AIRO JOURNALS
Free / Unpaid
Peer Reviewed
Multidisciplinary
International

ISSN: 2320-3714
Volume 3 Issue 2
August 2023
Impact Factor: 10.1
Subject : Computer Science

AN EVALUATION OF THE HIV VECTOR MACHINE CLASSIFICATION STUDY

BHABANI SANKAR RATHA
Research Scholar
Dr Pratap Singh Patwal
GUIDE NAME

Abstract

The National Consortium for Biotechnology Information (NCBI), the European Molecular Biology Laboratory (EMBL), the DNA Databank of Japan (DDBJ), the SWISS-PROT, and other biological databases currently contain a wealth of information about the molecular sequences of bacteria, viruses, plants, and animals for analysis by scientists to produce the knowledge and information needed to solve problems with health care, food, energy, and the environment [97]. The databases are collections of DNA, RNA, and peptide sequences with single-letter codes for nucleotides or amino acids. Sequence names and comments are also permitted in the format before the sequences. Protein architectures, functions, cellular and subcellular localization, and interactions in various clinical and physiological situations can all be predicted by the examination of these sequences. Accurate computational models for the prediction and categorization of these sequences are required for the accurate prediction of the aforementioned features.

INTRODUCTION

Particularly infecting humans, the human immunodeficiency virus (HIV) depletes the body's ability to combat disease and infection by destroying key immune system cells. Its specialty is that it requires a host in order to procreate and grow. Hence the moniker HIV, a key factor in human AIDS. HIV1 and HIV2 are the two recognized forms of HIV. HIV1 and HIV2 are connected to one another. HIV-1 spreads more readily than HIV-2. Additionally, HIV1 causes the bulk of infections worldwide and mutates more quickly than HIV2.

The HIV RNA genome consists of nine genes (gag, pol, and env, tat, rev, nef, vif, vpr, vpu, and occasionally a tenth tev, which is a fusion of tat, env, and rev), encoding 19 proteins, and at least seven structural landmarks (LTR, TAR, RRE, PE, SLIP, CRS, and INS). Both HIV1 and HIV2 share three of these genes, gag, pol, and env, which carry the necessary instructions to create the structural proteins for new viral particles.

AIRO JOURNALS
Free / Unpaid
Peer Reviewed
Multidisciplinary
International

ISSN: 2320-3714
Volume 3 Issue 2
August 2023
Impact Factor: 10.1
Subject : Computer Science

There is an increasing need for data analysis as the amount and diversity of data available grow.

Utilizing computational or machine learning approaches, these data can be mined for knowledge. The two types of data analysis that can be used to derive models characterizing significant data classes or to forecast future data trends are classification and prediction. We may gain a better comprehension of the data as a whole via such study. It is clear from the literature review in the first chapter that no attempts have been made to establish models for the classification of HIV-1 and HIV-2. In the literature, a few models for predicting HIV and its targets have been published. The two distinct models for classifying HIV-1 and HIV-2 are presented in this chapter. The first model deals with categorizing HIV types 1 and 2 according to their amino acid and dipeptide compositions, while the second model addresses the structural and regulatory proteins of HIV types 1 and 2 according to their amino acid compositions.

HIV genome and structure: The structure of HIV is distinct from that of other retroviruses. It is generally spherical and has a diameter of 120 nm, which is huge for a virus but about 60 times smaller than a red blood cell. The virus's nine genes are encoded by two copies of positive single-stranded RNA, which are encased in a conical capsid made up of 2,000 copies of the viral protein p24. Reverse transcriptase, proteases, ribonucleases, and integrates are a few of the enzymes that are strongly attached to single-stranded RNA, along with p7 and nucleocapsid proteins. The integrity of the virion particle is maintained by a matrix made of the viral protein p17 that surrounds the capsid. The matrix is then encircled by the viral envelope, which is made up of two layers of phospholipids, which are fatty molecules that are extracted from the membrane of a human cell when a freshly created virus particle buds from the cell. Proteins from the host cell and around 70 copies of a complex HIV protein that extends through the virus particle's surface are both contained within the viral envelope. This protein, known as Env, is made up of three glycoprotein (gp) 120 molecules for the cap and three gp41 molecules for the stem, which anchor the structure to the viral envelope. The virus may bind to and fuse with target cells thanks to this glycoprotein complex, which starts the infectious cycle. Future HIV therapies or vaccines have been proposed to target both of these surface proteins, particularly gp120.

LITERATURE REVIEW

In 2008, Samir Lakhashe et al. investigated the HIV epidemic in India. Indian patients who with HIV have both neutralising antibody responses and HIV-specific CTL responses, according to the few immunological investigations done on them. For this investigation, they used CTL epitope mapping. Cathy H. Wu examined the functionally annotated protein sequences in the Protein Sequence Database in 2002. An integrated knowledge base system that is currently being built, along with a classification-driven and rule-based approach with evidence attribution, addresses the annotating issues. The method enables the differentiation of experimentally validated and computationally anticipated features, as well as sensitive identification, consistent and rich annotation, and systematic error detection in annotation. Using the p24 antigen test and viral load test for

ISSN: 2320-3714
Volume 3 Issue 2
August 2023
Impact Factor: 10.1
Subject : Computer Science

antiretroviral therapy, Jörg Schüpbach (2003) investigated HIV progression. These tests reveal a distinctive expression pattern shared by a number of well-characterized genes categorised according to pertinent biological roles.

A study on the use of machine learning to enhance the outcomes of high throughput docking against the HIV-1 protease by Naive Bayes classifier was published in 2004 by Glick M et al. Shegaki Mitaku et al. (2004) provide SOSUI with a transmembrane helices prediction programme to explore membrane proteins. Using hidden markov models, it can also identify the kind of transmembrane in membrane proteins with 80% accuracy.

A.Lumini and L.Nanni 2005 introduced hierarchical classifiers architecture which is a successful attempt to produce a dramatically error reduction with regard to the performance of linear classifiers. For more precise prediction of the HIV-1 protease cleavage site, this hierarchy is helpful. The miRNA systems of plants and animals were compared and contrasted in Anthony A. Millar's 2005 [93] study to determine whether there were any basic differences or variations. This sheds light on the study of miRNAs and their biological significance. Machine learning techniques for predicting linear B-cell epitopes on proteins were examined by Söllner J. and Mayer B. in 2006. This method combines a number of characteristics that have previously been linked to antigenicity with some unique measures based on amino acid frequencies and neighbourhood propensities. On the HIV epitope validation set, machine learning classifiers perform significantly better than the reference classification systems.

A hybrid prediction method for Gram-negative bacteria that combines a one-versus-one support vector machines (SVM) model and a structural homology approach was proposed in 2007 by Emily Chia-Yu Su and others. The SVM model consists of a variety of binary classifiers that combine biological traits generated from translocation pathways in Gram-negative bacteria.

Sebastien et al. (2008) looked into how the SVM can forecast HIV-1 co receptor usage with increased accuracy when it is fitted with the proper string kernel. Data Mining Learning Models and Algorithms for Medical Applications were researched by Plamena Andreeva in 2008 [5]. The Bayes classifier and SMOmodel exhibit the highest accuracy among the studied algorithms and classifiers in WEKA, and the best correctly categorised examples have been attained.

In 2009, Prosperi et al. investigated the relationships between clinical indicators and genetic characteristics of the entire HIV-1 envelope and viral tropism. The assessment of mutational co-variation was performed using bootstrapped hierarchical clustering. For the classification of X4 variations, various machine learning techniques, including logistic regression, SVM, decision trees, rule-based reasoning, and feature selection methods, as well as loss functions (accuracy, ROC curves, and f-measure), were used and compared. With 92.7% accuracy, the logistic regression model was created.

For the categorization of HIV-1 protease inhibitors using molecular structure, Hanbing Rao et al. 2009 examined machine learning algorithms including SVM, K-Nearest Neighbor (K-NN), artificial neural network (ANN), and logistic regression (LR). SVM exhibits superior generalisation capabilities

AIRO JOURNALS
Free / Unpaid
Peer Reviewed
Multidisciplinary
International

ISSN: 2320-3714
Volume 3 Issue 2
August 2023
Impact Factor: 10.1
Subject : Computer Science

and can be applied as a rapid filter substitute in the virtual screening of sizable chemical databases.

Objectives

- Model based on Support Vector Machines for HIV-1 and HIV-2 classification.
- Support Vector Machine to Classify HIV Based on Micro RNAs & G-Protein Coupled Receptors.
- Machine Learning Model for HIV-1 and HIV-2 Enzyme Classification Based on Structure.
- Machine Learning Model for HIV Membrane Proteins and Apoptosis Protein Prediction and Classification.

RESEARCH METHODOLOGY

Machine Learning Techniques and Data Mining

Large biological data sets present potential for data mining as well as difficult difficulties that require creative solutions. While traditional computer science methods have proven helpful in this sense, they are becoming less and less competent to handle many of the intriguing sequence analysis tasks. This is because biological systems are inherently complex as a result of evolution's tinkering, and because we don't have a complete idea of how life is structured at the molecular level. On the other hand, machine learning approaches (such as neural networks, hidden markov models, support vector machines, and belief networks) are perfectly suited for fields where there are lots of data, "noisy" patterns, and no overarching theories. These methods' underlying concepts involve automatically learning the theory from the data through a process of inference, model fitting, or learning from instances. As a result, they create a strong complement to traditional approaches. Because they require a lot of calculation, machine learning techniques tremendously profit from increases in computer speed.

The technical foundation for data mining is provided by machine learning. It is used to extract information that is expressed in comprehensible form from the raw data in databases, information that may be used for a number of applications. Arthur Samuel coined the term "machine learning" in 1959 to describe a field of study that allows computers to learn without being explicitly taught. Machine learning is a branch of science that deals with creating algorithms that let computers adapt behaviours based on real-world information, including sensor data or databases. Machine learning requires cross-disciplinary expertise in a number of fields, including probability theory, statistics, pattern recognition, cognitive science, data mining, adaptive control, computational neuroscience, and theoretical computer science, just like all other artificial intelligence-related fields do.

.CLASSIFICATION

On the basis of a training set of data that includes observations whose sub-population is known, statistical classification is the problem in statistics of determining the sub-population to which fresh observations belong when the identity of the sub-population is unknown. Therefore, it is necessary that fresh individual items be grouped according to quantitative data on one or more measures, qualities, or attributes (among other things) and according to the training set, where previously chosen groupings are already created.

AIRO JOURNALS
Free / Unpaid
Peer Reviewed
Multidisciplinary
International

ISSN: 2320-3714
Volume 3 Issue 2
August 2023
Impact Factor: 10.1
Subject : Computer Science

Bayes's simple probabilistic model

A conditional model can be compared to a classifier's probability model.

$$p(C \mid F_1, \ldots, F_n)$$

Several feature variables through over a dependent class variable with a limited number of outcomes or classes. The issue is that building such a model on probability tables is impractical when there are many features or when a feature can take on a lot of different values. To make the model more manageable, it is reformed.

VECTOR SUPPORT MACHINES

In order to learn separation functions for pattern recognition (classification) tasks and to do functional estimate in regression issues, Support Vector Machine (SVM) is a potentially useful idea. It is a statistical learning technique that was introduced in the Vapnik Chervonenkis dimension theory and structural risk minimization framework as unique learning techniques (VC).

SVMs are a group of connected supervised learning techniques used for regression and classification. Data must often be divided into training and testing sets when performing a classification task. There are numerous "attributes" and one "target value" (i.e. class labels) for each instance in the training set (i.e the features or observed variables). By using only the features of the test data, SVM attempts to create a model that predicts the target values of the test data.

ANALYSIS

Model I:A Hybrid Approach For Classification Of Hiv1 And Hiv2 On The Basis Of Amino Acid Composition And Dipeptide Composition

There are two types of HIV: HIV1 and HIV2. The majority of HIV infections worldwide are caused by HIV1, which is more aggressive and contagious. HIV-2 is primarily restricted to West Africa due to its comparatively limited ability for transmission. Due to its high mutagenecity and capacity for recombination, HIV1 has three types and eight subtypes. These factors explain why there isn't a publicly accessible HIV/AIDS vaccine or treatment at the moment. HAART, also known as highly active antiretroviral therapy, is the standard course of treatment for HIV infection. Since its inception in 1996, when the protease inhibitor-based HAART was first made available, this has been extremely helpful to numerous HIV-infected people.

Biomolecular researchers are working to create novel HIV vaccines and anti-retroviral medications. Since proteins are essential for life and are crucial for medication targeting and design, current research focuses on developing models based on the amino acid and dipeptide composition of HIV proteins. And this aids in the reformulation of antiretroviral medications.

The building components of proteins are amino acids. The percentage of all amino acids in a protein is known as the amino acid composition. Understanding the makeup of amino acids makes it easier to comprehend how proteins interact and function. Two amino acids are linked by a single peptide bond to form the molecule known as a dipeptide (e.g. ala–ala, ala–leu, and val–ser). A set pattern length of 400 is provided by the dipeptide composition. Dipeptide

AIRO JOURNALS
Free / Unpaid
Peer Reviewed
Multidisciplinary
International

ISSN: 2320-3714
Volume 3 Issue 2
August 2023
Impact Factor: 10.1
Subject : Computer Science

composition is frequently employed in the creation of fold identification techniques. A protein's dipeptide composition can give general information about the protein. The nucleotide sequence of the gene generating each protein determines the specific amino acid sequence that makes up that protein. The quantity of amino acids a produced protein contains as well as its total molecular mass, which is often expressed in units of dalton, can be used to determine the size of the protein.

According to the experimental endeavours, there are three types of HIV-1 depending on where they are found in the world. However, no computational method for classifying HIV1 and HIV2 based on other factors including dipeptide composition, amino acid composition, and physiochemical properties is documented in the literature.

It is necessary to create quick and precise computational algorithms for the prediction and classification of HIV1 and HIV2 sequences in order to better understand their mechanism of advancement because different groups of HIV1 and HIV2 proteins have overlapping patterns.

RESULTS

The three datasets were created by taking the amino acid composition and dipeptide compositions of the HIV-1 and HIV-2 proteins from the PROCOS software. These datasets were loaded into the SVM for HIV-1 and HIV-2 classification. The set of data that was utilized to convert protein length variations into fixed length patterns. RBF Kernel and Linear Kernel were used to categorise the HIV1 and HIV2 strains.

The outcomes obtained unequivocally demonstrate the significance of dipeptide composition in distinguishing HIV1 from HIV2. Consequently, this is a step toward supporting different wet lab procedures in developing new medications and therapeutic agents to combat these two. Here, the relationship between HIV-1 and HIV-2 and their amino acid and dipeptide compositions is studied. This information can help us understand these proteins better. Their physiological and molecular functions, as well as the substrate affinity, may be connected to the dipeptide composition. The dipeptide composition-based classifier had an overall accuracy of 99.54% and MCC 0.937 with RBF kernel for classifying HIV1 and HIV2. HIV1 has a classification accuracy of 97% and an MCC of 0.940. Because the training and testing datasets are frequently nearly similar when the amino acid composition parameter is used for classification, linear kernel identifies HIV1 instances with an accuracy of 95.89, and MCC is discovered to be infinite. According to RBF Kernel, HIV2 is classified with 99.43% accuracy, and MCC is 0.925. The total accuracy of HIV1 and HIV2 in the linear kernel is found to be 95.85 with MCC 0.85. It shown that HIV1 and HIV2 could be easily identified based on the amino acid and dipeptide composition. Table 4.2 demonstrates that RBF kernel outperforms linear kernel at predicting outcomes. The significance of RBF kernel in the classification of biological data is thus demonstrated by this work.

CONCLUSION

The investigation of a hybrid model for HIV1 and HIV2 classification based on amino acid and dipeptide composition reveals the connection between these two. The

ISSN: 2320-3714
Volume 3 Issue 2
August 2023
Impact Factor: 10.1
Subject : Computer Science

classification of HIV1 and HIV2 structural and regulatory proteins is also based on the amino acid content, demonstrating the significance of amino acids in proteins. These forecasts aid in the discovery of dipeptide motifs, interactions between domains, protein interactions, and protein folding. Since it offers a protein's whole information, using the network-based technique to increase prediction accuracy will also be taken into consideration. This demonstrates a step in the direction of helping different wet lab approaches develop innovative medicines and therapeutic agents against these two.

Several attempts to predict protein function have previously employed amino acid compositions as well as scant knowledge about sequence order. Because amino acid composition is straightforward, it is an effective characteristic for foretelling protein interactions. Any protein's amino acid composition reveals its chemical make-up, which in turn influences how putative biological functions for uncharacterized proteins should be predicted. its characteristics and purpose. This will provide fresh information for finding or identifying HIV medicines. More crucially, amino acid composition alone outperforms other, additional complicated features, showing the presence of sequence-level information. In other words, as more information about HIV-1 and HIV-2 protein sequences is provided in the future, the classifier's accuracy will continue to increase.

REFERENCES

1. Translocon-Assisted Membrane Protein Folding from the Stephen White laboratory at UC Irvine, (2008).
2. UNAIDS, WHO, "UNAIDS Report on the global AIDS epidemic", Pages 16-34, (2010).
3. Perrin L, Kaiser L, Yerly S. "Travel and the spread of HIV-1 genetic variants". *Lancet Infect Dis.* vol 3, no. 1 pages 22–27, (2003).
4. Ménétret JF, Hegde RS, Heinrich SU, Chandramouli P, Ludtke SJ, Rapoport TA, Akey CW, "Architecture of the ribosome-channel complex derived from native membranes", *J. Mol. Biol.* Vol. 348, pages 445-57, (2006).
5. Ménétret JF, Hegde RS, Heinrich SU, Chandramouli P, Ludtke SJ, Rapoport TA, Akey CW, "Architecture of the ribosome-channel complex derived from native membranes", *J. Mol. Biol.* Vol. 348, pages 445-57, (2006).
6. Gibas C. and Jambeck P., "Developing bioinformatics computer skills", O'Reilly Publication, ed. 1, pages 50-100, (2001).
7. Cluster Analysis: Basic Concepts and Algorithms, (2011).
8. Compared with overview in: Fisher, Bruce; Harvey, Richard P.; Champe, Pamela C. Lipagesincott's Illustrated Reviews: Microbiology (Lipagesincott's Illustrated Reviews Series). Hagerstown, MD: Lipagesincott Williams & Wilkins. ISBN 0-7817-8215-5. Page 3, (2007).

ISSN: 2320-3714
Volume 3 Issue 2
August 2023
Impact Factor: 10.1
Subject : Computer Science

9. Boisvert S, Marchand M., LavioletteF. and Corbeil J., "HIV-1 coreceptor usage prediction without multiple alignments: an applications of string kernels", *Retrovirology, vol.* 5, no.110, (2008).

10. Zazzi M et al. "Predicting response to antiretroviral treatment by machine learning", The EuResist project, *Intervirology*, vol.55, no.2, pages 123-127, (2012).

Author's Declaration

I as an author of the above research paper/article, hereby, declare that the content of this paper is prepared by me and if any person having copyright issue or patentor anything otherwise related to the content, I shall always be legally responsible for any issue. For the reason of invisibility of my research paper on the website/amendments/updates, I have resubmitted my paper for publication on the same date. If any data or information given by me is not correct, I shall always be legally responsible. With my whole responsibility legally and formally I have intimated the publisher (Publisher) that my paper has been checked by my guide (if any) or expert to make it sure that paper is technically right and there is no unaccepted plagiarism and hentriacontane is genuinely mine. If any issue arises related to Plagiarism /Guide Name /Educational Qualification /Designation /Address of my university/college/institution/Structure or Formatting/ Resubmission / Submission /Copyright / Patent/Submission for any higher degree or Job/Primary Data/Secondary Data Issues. I will be solely/entirely responsible for any legal issues. I have been informed that the most of the data from the website is invisible or shuffled or vanished from the data base due to some technical fault or hacking and therefore the process of resubmission is there for the scholars/students who finds trouble in getting their paper on the website. At the time of resubmission of my paper I take all the legal and formal responsibilities, If I hide or do not submit the copy of my original documents (Aadhar/Driving License/Any Identity Proof and Photo) in spite of demand from the publisher then my paper may be rejected or removed from the website anytime and may not be consider for verification. I accept the fact that as the content of this paper and the resubmission legal responsibilities and reasons are only mine then the Publisher (Airo International Journal/Airo National Research Journal) is never responsible. I also declare that if publisher finds any complication or error or anything hidden or implemented otherwise, my paper maybe removed from the website or the watermark of remark/actuality maybe mentioned on my paper. Even if anything is found illegal publisher may also take legal action against me

BHABANI SANKAR RATHA
Dr Pratap Singh Patwal