ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue III, March 2025

Support Vector Machine Implementation in the Analysis of HIV Species Based on Amino Acid and Dipeptide Sequences

[1]Dr. Kailash Chandra Nayak., [2]Dr. Bhabani Sankar Ratha

[1]Associate Professor, USBM, Bhubaneswar

[2]Associate Professor, NIIS Institute of Business Administration, Bhubaneswar

DOI : https://doi.org/10.51583/IJLTEMAS.2025.140300024

Received: 26 March 2025; Accepted: 29 March 2025; Published: 07 April 2025

Abstract: Infections are little irresistible specialists that can reproduce inside the cells in living life form. Human Immunodeficiency (HIV) is an infection that attacks human cells, debilitates the insusceptible framework and making individuals more vulnerable to disease. The favourable environment created by such virus allows grow of its content and rapidly increases within the hosts. HIV is a progressive sickness that causes a wide range of medical problems as a result of a widening array of bodily deficiencies and also considered as a lentivirus responsible for AIDS. The progressive breakdown of safe structures and the increase of dangerous breakthrough contamination characterise such human situation. In the study we used support vector machine (SVM) with Uniprot / Swiss-prot database to know the discrimination in HIV species based on Amino Acid and Dipeptide sequences to know the proportion and suggest the doctors to inject favourable protein to overcome the contamination. This also encourages the pharmaceutical industries to produce low cost medicines to cure of such diseases and also allow the patients to recover from diseases in low cost.

Keywords: Support Vector Machine, HIV, Insusceptible Framework, Progressive sickness

I. Introduction

Human immunodeficiency virus (HIV) particularly infect human beings by destroying important cells that fight disease and infection, hence makes body deficient. It has the specialty that it needs host to reproduce and multiply. Hence the name HIV is a major cause of AIDS in humans. At least two distinct strains of HIV have been identified i.e. Influenza A (HIV-1) and Influenza B (HIV-2). Both HIV1 and HIV2 are closely related. HIV-1 is easily transmitted than HIV-2. HIV1 also mutates more efficiently than HIV2 and causes majority of infection globally.

As the amount and variety of data available increases, computational or machine learning techniques are needed to extract this data for knowledge extraction. Different forms of information tests, such as alignment and prediction, can be used to rule out models that reflect crucial categories of data or to predict how that data will be used in the future, respectively. Such surveys help us to better understand general information. From the writing review in the primary part it is obvious that no endeavour is accounted for improvement of models for arrangement of HIV1 and HIV2 and the study model arrangements with the order of HIV1 and HIV2 based on amino corrosive creation and dipeptide piece. Amino acids are the building blocks of protein. These molecules come together to form proteins. Our body uses them for many important functions, such as making hormones, building muscle, and repairing tissue. Our body needs 20 amino acids to work properly and makes many of these amino acids on its own.

II. Research Model

is more destructive and highly contagious and is responsible for most HIV contamination internationally. Unfortunately, Due to its low rate of spread, HIV-2 is mostly a West African disease. HIV1 has 3 types and 8 subtypes due to its high mutagenicity and recombination. For these reasons, there are currently no publicly available HIV or help antibodies or treatments.

Biomolecular researchers are making strides toward an HIV vaccine and improved antiretroviral treatments. Research is currently focusing on constructing models based on amino acid and dipeptide composition to gain a deeper understanding of HIV proteins, as they are the fundamental units of life and play a crucial role in the targeting and design of drug discovery. Let me assure you, I am present. Because of this, antiretroviral medication is being prescribed again. Proteins cannot be made without amino acids. Only a small percentage of protein amino corrosives are synthesized. It is possible to gain a deeper comprehension of protein partnerships and abilities by concentrating on amino acid structures that are particularly susceptible to corrosion. A sub-atomic particle formed by the union of two amino acids. A sample size of 400 is sufficient when constructing a dipeptide. Overlay confirmation procedures are often improved with the help of dipeptide structures. As a means of providing global data about proteins, the dipeptide structure is one. When it comes to proteins, the sequence of amino acids is as unique as the groups of nucleotides used to code for them. The number of amino acids and the total subatomic mass, provided in Daltons, can be used to approximate the size of the bound protein.

The cover design of different clusters of HIV1 and HIV2 proteins creates a need to facilitate fast and accurate methods of computation of the expected value and order of HIV1 and HIV2 clusters to more easily comprehend their migration systems. These proteins are analysed by forming amino-corrosion products and dipeptides. Previously, these properties were used to treat other natural problems, such as GPCRs and harmful protein assembly, as disclosed in writing.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue III, March 2025

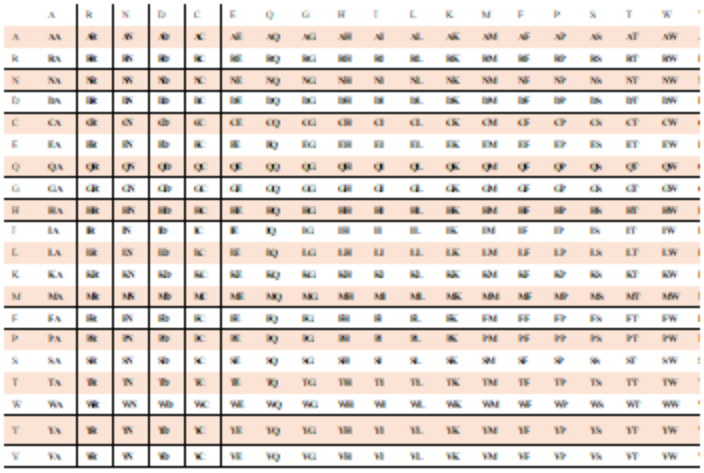

Table 1 the possible dipeptides with their symbolic representation

Amino-corrosion tissues embody data on some of the amino acids present in proteins, while dipeptides are used to represent global data on each protein assembly for degradation of these proteins. The data contained in the protein's variable essential amino-corrosion structures are converted into high-value vectors of the correct length expected for SVM, where they participate in HIV protein assembly.

One problem with protein characterization is that proteins vary in length. Amino-corrosion moieties and dipeptide synthesis provide protein data as 20-sided and 400-sided suitable vectors.

III. Data Collection and Implementation

Data: The data of HIV1 and HIV2 proteins in the form of amino acids in FASTA format is taken from Uniprot/Swiss-prot database. All the redundancies in the dataset were removed.

Dataset 1: It consists of proteins derived from HIV1, 356 were HIV1-positive cases and 356 were HIV-negative cases belonging to another enzyme group.

Dataset 2: It consists of proteins of the HIV2 group.150 positive cases of HIV2 and other enzyme group 150 proteins were considered negative.

Dataset 3: It consists of the proteins of HIV1 and HIV2. The total instances were 506 which are positive instances and 506 were protein sequences of enzymes taken as negative.

Protein sequence—Information about a protein can be encoded as a 20-dimensional vector based on its amino acid sequence. Hua and Sun previously used this method to predict the subcellular localization of proteins. The proportion of different types of amino acids in a protein is known as its amino acid composition. The relative abundances of the 20 essential amino acids were calculated using Equation 1.

$$\text{Fraction of } aai = \frac{\text{total number of amino acids of type (i)}}{\text{total number of amino acids in protein}} \qquad (1)$$

Where i is a specific type of amino acid (AA).

Dipeptide conjugates have been used to change proteins of varying lengths into transgenic vectors of a constant size. Some researchers have tried using dipeptide compositions to create fold detection methods in the past, and it has been useful. Because of the nature of the dipeptide composition, the length of the resulting patterns was always 400. The relative abundance and local sequence of amino acids are encoded in dipeptide structures. Equation 2 was used to determine this.

$$\text{Fraction of } \text{dep (i)} = \frac{\text{total number of dep (i)}}{\text{total number of all possible dipeptides}} \qquad (2)$$

Where dep (i) is one dipeptide i of 400 dipeptides.

PROCOS: For the aforementioned data set, the amino acid and dipeptide compositions were pulled from the PROCOS service. There are 450 vectors in total to determine the amino acid and dipeptide content of protein sequences of varying lengths with ease. As a result of using this program, we can avoid waste of time or effort to structure hybrids for use with a classifier (support vector machine).

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue III, March 2025

SVMs:

We use large-margin separation and kernel functions to find solutions to classification difficulties. Separation by a considerable margin is inspired by the two-dimensional point-classification scheme. This hunch is right on, as it captures the essence of the mathematically stated concept of large-margin separation. Support vector machines are useful for classifying biological data for two reasons (protein sequences):

First, SVM is known to perform well relative to other statistical or machine learning techniques, and many biological problems have highly dimensional noisy data. Second, kernel techniques such as support vector machines can readily deal with non-vector inputs, including variable-length sequences and graphs. These records are frequently used in computational biology and bioinformatics.

The process of cross-validating datasets is carried out in this lab. The proteins in a dataset are randomly split into M equal-sized groups in cross-validation. Proteins are used for training or developing methods, which are then applied to the remaining N/M proteins. This is done an infinite number of times (M) until every possible combination of elements has been tried. Here, we use 5-fold cross-validation to assess the accuracy of classifiers that use amino acid and dipeptide composition as features. [Measurement of HIV-1 and HIV-2 Proteins] Standard threshold-dependent measures like sensitivity, specificity, accuracy, and Mathieu's correlation coefficient [MCC] were used to evaluate the first-level classifiers' performance is analysed since Lib SVM offers a tool for picking parameters based on RBF kernels.

Grid-based cross-checking of data:

LibSVM-in-built Tools' modules were used to conduct grid searches in C and gamma. The best C and gamma pair is selected based on its performance in cross-validation.

Prediction system analysis:

Specimens that actually were positive and those that actually were negative were determined to be such. In cases of false-positive results, previously negative samples were incorrectly classified as positive. Incorrectly identifying a positive sample as negative is a false negative. Sensitivity, specificity, total accuracy, and the Matthews correlation coefficient [MCC] were used to evaluate the predictive performance.

In machine learning, the Matthews correlation coefficient is a useful metric for evaluating the accuracy of two-class (binary) classifications. It is considered a fair measure because it takes into account both good and negative results and may be used with a wide range of class sizes. To put it simply, MCC is the correlation coefficient between the actual and anticipated binary classifications equates to a value between 1 and +1. In this context, a coefficient of +1 denotes a perfectly accurate forecast, 0 indicates that the prediction is not random, and -1 implies an absolutely opposite result. A statistical metric that can be used to approximate prediction accuracy is the Matthews correlation coefficient (MCC). In general, higher MCC values correspond to more accurate forecasts.

IV. Results and Discussions

Three datasets were constructed by pulling information from the PROCOS program about the amino acid and dipeptide compositions of HIV-1 and HIV-2 proteins. The SVM was fed with these datasets to classify HIV-1 from HIV-2. Information utilized in the process of translating proteins of varying lengths into sequences of known length. Both RBF and linear kernels were used to categorize HIV1 and HIV2.

Sl.No.	Method	RBF Kernel		Linear kernel	
		ACC	MCC	ACC	MCC
Dataset 1	HIV1	97%	0.940	95.89%	INFINITE
Dataset 2	HIV2	99.43%	0.925	98.89%	0.9796
Dataset 3	All positives HIV1 & 2	99.54%	0.937	95.85%	0.85

Table 2 the accuracies of HIV1 & HIV2 in both RBF kernel and linear kernel

Table 2 displays the results of an RBF and linear kernel comparison based on datasets 1, 2, and 3. The observed results demonstrate the significance of dipeptide composition in discriminating HIV1 and HIV2. As a result, there is a call for a greater use of wet lab techniques in the discovery of new treatments for these conditions. In this work, we explore how differences in the amino acid and dipeptide compositions of HIV1 and HIV2 may shed light on their functions. Dipeptide composition may be related to molecular and physiological roles as well as substrate affinity. With an MCC of 0.937, RBF kernel, the overall accuracy of the dipeptide composition-based classifier for HIV1 and HIV2 classification was 99.54 percent. HIV1 classifies with 97% accuracy and MCC of 0.940. Linear kernel classifies HIV1 instances with an accuracy of 95.89 and MCC is found to be infinite because as the amino acid composition parameter is taken for classification, training and testing datasets are almost identical in

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue III, March 2025

some cases. HIV2 classifies with 99.43% and MCC is 0.925 as shown by RBF Kernel. In linear kernel the overall accuracy of HIV1 and HIV2 are found to be 95.85 with MCC 0.85. HIV1 and HIV2 have been shown to be readily distinguishable based on amino acid and dipeptide composition. Thus table 2 shows that RBF kernel predicts better results as compared to linear kernel. Therefore this study proves the importance of RBF kernel in biological data classification.

The establishment of such methods will accelerate the identification of HIV types and subtypes. Amino acid and dipeptide compositions provide global information for proteins to facilitate drug discovery in inflammatory diseases, osteoporosis or HIV, proteomics, and systems biology.

V. Conclusion

Protein function has traditionally been predicted using amino acid composition and rudimentary sequence order information. Protein interaction prediction can benefit greatly from the relative simplicity of the amino acid composition. Predicting the likely biological role of an uncharacterized protein is based on the amino acid composition of the protein, which provides a clue to the molecule's chemical nature. This sheds new light on the search for HIV treatments. The sequence level information can be inferred from the amino acid composition alone, and there are other existing, more sophisticated properties that make this inference possible. This indicates that as more protein sequence data for HIV1 and HIV2 is collected, the classifier will be improved.

Machine learning algorithms like SVM and classifier can prove to be an excellent way for functional as well as structural annotation of newly discovered proteins. Many of the so called orphan proteins of any class can be designated as belonging to it, if rigorous testing will be done by various parameters and various kernels. With the detailed and quick analysis helps the pharmaceutical industries to produce low cost medicines to cure of such diseases and also allow the patients to recover from diseases in low cost.

Similarly the genomic data can be annotated with its various characteristics studied and analysed in the entire manner. It poses a very bright picture that in the future we will be able to unravel the mystery of life with all the macromolecules deciphered for their structure and function.

Bibliography

1. Ardekani A.M. and Naeini M.M. “The role of miRNA in human diseases”, Avicenna J Med Biotech, vol 2, no.4, pages 161-179, (2010).
2. Asa Ben- Hur, Cheng Soon Ong, “Support Vector Machines and Kernels for Computational Biology”. 2009.
3. Bickel S, Bogojeska J, Lengauer J, Schiffer T., “Multi-Task Learning for HIV Therapy Screening”, Appearing in Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, (2010).
4. Carr JK, Foley BT, Leitner T, Salminen M, Korber B, McCutchan F, "Reference sequences representing the principal genetic diversity of HIV-1 in the pandemic”. In Los Alamos National Laboratory (ed.), HIV sequence compendium, pp. 10–19, (1998).
5. Caruana, R. and Niculescu-Mizil, A.: "An empirical comparison of supervised learning algorithms", Proceedings of the 23rd international conference on Machine learning, (2006).
6. Chang, C.-C., Hsu C.W. & Lin, C.-J., “A practical guide to support vector classification”, 2010.
7. Dolin, [edited by] Gerald L. Mandell, John E. Bennett, Raphael. Mandell, Douglas, and Bennett's principles and practice of infectious diseases (7th ed.). Philadelphia, PA: Churchill Livingstone/Elsevier. pp. Chapter 169, (2010).
8. Hanbing Rao, Guobing Yang, Ningxin Tan, Ping Li, Zerong Li, Xiangyuan Li, “Prediction of HIV-1 Protease Inhibitors Using Machine Learning Approaches”, QSAR & Combinatorial Science, vol. 28, Issue 11-12, pages 1346–1357, (2009).
9. Heider, D**,** Verheyen J. and Hoffmann D. Machine learning on normalized protein sequences”, *BMC Research Notes*, vol. 4**,** no.94 (2011).
10. Keele, B.; Jones, J.; Terio, K.; Estes, J.; Rudicell, R.; Wilson, M.; Li, Y.; Learn, G. et al. "Increased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz". Nature, vol.460, no.7254, pages 515–519, (2009).
11. Klon AE, Glick M and Davies JW, “Applications of machine learning to improve the results of high-throughput docking against the HIV-1 protease”, J Chem Inf Comput Sci., Vol.44, no.6, pages 2216-2224, (2010).
12. Kitahata MM, Gange SJ, Abraham AG, et al., "Effect of early versus deferred antiretroviral therapy for HIV on survival". N. Engl. J. Med. Vol.360, no. 18, pages 1815–26. (2009),
13. Lakhashe S.,Thakar M,Godbole S, Tripathy S. and Paranjape R. ,” HIV Infection in India: Epidemiology and Pathogenesis.” National AIDS Research Institute Pune, INDIA, Journal of Biosciences, vol. 33, no. pages515-25. (2008).
14. Los Alamos National Laboratory Established 1943. "Fighting the world’s most dangerous disease, Los Alamos Lab". Lanl.gov. (2010).
15. Philip E.Bourne, Helge Weissig, San Diego, Structural bioinformatics, University of California, San Diego. Wiley-liss publication, (2009).
16. Plantier JC, Leoz M, Dickerson JE, Oliveira F.D., Cordonnier F., Véronique Lemée V., Damond F., Robertson D.L., & Simon F., "A new human immunodeficiency virus derived from gorillas", Nature Medicine, vol.15, no.8, pages 871–872, (2009).

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue III, March 2025

17. Prosperi MC, "Robust supervised and unsupervised statistical learning for HIV type 1 coreceptor usage analysis", AIDS Res Human retroviruses, vol.25, no.3, pages 305-14, (2009).
18. Ratha B.S. and Patwal P. S., "An Evaluation of the HIV Vector Machine Classification Study", Airo International Journal, Vol. 3, Issue 2, pages 570-577(2023).
19. Ratha B.S. and Patwal P. S., "An Examination of the Effects of HIV on Permanent System", Airo National Journal, Vol. 3, Issue 2, pages 140-145(2023).
20. Rishishwar L., Mishra N., Pant B., Pant K., and Pardasani K.R., "ProCoS: Protein composition server", Bioinformation, vol.5, no.5, 227 (2010)
21. Septiembre J., "Transmembrane glycoproteins cross reactive HIV-1/HIV-2 epitope", Rev Biomed, vol.11, No. 3, pages 155-160, (2000).
22. Sharp, PM; Hahn, BH, "Origins of HIV and the AIDS Pandemic", Cold Spring Harbor perspectives in medicine 1 (1): a006841, (2011).
23. Singh Y. and Mars M. "Support vector machines to forecast changes in CD4 count of HIV-1 positive patients", Scientific Research and Essays, vol. 5, no.17, pages 2384-2390, (2010).
24. Söllner J, Mayer B, "Machine learning approaches for prediction of linear B-cell epitopes on proteins", J Mol Recognition, vol.19, no.3, pages 200-208, (2006).
25. Song C. and Shi F., "Prediction of Subcellular Localization of Apoptosis Proteins by Dipeptide Composition", International Journal of Digital content Technology and its Applications, vol. 4, no. 1, pages 32-36, (2010).
26. Tierney, A.J., "Structure and function of invertebrate 5-HT receptors: a review" Comp. Biochem. Physiol. A Mol. Integr. Physiol., vol.128, pages 791–804, (2001).
27. Translocon-Assisted Membrane Protein Folding from the Stephen White laboratory at UC Irvine, (2008).
28. White S.H. and Heigne G.V. "How translocon selects transmembrane helices?", Annual review of Biophysics, vol. 37, pages 23-42, (2008).
29. Yavuz O. and Ozyilmaz L., "Analysis and Classification of Hiv-1 Sub-Type Viruses by AR Model through Artificial Neural Networks", World Academy of Science, Engineering and Technology, vol. 49, pages 826-831, (2009).
30. Zazzi M et al. "Predicting response to antiretroviral treatment by machine learning", The EuResist project, Intervirology, vol.55, no.2, pages 123-127, (2012).
31. http:// www.pdb.org
32. http://bioweb.pasteur.fr/seqanal/interfaces/topagesred.html
33. http://www.cbs.dtu.dk/services/TMHMM/or
34. http://www.cbs.dtu.dk/services/TMHMM-2.0/
35. http://www.dtreg.com/svm.html
36. http://www.csbio.sjtu.edu.cn/bioinf/virus
37. http://www.niaid.nih.gov
38. http://www.uniprot.org
39. http://www.misangar.ac.uk/ microRNA registry
40. http://www.yale.edu/bio243/HIV/genome.html
41. http://i_base.info/qa/36