



# A Study On Applications Of Machine Learning In The Classification Of Hiv Structural And Regulatory Proteins Based On Amino Acid

<sup>1</sup>Dr. Kailash Chandra Nayak

<sup>1</sup>Associate Professor, USBM, Bhubaneswar

<sup>2</sup>Dr. Bhabani Sankar Ratha

<sup>2</sup>Associate Professor, NIIS Institute of Business Administration, Bhubaneswar

## Abstract:

Identification of diseases and the strategies to be made to overcome it are the medicinal research to bring a healthy society and where people can find good doctors around them and also avail low cost medical services for a long life. It can be achieved only through technological advancement and its availability to the society. Because technology can detect the disease in a short time and find the composition of antibodies in a cell. Basing on the identification, doctors can inject needful medicine to a patient to bring healthy in a short time. The use of machine learning model with Support Vector Machine (SVM) helps to detect the HIV disease in the human body with the variation in amino acid compositions in the cells and tries to enhance the protein level to make the cell active and to fight against the infection. The study finds the different structural and regulatory proteins based on amino acid compositions in a quick time and provide a positive response for doctors to cure the patients.

**Keywords:** SVM, Machine Learning, HIV, Structural protein, regulatory protein

## Vector-based support system for binary classification:

### Support Vector Machine:

Machine learning techniques are supervised learning algorithms that correctly predict and classify biometric data. Such challenges are well-suited to the application of support vector machines and associated kernel approaches. The decision levels at the heart of support vector machines establish the limits within which decisions can be made. SVM classification relies heavily on large-margin classification and kernel operations.

Predictions are at the root of many of computational biology's difficulties. Effective prediction and classification of biometric data can be achieved through the use of machine learning techniques, which are supervised learning algorithms. Such challenges are well-suited to the application of support vector machines and associated kernel approaches. The decision levels at the heart of support vector machines establish the limits within which decisions can be made. A decision level is a hierarchical partition that distinguishes between classes of items. Whether an item is (+1) positive or (-1) negative determines whether it is a plus or minus, making the distinction between the two items clear. Kernel functions and large-margin classification both play crucial roles in SVM classification. Regularisation, RBF kernel, and linear kernel parameters were all optimised during SVM training.

The performance of each classifier was assessed using 5-fold cross-validation. The dataset was replicated five times, each time at random, for 5-fold cross-validation. Four different data sets were used to teach each classifier, and just one was used to test it. To assess the classifier's efficacy, we computed the accuracy and MCC for each nuclear receptor subfamily, as explained in detail by Hua and Sun and illustrated by Equation 1. True positive (tp) and false negative (tn) designations were made for the samples. A false positive (fp) occurs when a positive result is reported for a sample that should have tested negative. When a sample that should have tested positive was incorrectly labelled as negative, a false negative occurred. Several metrics, such as the Matthews correlation coefficient (MCC), the specificity (equation 1), and the accuracy (equation 2), were used to assess the model's predictive ability (Equation 3). Using a two-class example, MCC shows how the predicted a-class relates to the actual a-class. It can be used to measure the efficacy of a classifier regardless of how evenly distributed the classes are.

$$\text{Sensitivity} = \frac{tp}{tp + fn} \quad (1)$$

$$\text{Specificity} = \frac{tn}{tn + fp} \quad (2)$$

$$\text{MCC} = \frac{tp \times tn - fp \times fn}{\{(tp + fp)(tn + fp)(tn + fp)(tn + fn)\}^{1/2}} \quad (3)$$

### **Classification of HIV1 and HIV2 structural and regulatory proteins based on amino acid composition.**

The human immunodeficiency virus (HIV) genome consists of 9,741 base pairs distributed over 9 genes. There are 19 proteins encoded by the genes present in all retroviruses (gag, pol, env, tat, rev, vif, vpr, vpu, and nef).

All retroviruses contain genes for the Gag category of antigens. Produces a number of viral defence proteins. Matrix (MA), Capsid (CA), and Nucleocapsid (NC) are the three components of HIV (nucleocapsid). About 2000 base pairs are enlisted by the capsid protein.

The Pol polymerase gene is shared by all retroviruses and provides the viral replication enzymes. PR (protease), IN (endonuclease), and RT make up the three components of HIV (reverse transcriptase).

All retroviruses have the Env envelope gene. Produces a protein coat for viruses. There are two components to HIV. Surface Umbrella (gp120) and Transmembrane (TM) (transmembrane envelope, gp41).

The activity of far-flung genes can be altered by tran-activated genes. Controls the ability of all HIV proteins to transactivate.

- A gene that controls the production of the viral protein Rev.

Cell-free viral infectivity is dependent on the Vif virus infectious factor gene. Nef Is a Misunderstood Amplifier, Often Mislabelled as a Negative Regulator

- Vpr (viral protein R): Not much is known about the gene encoding the viral protein R.
- Vpu (viral protein U): There can be no successful viral replication or shedding without the viral protein U gene. Specifically in HIV-1.
- The Vpx virus has a protein X gene, but nobody knows what it does. Exclusive to HIV type 2 and SIV.

Three of these genes, namely gag, pol, and env, encode structural proteins for HIV1 and HIV2 viral particles. A protein termed gp160 is encoded by the env gene and is broken into gp120 and gp41 by cellular proteases. The ability of HIV to infect cells and multiply itself is regulated by the remaining six genes: tat, rev, nef, vif, vpr, and vpu (or vpx for HIV-2) (replicate). Gene that acts as a regulator or helper for the protein it regulates or that causes disease.

Since HIV-1 and HIV-2 are distinguished from one another by the amino acid makeup of their respective structural and regulatory proteins, these factors are used to categorize the two viruses. Protein function can be studied in great detail by varying the concentration of various amino acids, which are the proteins' building blocks. There are a number of benefits to amino acid composition (AAC). B. They are simple to implement, useful for predicting protein interactions, computationally inexpensive when dealing with large protein sequences from higher organisms, and easy to apply in general.

In this part, we introduce the support vector machine (SVM) models we've created to predict, identify, and classify structural and regulatory proteins in HIV1 and HIV2 based on their amino acid composition.

## Model and Method

**DATA:** From the UNIPROT/SWISS-PROT database, we obtain the protein sequence data in FASTA format. It might be more productive to eliminate unnecessary repetition. The following data sets were created so that structural and regulatory proteins in HIV1 and HIV2 may be studied in greater depth.

Dataset 1: It contains all of HIV-1 and HIV-2's structural and regulatory proteins. Dataset entries that were flagged as fragments were omitted. There were 502 total cases, with 430 instances chosen for structural

proteins and 72 instances chosen for regulatory proteins. There were 502 HIV-1 and HIV-2 positive patients and 502 HIV-1 and HIV-2 negative cases.

Included in dataset 2 are all protein occurrences outside of the HIV family. Fifty-two cases in total.

HIV-1 and HIV-2 structural proteins are fully described in dataset 3. The data collection excludes any records that have been flagged as fragmentary. There are a total of 430 hits, with 430 positive hits corresponding to structural proteins and 430 negative hits corresponding to another class of enzymes.

The fourth dataset includes both the HIV-1 and HIV-2 regulatory proteins. Fragmented entries were excluded from the analysis. There were 72 regulatory proteins that were considered positive and belonged to one enzyme family, and 72 that were considered negative and belonged to another.

The Procos server facilitates protein synthesis. The amino acid profiles used in the aforementioned dataset were retrieved from PROCOS. The 450 input vectors were created in an SVM like format. As a result of using this program, you won't have to waste time or effort structuring hybrids for use with a classifier (support vector machine).

Machines that use support vectors: To classify data that is linearly inseparable, it uses a hyperplane and a kernel function, making it ideal for usage in the field of biology (further details are given in previous section).

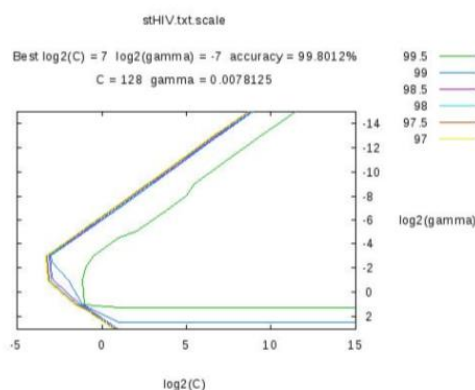


**Figure 1 Grid search across instances of HIV1 and HIV2 structural and regulatory proteins.**

Measurement of Success: Both structural and regulatory protein sequences are taken into account in the training set. Classification can be accomplished with the help of support vector machines (binary classification).  $C$  and  $\gamma$  are two parameters for the RBF kernel (radial basis functions:  $\exp(-\gamma \cdot \|u-v\|^2)$ ). No one can say in advance which values of  $C$  and  $\gamma$  will yield the best results. Because of this, it is necessary to undertake some model selection (parameter search). Find optimal parameters such that the classifier can make reliable predictions about previously unknown information (that is, test data). Some of the characteristics

that go into feature vectors are converted to binary form, and these can include things like the amino acid content of proteins. Each of these feature vectors is built for each of the 100 accessible instances, making the total feature vector size 100 x 5. Ten-fold cross-validation is used in conjunction with a grid search on  $C$  and  $\gamma$  to get reliable model predictions. Grid searching was used on one of the training sets to determine the best values for  $C$  and  $\gamma$  in the RBF kernel function (see Figures 1, 2, 3, and 4).

With just two options, yes and no, one prediction has four distinct outcomes. Positive, Negative, Positive, and False Negative. The right categories are "true positives" and "true negatives." The overall success rate

is calculated by taking the number of right classifications and dividing by the total number of possible ones.

### Prediction system analysis:

This is calculated using various accuracy metrics, including the proportion of correct diagnoses, the number of missed diagnoses, the rate of false positives, the rate of false negatives, the specificity, sensitivity, and precision of the test (details are described in the previous section).

### Discussion of the Survey's Findings

HIV-1 and HIV-2 are distinct because of their different geographic ranges and the considerable recombination and mutation that has occurred between them. Because of this, HIV-1 and HIV-2 could have amino acid profiles that are distinct from one another. Thus, an effort was undertaken to categorize the structural and regulatory proteins of HIV1 and HIV2 based on their amino acid makeup.

Ten-fold cross-validation was used to assess the classifier's efficacy at separating structural and regulatory proteins from HIV1 and HIV2. There is an RBF kernel-based parameter selection tool included in LibSVM.

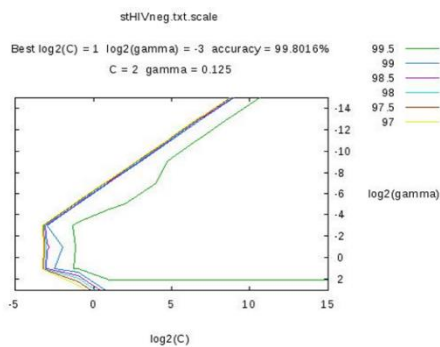


**Figure 2** Grid search of Other than HIV (Negative instances)

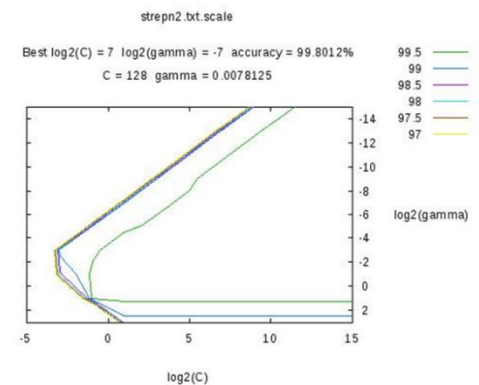


**Figure 3** Grid search for structural proteins instances of HIV1 and HIV2

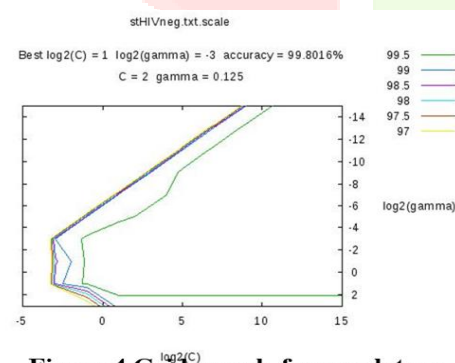


**Figure 4** Grid search for regulatory protein instances of HIV1 and HIV2.

**Grid search for cross-validation.** We used the in-built modules of libSVM tools to do a grid search for C and Gamma on Datasets 1, 2, 3, and 4, and we received the following charts and diagrams.

To follow the process, a rescaling is done to determine which grid path is best. For each data set, we test out several combinations of C and gamma and pick the one with the highest cross-validation accuracy. Grid search yields  $C=128$  and  $\gamma=0.0078$ , leading to a 99.80% accuracy for dataset 1, as displayed in Figure 4.1, and  $C=2$  and  $\gamma=0.0078$ , yielding a 100% accuracy for dataset 2. In this case, we got a result of  $\gamma=0.125$ . Figure 4.2 shows set 2 with  $C = 2$  and  $\gamma = 0.125$  from a grid search; Figure 4.3 shows set 3 with  $C = 128$  and  $\gamma = 0.125$  from a grid search; and Figure 4.4 shows set 4 with  $C = 2$  and  $\gamma = 0.125$  from a grid search. Figure 4.4 has a gamma value of 0.0078.

In Table 1, we can see the C, gamma (g, or ), and fidelity values of structural and regulatory proteins in HIV-1 and HIV-2.

Therefore, the structural and regulatory proteins of HIV1 and HIV2 have an amino acid composition that

| Sl. No. | Datasets                                            | C   | gamma  | Accuracy |
|---------|-----------------------------------------------------|-----|--------|----------|
| 1.      | Structural and Regulatory proteins of HIV1 and HIV2 | 128 | 0.0078 | 99.80%   |
| 2.      | Others(not related to HIV)                          | 2   | 0.125  | 99.80%   |
| 3.      | Structural Proteins of HIV1 and HIV2                | 128 | 0.0078 | 100%     |
| 4.      | Regulatory Proteins of HIV1 and HIV2                | 2   | 0.125  | 99.80%   |

**Table 1: C, gamma, and precision of HIV1 and HIV2 structural and regulatory proteins**

corresponds to a Curacy of 99.80 - C=128 and gamma=0.0078. The significance of protein sequences' amino acid content was empirically demonstrated in this work. With an accuracy of 99.80, we were able to classify the structural and regulatory protein dataset 1 with the other datasets, suggesting that there is some variation between the two. The third dataset of HIV1 and

HIV2 structural proteins is perfect, indicating that their amino acid makeup is identical. The difference between HIV1 and HIV2 regulatory proteins in dataset 4 is 99.80 percent attributable to genetic variation (vpx for HIV2). Vpu (viral protein unique) is an exclusive HIV1 protein that facilitates the cleavage of CD4 in the endoplasmic reticulum and the subsequent release of infectious virus particles from infected cells' plasma membranes. There is a viral protein x (Vpx) that is unique to HIV2. Vpx is incorporated into the virion particle and therefore present at a site during the early stages of viral infection, and it is closely related to vpr (comparing their gene sequences) and appears as a duplication of the vpr gene suggests that it helps the viral nucleoprotein complex enter the nucleus. Genestat, rev, nef, vif, and vpr are the other five genes that encode regulatory proteins that limit HIV's capacity to replicate and infect cells increase. Support vector machines (SVMs) were used to create the classifier since they have previously been proved to perform better than random methods for classifying biological data. Implemented are algorithms for the classification of structural and regulatory proteins in HIV1 and HIV2 datasets. The data acquired here is useful for identifying subtypes of HIV-1. There is evidence that new proteins found in HIV1 and HIV2 are either structural proteins or regulatory proteins. This data shows how crucial HIV proteins are to the virus's functionality and offers fresh perspective on the search for HIV1 and HIV2 treatments. To gain a deeper knowledge of how HIV develops, scientists are focusing on structural and regulatory proteins.

## Conclusion:

By comparing a hybrid model that uses both amino acid and dipeptide composition to categorize HIV1 and HIV2, we can see how the two interact with one another. The relevance of amino acids in proteins is further demonstrated by research that categorizes structural and regulatory proteins of HIV1 and HIV2. Using these forecasts, researchers can more easily identify Dipeptide patterns, domain connections, protein-protein interactions, and protein folding. In order to further enhance the precision of predictions, network-based methods are being studied. These methods provide comprehensive data on proteins. This is a step toward bolstering a wide range of wet lab procedures now being employed to create novel treatments and medications for these conditions.

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 based on them 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 based on the rigorous testing done by various parameters and various kernels.

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.

## 6. 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 25<sup>th</sup> 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. pages 515-25. (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).
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).
39. 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).
26. Translocon-Assisted Membrane Protein Folding from the Stephen White laboratory at UC Irvine, (2008).
27. White S.H. and Heigne G.V. "How translocon selects transmembrane helices?", Annual review of Biophysics, vol. 37, pages 23-42, (2008).
28. 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).
29. Zazzi M et al. "Predicting response to antiretroviral treatment by machine learning", The EuResist project, Intervirology, vol.55, no.2, pages 123-127, (2012).
30. [http:// www.pdb.org](http://www.pdb.org)
31. <http://bioweb.pasteur.fr/seqanal/interfaces/topagesred.html>
32. <http://www.cbs.dtu.dk/services/TMHMM/>  
or  
<http://www.cbs.dtu.dk/services/TMHMM-2.0/>
33. <http://www.dtrek.com/svm.html>
34. <http://www.csbio.sjtu.edu.cn/bioinf/virus>
35. <http://www.niaid.nih.gov>
36. <http://www.uniprot.org>
37. <http://www.misangar.ac.uk/microRNA/registry>
38. <http://www.yale.edu/bio243/HIV/genome.html>
39. [http://i\\_base.info/qa/36](http://i_base.info/qa/36)