A Computational Approach To Classify HIV Secondary Structure Of Enzymes

A Dubey, U Chouhan

Citation

A Dubey, U Chouhan. *A Computational Approach To Classify HIV Secondary Structure Of Enzymes*. The Internet Journal of Medical Informatics. 2009 Volume 5 Number 2.

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.

INTRODUCTION

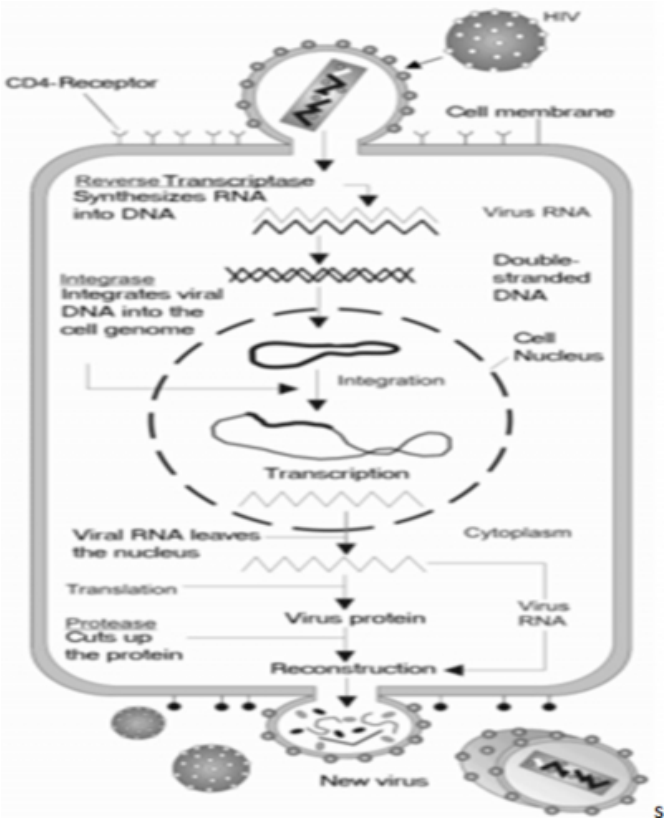
Human immunodeficiency virus (HIV) is a lentivirus (a member of the retrovirus family) that causes acquired immunodeficiency syndrome (AIDS) [1,2]. HIV is of two types-HIV-1 & HIV-2 HIV-1 is the virus that was initially discovered and termed both LAV and HTLV-III. It is more virulent, more infective, [3] and is the cause of the majority of HIV infections globally. The lower infectivity of HIV-2 compared to HIV-1 implies that fewer of those exposed to HIV-2 will be infected per exposure. Because of its relatively poor capacity for transmission, HIV-2 is largely confined to West Africa [4]. HIV is different in structure from other retroviruses. It is roughly spherical [5] with a diameter of about 120 nm, around 60 times smaller than a red blood cell, yet large for a virus [6]. 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 [7] 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 [8]. 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 [9,10] After HIV has bound to the target cell, the HIV RNA and various enzymes, including reverse transcriptase, integrase, ribonuclease, and protease, are injected into the cell during the microtubule based transport to the nucleus, the viral single strand RNA genome is transcribed into double strand DNA, which is then integrated into a host chromosome [11] 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 [12]. 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 [13]. 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 [14]. 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 [15]. 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 plays an important role.

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 [19,20,21,22] 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 and to understand HIV structure better. As these are the better drug targets.

Figure 1

Figure1: Replication cycle of HIV shows role of enzymes:

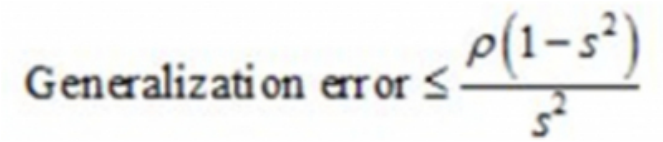


Method: Here the protein secondary structure data has been taken from PDB (Protein data bank) [15] of which the present work focuses on the further classification of according to alpha, beta and residue. Various algorithms of machine learning are available for classification and prediction of alpha, beta and residues. It has been developed using different algorithms of WEKA classifier [16]. 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 (or tree diagram) is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Another use of decision trees is as a descriptive means for calculating conditional probabilities. In data mining and machine learning, a decision tree is a predictive model; that is, a mapping from observations about an item to conclusions about its target value. More descriptive names for such tree models are classification tree (discrete outcome) or regression tree (continuous outcome). In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. The machine learning technique for inducing a decision tree from data is called decision tree learning, or (colloquially) decision trees [17].

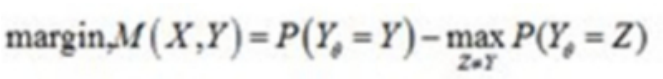
Random Forest is a class of ensemble method specially designed for decision tree classifiers .It combines the prediction made by multiple decision trees where each tree is generated based on the value of an independent set of random vectors .The random vectors are generated from a fixed probability distribution .Bagging using decision trees is a special case of random forests ,where randomness is injected into the model building process by randomly choosing N samples with replacement ,from the original training set. It has been theoretically proved that the upper bound for generalization error of random forests converges to the following expression when the number of trees is sufficiently large.

Figure 2



Where is the average correlation among the trees and s is a quantity that measures the strength of the tree classifier. The strength of a set of classifier refers to the average performance of the classifier where performance is measured probabilistically in terms of the classifier margin.

Figure 3



Where $Y_{}$ is the predicted class of X according to a classifier built from some random vector . The higher the margin is, the more likely it is that the classifier correctly predicts a given example X [17].

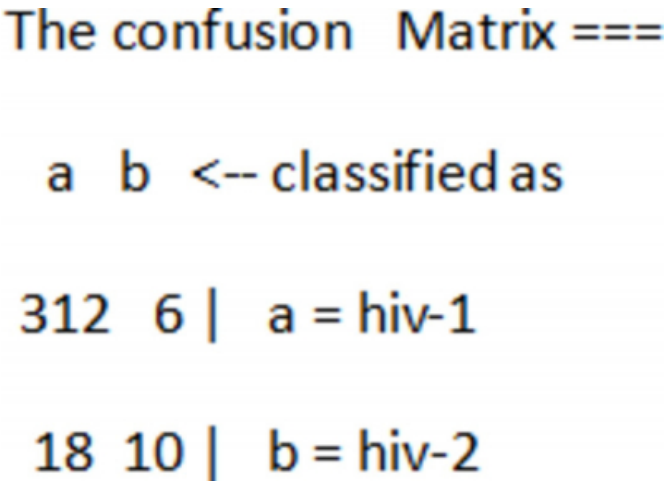
Rotation Forest: It is built with a set of decision trees. For each tree, the bootstrap samples extracted from the original training set are adopted to construct a new training set. Then the feature set of the new training set is randomly split into some subsets, which are transformed with a linear transformation 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.[18]

RESULT & DISCUSSION

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

The confusion matrix of alpha+beta of HIV-1 & HIV-2 generated from the above is given as under:

Figure 4



The Detailed Accuracy developed By Class is shown as-

Figure 5

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	

ROC: Receiver Operating Curve (ROC) is a graphical technique for evaluating data mining schemes. ROC curves 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 (1,2,). As ROC depicts the performance, we can refer from the confusion matrix that in case of HIV-1 class, the false positive ratio is 0.643, which clearly indicates that the true positive ratio is 0.981. The accuracy of results for these two classes obtained from all the three classifiers is presented as follows and accuracies of the above classifiers are also presented.

Figure 6

Figure 1: ROC of Rotation Forest

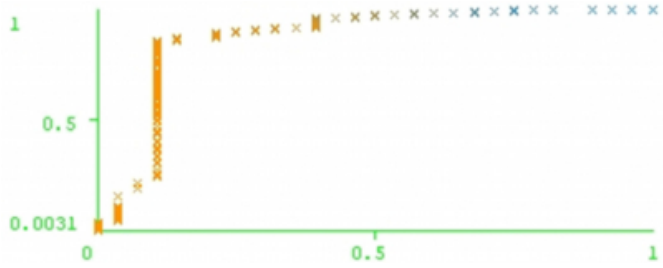


Figure 7

Figure 2: ROC OF Random forest

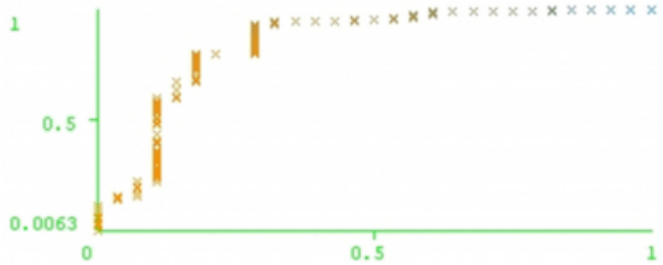
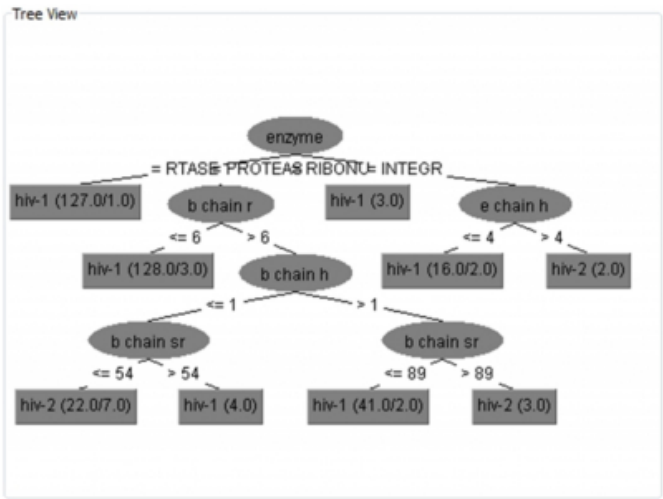


Figure 8

Figure 3: The J48 algorithm of Weka is used to obtain the tree view of classes as



The accuracy of results obtained by different algorithms is presented in Table -1

Figure 9

Algorithm	Accuracy
Rotation Forest	93.0636%
Random Forest	92.4855%
J48	92.7746%

Thus we observe that out of the 346 instances of HIV reverse transcriptase, protease, integrase, and ribonuclease taken for cross validation 322 were classified correctly whereas 24 were classified incorrectly by rotation forest classifier. This accounts to 93.0636 % accuracy which was the highest among all the three classifier used here so far. Thus the above classifier is able to classify HIV-1 and HIV-2 for which no algorithm has been reported in the literature so far. We can increase the instances by adding secondary structure data of other organisms like mouse, rat, pig and others but it does not give any significant change. This implies that the human instances are alone sufficient to develop the classifier. The reason is that similarity is 75-85% for enzyme structures among human and other organism. Hence inclusion of secondary structure data of other organisms will not only increase the instances but also increase the redundancy. The same model can be applied for organism like mouse, rat etc. for which secondary structure information is available in Protein Data Bank which is structure database of protein.

CONCLUSION

The above classifier takes into account the secondary structure of all the known 346 HIV enzymes as the rotation forest classifier performs the best among all the three classifiers, it qualifies as most suitable choice for classification and prediction. The authors wish to incorporate it as soon as more information is available in the future. The above model is useful for generating information which can be of great use in prediction of structure and function of all the enzyme structures present since they are key drug targets. The protein structure belonging to a particular class will have functional domains, alpha and beta sheet

corresponding to that class which will ease in locating the active site(s) as well as the binding site(s) in the classified protein and hence it can be the potential active site or binding site for the drug. As more structures of HIV enzymes are discovered the above classifier can be trained to improve the accuracy of results.

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). "How does HIV cause AIDS?". Science 260 (5112): 1273–9. doi 10.1126/science.8493571. PMID 8493571
2. Douek DC, Roederer M, Koup RA (2009).”Emerging concepts in the immunopathogenesis of AIDS”. Annu. Rev. Med. 60: 471–84. Doi: 10.1146/annurev.med.60.041807.123549 PMID 18947296.
3. Gilbert, PB et al; McKeague, IW; Eisen, G; Mullins, C; Guéye-Ndiaye, A; Mboup, S; Kanki, PJ (28 February 2003). "Comparison of HIV-1 and HIV-2 infectivity from a prospective cohort study in Senegal". Statistics in Medicine 22 (4): 573–593. Doi: 10.1002/sim.1342.PMID
4. Reeves, J. D. and Doms, R. W (2002). "Human Immunodeficiency Virus Type 2". J. Gen. Virol. 83 (Pt 6): 1253–65. Doi 10.1099/vir.0.18253-0PMID12029140.
5. McGovern SL, Caselli E, Grigorieff N, Shoichet BK (2002). "A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening". J Med Chem 45 (8): 1712–22. Doi:10.1021/jm0121/jm010533y. PMID 11931626.
6. 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
7. McGovern SL, Caselli E, Grigorieff N, Shoichet BK (2002). "A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening". J Med Chem 45 (8): 1712–22. Doi: 10.1021/jm010533y.PMID11931626.
8. 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
9. Various (2008) (PDF).HIV Sequence Compendium 2008 Introduction.http://www.hiv.lanl.gov/content/Sequence/HIV/Compendium/2008/frontmatter.pdf Retrieved 2009-03-31.
10. Chan D, Kim P (1998). "HIV entry and its inhibition". Cell 93 (5): 681–4.doi:10.1016/s0092-8674(00)81430-0.PMID 9630213.
11. Wyatt R, Sodroski J (1998). "The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens". Science 280 (5371): 1884–8. Doi:10.1126/science.280.5371.1884.PMID96322381.
12. Zheng, Y. H., Lovsin, N. and Peterlin, B. M. (2005). "Newly identified host factors modulate HIV replication". Immunol. Lett. 97 (2): 225–34. Doi: 10.1016/j.imlet.2004.11.026.PMID 15752562
13. 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
14. 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. http://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/1997/partIII/Gelderblom.pdf
15. Protein Data Bank
16. .http://www.cs.waikato.ac.nz/ml/weka
17. Pang-Ning, Tan.M.Steinbach, V.Kumar. Introduction to Data Mining, 2008. (s)
18. 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
19. B.Pant, K.Pant and K.R.Pardasani,”SVM classifier for classification of MMPs and ADAMs accepted for publication in ICMLC 2010, Banglore.
20. 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.
21. A.Dubey, B.Pant and Neeru Adlakha,”SVM Model for Amino Acid Composition based Classification of HIV1 Groups”. IEEE digital library published.

Author Information

Anubha Dubey
Research Scholar, Department of Bioinformatics

Usha Chouhan
Department of Mathematics