell nucleus. The integration of the viral DNA into the host cell's genome is carried out by another viral enzyme called *integrase* [7]. The final step of the viral cycle, assembly of new HIV-1 virons, begins at the plasma membrane of the host cell. During maturation, HIV proteases cleave the polyproteins into individual functional HIV proteins and enzymes. The various structural components then assemble to produce a mature HIV virion [9]. This cleavage step can be inhibited by protease inhibitors. The mature virus is then able to infect another cell. Enzymes made of proteins. Hence secondary structure play an important role.

Figure 1 shows replication cycle of HIV-1 &HIV-2.



Secondary structures of protein are compact with helices and strands. Hence there is a need for development of computational techniques for prediction and classification of HIV-1and HIV-2 protein (enzymes) structures. In this paper a machine learning model has been developed for classification of alpha, beta and residues of HIV ribonuclease, HIV reverse transcriptase, protease, integrase, and these four types of HIV enzymes are present in HIV1 &HIV2 cycle [16,17,18,19] as given in Figure1. .Various machine learning algorithms such as J48, Rotation Forest, and Random Forest have been used to classify alpha, beta and residues of HIV reverse transcriptase,

protease, ribonuclease, integrase and model developed gives fair accuracy. The information generated from these models can be of great use in clinical applications.

METHODS

Here the protein secondary structure data has been taken from PDB (Protein data bank) [13]of which the present work focuses on the further classification of according to alpha, beta and residue. Various algorithms of machine learning are available for classification and prediction of alpha, beta and residues. It has been developed using different algorithms of WEKA classifier [12]. Thus, for the same input they give different result and also differ in accuracy. This variation in result and accuracy leads to dilemma of choosing algorithm for classification and prediction of alpha, beta and residues. Classification using merely the predicted domain from the input sequence. From the various algorithms J48, Random Forest and Rotation Forest gives the better result with fair accuracies.

J48: A decision tree is a flowchart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. Internal nodes are denoted by rectangles, and leaf nodes are denoted by ovals. The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery **[13,14].**

Random forest (or **random forests**) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler, and "Random Forests" is their trademark. The term came from **random decision a forest that was** first proposed by Tin Kam Ho of Bell Labs in 1995. The method combines Breiman's "bagging" idea and the random selection of features, introduced independently by Ho and Amit and Geman in order to construct a collection of decision trees with controlled variation [13,14].

Rotation Forest: It is built with a set of decision trees. For each tree, the bootstrap samples extracted from the original training set are adopted to construct a new training set. Then the feature set of the new training set is randomly split into some subsets, which are transformed with a linear transformation method individually. Consequently, a full feature set is reconstructed with all the transformed features for each tree in the ensemble. Since a small rotation of axes may build a complete different tree, the diversity of the ensemble system can be guaranteed by the transformation. [15]

RESULT: To achieve our goal and develop our methodology we obtained the dataset from Protein Data Bank (PDB) for both HIV-1 & HIV-2 [13]. The following six cases arises for classification of HIV-1 & HIV-2 enzymes. PDB Classification according to HIV Reverse Transcriptase, HIV Protease, HIV ribonuclease by J48, Random forest, Rotation Forest will give the following results.

CASE1- All chains including alpha and beta

Algorithm	Accuracy	Average
Rotation Forest	93.0636%	0.931
J48	92.7746%	0.928
Random Forest	92.4855%	0.925

Rotation forest gives better results with accuracy 93.0636%

Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.981	0.643	0.945	0.981	0.963	0.868	hiv-1
0.357	0.019	0.625	0.357	0.455	0.868	hiv-2
Weighted Avg.	0.931	0.592	0.92	0.931	0.922	0.868

=== Confusion Matrix ===

a	b	<-- classified as
312	6	a = hiv-1
18	10	b = hiv-2

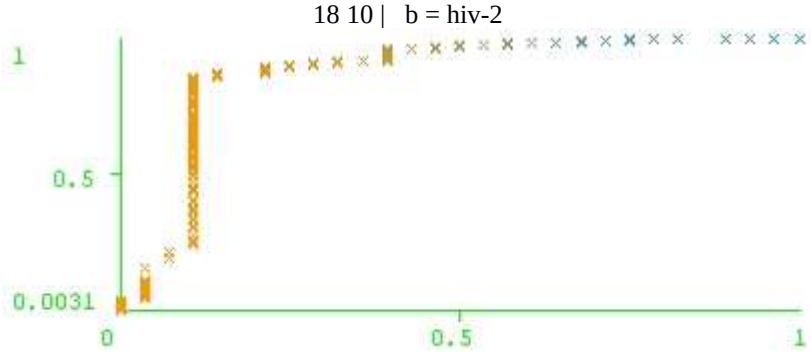


Figure 2: Shows ROC of all chains including alpha and beta.

CASE 2: All chains including alpha (without residues).

Algorithm	Accuracy	Average
Rotation Forest	91.6185%	0.916
J48	91.6185%	0.919
Random Forest	92.4855%	0.925

J48 predicts better results.

Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.997	0.893	0.927	0.997	0.961	0.844	hiv-1
0.107	0.003	0.75	0.107	0.188	0.844	hiv-2
Weighted Avg.	0.925	0.821	0.913	0.925	0.898	0.844

=== Confusion Matrix ===

a	b	<-- classified as
317	1	a = hiv-1
25	3	b = hiv-2

Figure3: Shows ROC of all chains including alpha (without residues).

Case 3: All alpha only (with residues)

Algorithm	Accuracy	Average
J48	90.7514%	0.908
Rotation Forest	91.3295%	0.913
Random Forest	92.4855%	0.931

Random Forest predicts better results as accuracy is 92.4855%.

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.969	0.5	0.957	0.969	0.963	0.815	hiv-1
0.5	0.031	0.583	0.5	0.538	0.815	hiv-2
Weighted Avg.	0.931	0.462	0.926	0.931	0.928	0.815

=== Confusion Matrix ===

a	b	<-- classified as
308	10	a = hiv-1
14	14	b = hiv-2

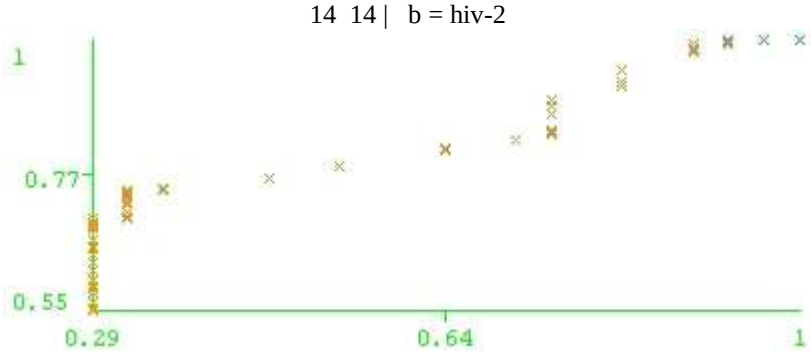


Figure 4: Shows ROC of all chains including alpha (without residues).

CASE 4: All betas without residues:

Algorithm	Accuracy	Average	Time taken(in second)
Rotation Forest	91.9075%	0.919	0.37 sec
J48	91.9075%	0.919	0.03 sec
Random Forest	92.1965%	0.922	0.09 sec

Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.994	0.893	0.927	0.994	0.959	0.691	hiv-1
0.107	0.006	0.6	0.107	0.182	0.691	hiv-2
Weighted Avg.	0.922	0.821	0.9	0.922	0.896	0.691

=== Confusion Matrix ===

a	b	<-- classified as
316	2	a = hiv-1
25	3	b = hiv-2

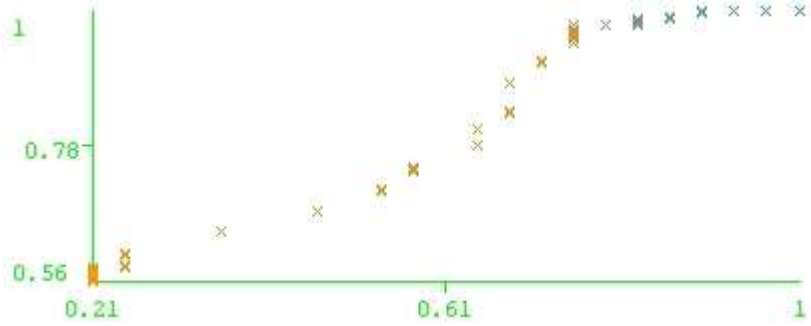


Figure 5. Shows ROC of all chains including beta (without residues).

CASE 5: All betas with residues:

Algorithm	Accuracy	Average
Rotation Forest	91.9075%	0.919
J48	91.9075%	0.919
Random Forest	90.1734%	0.902

Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	1	0.919	1	0.958	0.496	hiv-1
0	0	0	0	0	0.496	hiv-2
Weighted Avg.	0.919	0.919	0.845	0.919	0.88	0.496

=== Confusion Matrix ===

a	b	<-- classified as
318	0	a = hiv-1
28	0	b = hiv-2

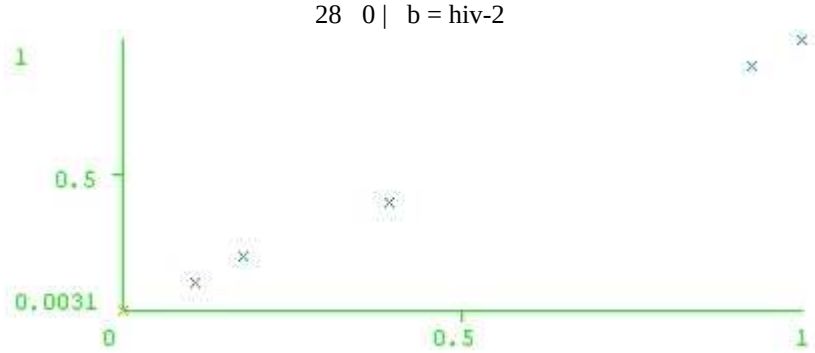


Figure 6. Shows ROC of all betas including beta with residues.

DISCUSSION

Rotation forest predicts better results because it works better with large datasets and generating classifier ensembles based on feature extraction as in case1 all chains including alpha and beta. Random Forest shows better results in single case 2,3,4 because it gives estimates of what variables are important in the classification. Receiver Operating Curve (ROC) is a graphical technique for evaluating data mining schemes. A ROC curve depicts the performance of a classifier without regard to class distribution or error costs. They plot the number of positives included in the samples on the vertical axis, expressed as a percentage of the total number of positives, against the total number of negatives on the horizontal axis. For each fold of a 10 fold cross validation, weight the instances for a selection of different cost ratios train the scheme on each weighted set, count the true positives and false positives in the test set, and plot the resulting point on the ROC axes. The ROC curves for different classes have been plotted as shown in Figures **(**2-6**)**. As ROC depicts the performance, we can refer from the confusion matrix that in case 1, the false positive ratio is 0.643, which clearly indicates that the true positive ratio is 0.981.In case 2, the false positive value is 0.893 and true positive is 0.981.Case3 shows false positives of 0.5 and true positives of 0.969. Case4 shows false positives 0.893 and true positives 0.994. Case5 shows false positives 1 and true positives 1. The accuracy of results for the five cases obtained from all the three classifiers with input as alpha or beta with chains as predicted from three different classifier and their comparison is presented in (Tables 1-5). In the case1 (see Table 1), when predicted alpha and beta from all the three classifiers are taken, the accuracy of is 93.0636% % is achieved through rotation forest. Case2: All alpha chain (without residues) random forest predicts better accuracy with 92.4855%. Case 3: All alpha chain with residues random forest predicts better accuracy with 92.4855%. Case4: All beta chain (without residues) random forest predicts better accuracy with 92.1965%. Case 5: All beta chain with residues J48 predicts better accuracy 91.9075% as rotation forest also predicts the same result but time taken by J48 is less. Hence Random Forest found suitable for case 2, 3, 4.

CONCLUSION

Among all the three classifiers, the classification of alpha, beta and alpha+beta with residues have five cases. So it is concluded that Random Forest found suitable for case 2, 3, 4. As it gives estimates of what variables are important in the classification. J48 predicts better result in case 5 as its speed is good and performs better calculation and has better memory. As more proteins have discovered the accuracy of the model is maintained and server is also developed. Database can also be redesigned to provide more scalable system. The challenge now is to organize these data in a way that evolutionary relationships between proteins can be uncovered and used to understand better protein function. Because protein structures are more highly conserved than are protein sequences, there is a growing interest in studying evolution based on an understanding of the protein structure space. The first steps common to the analysis of any large set of data are to group together data points that are similar, and then to identify connections between those elementary groups. These steps are usually performed with classification techniques. Hence structural classification of proteins leads to drug discovery and also helpful to biomedical scientists to develop protocols for identification of HIV.

Acknowledgement:
The authors are highly thankful to Department of biotechnology, New Delhi for providing Bioinformatics Infra Structures Facility at MANIT, Bhopal for carrying out this work.

REFERENCES

[1] Weiss RA (May **1993**). *Science* **260** (5112): 1273–9.doi:10.1126/Science.8493571.PMID 8493571
[2] Douek DC, Roederer M, Koup RA (**2009**). *Annu. Rev. Med.* **60**: 471–84. Doi: 10.1146/annurev.med.60.041807.123549 PMID 18947296.
[3] McGovern SL, Caselli E, Grigorieff N, Shoichet BK (**2002**). *J Med Chem* 45 (8): 1712–22.doi:10.1021/jm010533y.PMID 11931626
[4] Compared with overview in: Fisher, Bruce; Harvey, Richard P.; Champe, Pamela C. (**2007**). *Lippincott's Illustrated Reviews: Microbiology (Lippincott's Illustrated Reviews Series)*. Hagerstown, MD: Lippincott Williams & Wilkins. ISBN 0-7817-8215-5. Page 3
[5] Various (**2008**) (PDF).HIV Sequence Compendium **2008** Introduction.http://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/2008/frontmatter.pdf Retrieved 2009-03-31.
[6] Chan D, Kim P (**1998**). *Cell* **93** (5): 681–4. Doi: 10.1016/S0092-8674(00)81430-0. PMID9630213
[7] Wyatt R, Sodroski J (**1998**). *Science* **280** (5371): 1884–8.doi:10.1126/science.280.5371.1884.PMID9632381.
[8] Zheng, Y. H., Lovsin, N. and Peterlin, B. M. (**2005**). *Immunol. Lett.* **97** (2): 225–34. doi:10.1016/j.imlet.2004.11.026.PMID15752562
[9] Doc Kaiser's Microbiology Home Page>IV.VIRUSES>F.ANIMAL VIRUSLIFE CYCLE>3.The Life Cycle of HIV Community College of Baltimore County. Updated: Jan., 2008
[10] 11. Gelderblom, H. R (**1997**)."Fine structure of HIV and SIV". In Los Alamos National Laboratory (ed.) (PDF). *HIV Sequence Compendium*.Los Alamos, NewMexico: Los Alamos National Laboratory.pp. 31–44.
[11] http://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/1997/partIII/Gelderblom.pdf15.
[12] www.pdb.org

[13] http://www.cs.waikato.ac.nz/ml/weka
[14] Pang-Ning, Tan.M.Steinbach, V.Kumar. Introduction to Data Mining, **2008**. (s)
[15] Juan J, Rodr? Genz, Ludmila I.Kuncheva, Carlos J.Alonso "Rotation Forest: A New Classifier Ensemble Method" IEEE Transaction on Pattern Analysis and Machine Intelligence. October **2006** Vol 28, No. 10 pp1619-1630.
[16] B.Pant, K.Pant and K.R.Pardasani,"SVM classifier for classification of MMPs and ADAMs accepted for publication in ICMLC **2010**, Banglore.
[17] B.Pant,K.Pant and K.R.Pardasani, "DiRiboPred: A Web tool for Classification and Prediction of Ribonucleases, accepted for publication in *Global Journal of Computer Science and Technology*, University of Wisconsin,USA Vol 10. Issue 6 July-August **2010**.
[18] A.Dubey, B.Pant and Neeru Adlakha,"SVM Model for Amino Acid Composition based Classification of HIV1 Groups". *IEEE digital library* published.
[19] A.Dubey, B.Pant and Usha Chouhan," SVM Model for Classification of Structural and Regulatory Proteins of HIV1 and HIV2 is accepted for publication in *Journal of Bioinformatics Applications and research.*