Journal of Advanced Bioinformatics Applications and Research ISSN 0976-2604
Vol 2, Issue 1, 2011, pp 84-88 http://www.bipublication.com

SVM MODEL FOR CLASSIFICATION OF STRUCTURAL AND REGULATORY PROTEINS OF HIV1 AND HIV2

Anubha Dubey*, Bhasker Pant*, UshaChouhan
*Department of Bioinformatics, MANIT, BHOPAL, INDIA
Department of Mathematics, MANIT, BHOPAL,INDIA

ABSTRACT

HIV is human immunodeficiency virus causes AIDS (Acquired Immunodeficiency Virus) which leads to life threatening opportunistic infections.HIV has two types HIV1 and HIV2. To understand the mechanism of disease progression structural and regulatory proteins play an important role. In the present paper an approach is being developed to classify structural and regulatory proteins of HIV1 andHIV2 by using Libsvm on the basis of amino acid composition. The performance of the method was evaluated using 10 fold cross validation where accuracy of 99.8% was obtained for structural and regulatory. Other group, structural and regulatory was 99.8%, 100% and 99.8% simultaneously.

Keywords- HIV, Structural, regulatory, Libsvm.

INTRODUCTION

One of the biggest challenges facing biologists today is the structural and functional classification and characterization of protein sequences. For example, in humans, the number of proteins for which the structures and functions are unknown makes up more than 40% of the total number of proteins. As a result, over the past couple of decades, extensive research has been done on trying to identify the structures and functions of proteins.

It is well known that proteins are essential parts of organisms and participate in virtually every process within cells. Many proteins are enzymes that catalyze biochemical reactions and are vital to metabolism. Proteins also have structural or mechanical functions, number of computational approaches have been developed over the years to predict the localization of proteins. Initial efforts relied on amino acid compositions [1,2], the prediction of signal peptides [3,4] or a combination of both.

Later efforts were targeted at incorporating sequence order information (in the form of dipeptide compositions etc.) in the prediction algorithms [7,8].

In this present approach we are trying to develop a classifier to classify HIV1and HIV2 structural and functional proteins on the basis of their amino acid composition. HIV contains nine gene made of 9749 base pairs. All retrovirus contain the genes gag (codes for internal structural proteins and capsid proteins using about 2000 base pairs), pol (codes for the three enzymes necessary for replication using 2900 base pairs). Other genes within HIV are tat (transactivation protein), rev (regulator of expression of Virus protein), vif (virus infectivity factor), nef (misnamed negative regulator factor, but really an enhancing factor), vpr (virus protein R), and vpu (Virus protein U) encoding 19 proteins [9]. The products of *gag*, *pol*, and *env* genes, which are essential components of the retroviral particle, are structural proteins. While regulatory and accessory gene are vif, vpu, nef, tat,rev, vpr. Tat and Rev proteins of HIV/SIV and Tax and Rex proteins of HTLVs. They modulate transcriptional and posttranscriptional steps of virus gene expression and are essential for virus propagation. As proteins are assembled from amino acid using information encoded specified by the some amino acids using information encoded in genes. Each protein has its own unique amino acid sequence

that is specified by the nucleotide sequence of the gene encoding this protein. The size of a synthesized protein can be measured by the number of amino acids it contains and by its total molecular mass, which is normally reported in units of Daltons [10].
Knowledge of protein structure plays a crucial role in analysis of protein function, simulation of protein-ligand interaction, rational drug discovery and in many other applications.

II Material and Methods:

To achieve our goal and develop our methodology we obtained the dataset from Swissprot/Uniprot databank of Expasy server (12). The following two data sets were used.

Dataset1: It consisted of all the structural and functional proteins of HIV1 and HIV2. All the entries marked as fragments were not included in the dataset. The total instances were for structural and regulatory are 502,430 for structural, for regulatory is 72. The 502 were positive belonging to HIV1and HIV2 and 502 were negative instances belonging to enzymatic group.
Dataset2: It consists of all the protein instances of group other then HIV. Total instances taken are 502.
Dataset3: It consisted of all the structural proteins of HIV1 and HIV2. All the entries marked as fragments were not included in the dataset. They were treated as negative instances. The total instances were 430 for structural
Dataset4: It consisted of all the regulatory proteins of HIV1 and HIV2. All the entries marked as fragments were not included in the dataset. They were treated as negative instances. They total instances were 72 for regulatory
For training dataset we consider sequences belonging to structural and regulatory proteins. Support vector machine (Binary classification) is used for classification.

Support vector machine

The concept of Support Vector Machine to achieve our goal and develop our methodology we obtained the dataset from Swissprot/Uniprot databank of Expasy server (11, 12). The following two data sets were used. Support Vector Machine (SVM) was first introduced by Vapnik [13, 14] and in recent times, the SVM approach has been used extensively in the areas of classification and regression. SVM is a learning algorithm which, upon training with a set of positively and negatively labeled samples, produces a classifier that can then be used to identify the correct label for unlabeled samples. SVM builds a classifier by constructing an optimal hyper plane that divides the positively and the negative labeled samples with the maximum margin of separation. Each sample is described by a feature vector. For a detailed description of the mathematics behind SVM, we refer the reader to an article by Burges [15]. For the present study, we used the SVM*light* package (version 6.01) created by Joachim's [16]. The package is available online and is free for scientific use.[17,18] Amino Acid Composition based classification of HIV1 groups study had been done by some researchers.[20]

Previously, this parameter has been used for predicting the sub cellular localization of proteins (19). The amino acid composition is the fraction of each amino acid type within a protein.

The fractions of all 20 natural amino acids were calculated by using Equation 1,

$$= \frac{Total\ Number\ of\ amino\ acid\ i}{Total\ number\ of\ amino\ acids\ in\ a\ protein}$$

Evaluation of Performance

The performance of our classifier was judged by 10 fold cross validation. The LibSVM provides a parameter selection tool using the RBF kernel: cross validation via grid search. A grid search was performed on C and Gamma using an inbuilt module of libSVM tools on Dataset 1, Dataset 2, Dataset 3 and Dataset 4 as shown inFigure1, Figure2, Figure3, and Figure4.

FIG1.Total Structural and Negative protein instances of HIV1 and HIV2

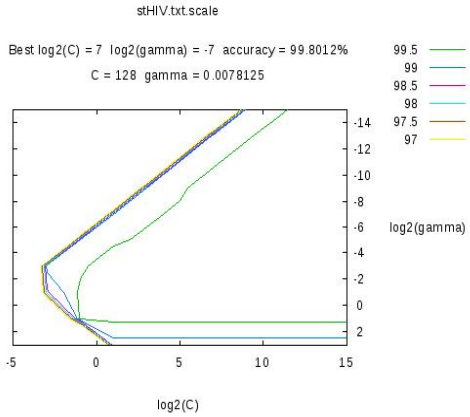


FIG2. Other than HIV (Negative instances)

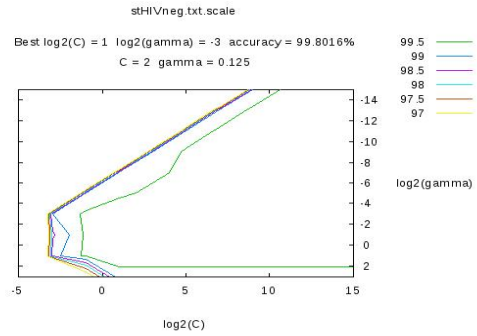


FIG 3: Structural proteins instances of HIV1 and HIV2

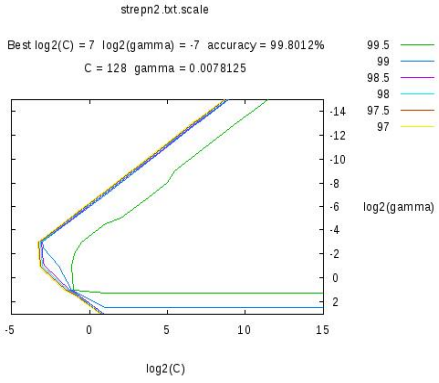
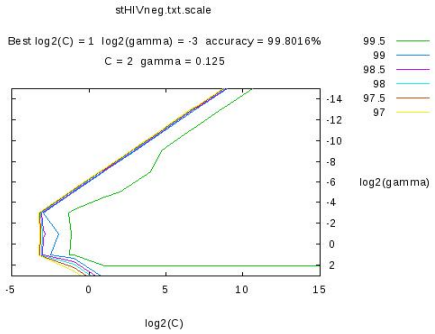


FIGURE 4. Regulatory protein instances of HIV1 and HIV2.



Here pairs of C and Gamma are tried and the one with the best cross validation accuracy is picked. On using the values of C=128 and Gamma=0.0078 obtained through grid search an accuracy of 99.80% was obtained on Dataset 1,the value of C=2 and Gamma= 0.125 obtained through grid search on Dataset 2,the value of C= 2 and Gamma= 0.125 obtained through grid search on Dataset 3,the value of C=128 and Gamma=0.0078 obtained through grid search on Dataset4.Prediction system assessment True positive (TP) and true negatives (TN) were identified as the positive and negative samples, respectively. False positives (FP) were negative samples identified as positive. False negatives (FN) were positive samples identified as negative. The prediction performance was tested with sensitivity (TP/ (TP+FN)), specificity (TN/ (TN+FP)), and overall accuracy (Q2). The accuracy for each group of HIV-1 was calculated as described by Hua and Sun [14] and shown below in equation 2.

$$Accuracy(x) = \frac{tp + tn}{tp + tn + fp + fn}$$

Polycomp: The input vector of 450 was generated directly in the format of SVM by software Polycomp developed under Department of Bioinformatics, MANIT, and Bhopal, India[21]. This software generates data which can be directly fed into the classifier hence saving valuable time and energy needed for formatting the hybrid .

Discussion: We implemented our algorithm for classification of HIV1 and HIV2 structural and regulator protein dataset The results obtained here will be helpful in differentiating between different groups of HIV-1. A new protein discovered can be shown to either belonging to structural and regulatory HIV1 and HIV2.The c, g and accuracy of structural and regulatory of HIV-1 and HIV2 are given in Table 1.

Table 1: The c, g and accuracy of structural and Functional protein of HIV1 and HIV2.

	c	g	Accuracy
Structural and regulatory HIV1and HIV2	128	0.0078	99.80

The HIV1 lead to faster disease progression in comparison to HIV2. HIV1 and HIV2 are classified using amino acid composition of structural and regulatory proteins which is given in Table2.

S.No	GROUPS	C	g	Accuracy
1)	Others	2	0.125	99.80%
2)	Structural Proteins	2	128	100%
3)	Regulatory proteins	2	0.125	99.8%

IV CONCLUSION

Structural and functional proteins have been an active area of research. A number of efforts have previously used amino acid compositions as well as limited sequence order information in order to predict protein function. In this work, we have developed a novel approach based on using amino acid composition of HIV1and HIV2 of structural and functional proteins. Our results clearly highlight the importance of amino acid composition in differentiating between these groups. This model can also be an important tool to understand the differences between structural and regulatory proteins of HIV1 and HIV2. Hence a step towards assisting various wet lab techniques in devising novel drugs and therapeutic agents against these two. The correlation of structural and regulatory proteins with their amino acid composition explored here can be useful to obtain better insight about these proteins. Their molecular and physiological roles along with the substrate affinity can also be correlated with amino acid composition. The accuracy of predicting group all the proteins of HIV1 and HIV2 was found to be 99.80 %.The overall accuracy of the amino acid composition-based classifier for classifying the structural and regulatory proteins was 99.80 respectively.

ACKNOWLEDGEMENT

The authors are highly thankful to the Department of Biotechnology, New Delhi, India and M.P. Council of Science and Technology M.P., Bhopal, India for providing support in the form of Bioinformatics infrastructure facility to carry out the research work.

V. REFERRENCES:

[1] Hua S, Sun Z: Support vector machine approach for protein sub cellular localization prediction**.** *Bioinformatics* 2001, 17**(**8**):**721-728.
[2]. Reinhardt A, Hubbard T: Using neural networks for prediction of the sub cellular location of proteins. *Nucleic Acids Res* 1998, 26(9)**:**2230-2236 [3].Claros M, Vincens P: Computational method to predict mitochondrial imported proteins and their targeting sequences. *Eur J Biochem* 1996, 241**:**779-786.
[4]. Emanuelsson O, Nielsen H, Brunak S, von Heine G: Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence**.** *Journal of Molecular Biology* 2000, 300(4)**:**1005-1016.
[5]. Emanuelsson O, Nielsen H, von Heine G: ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites**.** *Protein Sci* 1999, 8**:**978-984.
[6]. Fujiwara Y, Asogawa M, Nakai K: Prediction of Mitochondrial Targeting Signals using Hidden Markov Models**.** In *Genome Informatics 1997*. Edited by: Miyano S, Takagi T. Japanese Society for Bioinformatics, Tokyo: Universal Academy Press; 1997:53-60.
[7].Predotar: A prediction service for identifying putative mitochondrial and plastid targeting sequences[http://www.inra.fr/predotar] 1997.
[8]. Nakai K, Horton P: PSORT: a program for detecting the sorting signals of proteins and predicting their sub cellular localization. *Trends Biochem Sci* 1999, 24:34-35.
[9]. HIV Encyclopaedia.

[10]. *Protein Encyclopaedia.*
11. Chou KC, Zhang CT: Predicting protein folding types by distance functions that make allowances for amino acid interactions. *J Biol Chem* 1994, 269(35):22014-20.
12. Chou KC: A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins* 1995, 21(4):319-344. Chou KC, Elrod DW: Prediction of membrane protein types and sub cellular locations. *Proteins* 1999, 34:137-153.
[13]. Chou KC, Elrod DW: Protein Sub cellular location prediction. *Protein Eng* 1999, 12(2):107-118.
[14]. Chou KC: Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 2001, 43(3):246-255.
[15]. Cui Q, Jiang T, Liu B, Ma S: Esub8: A novel tool to predict protein subcellular localizations in eukaryotic organisms. *BMC Bioinformatics* 2004, 5:66.
[16]. Vapnik V: *The Nature of Statistical Learning Theory*. Springer; 1995.

[17]. Vapnik V: *Statistical Learning Theory*. Wiley; 1998.

[18]. Bock JR, Gough DA: Predicting protein-protein interactions from primary structure. *Bioinformatics* 2001, 17(5):455-460.
[19]. 11.Park KJ, Kanehisa M: Prediction of protein sub cellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics* 2003, Analysis, synthesis and diagnostics of array antennas through complex-valued neural networks", Microwave and Optical Technology
19(13):1656-1663
[20]. Dubey Anubha, Pant Bhasker SVM model for amino acid composition based classification of HIV1 groups.