Anubha Dubey, Dr. Usha Chouhan / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 2, Mar-Apr 2012, pp.555-559

COMPUTATIONAL MODEL FOR PREDICTION OF TRANSMEMBRANE HELICES IN HIV

Anubha Dubey* Dr.Usha Chouhan**

*Research Scholar
Department of Bioinformatics
**Assistant Professor
Department of Mathematics
MANIT, BHOPAL

ABSTRACT: Genomics and proteomics have added valuable information to our knowledgebase of the human biological system including the discovery of therapeutic targets and disease biomarkers. In order to aid in the identification of membrane proteins, a number of computational methods have been developed. These tools operate by predicting the presence of transmembrane segments. Here we utilize SOSUI prediction method to classify amino acid sequences by two types of transmembrane helices, primary and secondary transmembrane segments. In this work, we have analyzed HIV protein dataset using the SOSUI system which not only predicts transmembrane helix regions but also classifies them whether the protein is soluble or membrane. Further a computational model is being developed by machine learning with fair accuracy.

***Keywords*:** Genomics, Proteomics, Biomarker, Transmembrane

INTRODUCTION: A membrane protein is a protein molecule that is attached to, or associated with the membrane of a cell or an organelle. More than half of all proteins interact with membranes. Membrane proteins can be divided into several categories [1]

- Integral membrane proteins which are permanently bound to the lipid bilayer
- Peripheral membrane proteins that are temporarily associated with lipid bilayer or with integral membrane proteins
- Lipid-anchored proteins bound to lipid bilayer bound through lipidated amino acid residues

In addition, pore-forming toxins and many antibacterial peptides are water-soluble molecules, but undergo a conformational transition upon association with lipid bilayer and become reversibly or irreversibly membrane-associated. A slightly different classification is to divide all membrane proteins to integral and amphitropic [2]. The amphitropic are proteins that can exist in two alternative states: a water-soluble and a lipid bilayer-bound. The amphitropic protein category includes water-soluble channel-forming polypeptide toxins, which associate irreversibly with membranes, but excludes peripheral proteins that interact with other membrane proteins rather than with lipid bilayer. Membrane Proteins commonly function as complexes. These complexes are vital to cellular function. Understanding how these complexes are assembled degraded, and their composition are crucial to understanding their function and regulation. Reoccurring in recent literature are the ideas that: membrane protein complexes assemble in an orderly fashion, chaperones aid assembly by preventing unfavorable interactions, and membrane proteins can be interchanged in existing complexes. Membrane protein complexes assemble through the orderly assembly of intermediates. HIV HIV infection in humans is considered pandemic by the World Health Organization (WHO). Nevertheless, complacency about HIV may play a key role in HIV risk [3, 4]. The RNA genome consists of at least seven structural landmarks (LTR, TAR, RRE, PE, SLIP, CRS, and INS) and nine genes (*gag*, *pol*, and *env*, *tat*, *rev*, *nef*, *vif*, *vpr*, *vpu*, and sometimes a tenth *tev*, which is a fusion of tat env and rev) encoding 19 proteins. Three of these genes, *gag*, *pol*, and *env*, contain information needed to make the structural proteins for new virus particles [5]. For example, *env* codes for a protein called gp160 that is broken down by a viral enzyme to form gp120 and gp41. The six remaining genes, *tat*, *rev*, *nef*, *vif*, *vpr*, and *vpu* (or *vpx* in the case of HIV-2), are regulatory genes for proteins that control the ability of HIV to infect cells, produce new copies of virus (replicate), or cause disease [5]. A substantial amount of gp120 can be found released in the medium. gp120 contains the binding site for the CD4 receptor, and the seven transmembrane domain chemokine receptors that serve as co-receptors for HIV-1. Vif is a cytoplasmic protein, existing in both a soluble cytosolic form and a membrane-associated form. The latter form of Vif is a peripheral membrane protein that is tightly

associated with the cytoplasmic side of cellular membranes.

A multifunctional 27-kd myristoylated protein produced by an ORF located at the 3' end of the primate lentiviruses. Other forms of Nef are known, including nonmyristoylated variants. Nef is predominantly cytoplasmic and associated with the plasma membrane via the myristoyl residue linked to the conserved second amino acid (Gly). Nef has also been identified in the nucleus and found associated with the cytoskeleton in some experiments. One of the first HIV proteins to be produced in infected cells, it is the most immunogenic of the accessory proteins. Amino acids are the building blocks of proteins. There are many amino acid sequences in proteomes which are not homologous to any other sequences. Therefore, methods to classify proteins independent of the sequence homology are strongly required for computational analysis of proteomes. Proteins may be divided into two categories: soluble and membrane proteins. Since membrane proteins are characterized by the existence of long hydrophobic transmembrane helices, the classification of amino acid sequences into two types of proteins, soluble and membrane proteins, is possible with considerably high accuracy. Transmembrane protein is a protein that goes from one side of a membrane through to other side of the membrane. They permit the transport of specific substances across the biological membrane The classification of all amino acid sequences in several proteomes was reported recently, leading to the conclusion that the fraction of membrane proteins is about 30% [6]. However, those methods provided only the information about the number of transmembrane helices. We have previously proposed a new method (SOSUI) to classify amino acid sequences by two types of transmembrane helices, primary and secondary transmembrane segments. In this work, we have analyzed HIVdataset using the SOSUI system [7] which not only predicts transmembrane helix regions but also classifies them by the strength of interaction with lipid membranes and whether the protein is soluble or membrane. Further a computational model is being developed by machine learning supervised algorithms which correctly builds models according to number of amino acids, average hydrophobicity, and type of membrane protein, number of transmembrane segment, length of transmembrane segment, [12, 13, 14, and 15].

Methods:

Here the protein sequence data has been taken from Uniprot data bank [8] of which the present work focuses on the further classification of according to soluble and membrane proteins. Membrane proteins have transmembrane helices. 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 [9]. 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 is a flowchart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. Internal nodes are denoted by rectangles, and leaf nodes are denoted by ovals. The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery [9,10].

Logistic: In statistics, logistic regression (sometimes called the logistic model or logit model) is used for prediction of the probability of occurrence of an event by fitting data to a logit function logistic curve. It is a generalized linear model used for binomial regression. Like many forms of regression analysis, it makes use of several predictor variables that may be either numerical or categorical [11].

Bagging: Bagging also called as bootstrap aggregating, is a technique that repeatedly samples from a data set according to a uniform probability distribution. Each bootstrap sample has the same size as the original data [9,10].

The proteins used for this study were collected from Uniprot/Swiss Prot database. All the redundant data is removed and complete sequences are taken for study.

Result: The machine learning is a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviours based on empirical data, such as from sensor data or databases. A learner can take advantage of examples (data) to capture characteristics of interest of their unknown underlying probability distribution. Data can be seen as examples that illustrate relations between observed variables. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data. In this study we used

supervised learning algorithms of machine learning. Supervised learning generates a function that maps inputs to desired outputs (also called labels, because they are often provided by human experts labeling the training examples). The classifiers used for computational model are logistic, J48, Bagging which gives good results for classification of proteins into soluble or membrane. Bagging gives better result in all the cases except types of membrane proteins (case5).

CASE1: In average hydrophobicity Bagging has given better result. The result and comparative analysis is shown in table 1.

Table1 shows better result with bagging.

Soluble			Membrane		
Classifier	TP	FP	TP	FP	Accuracy
logistic	0.941	0.031	0.969	0.059	95.9184%
J48	0.971	0.047	0.953	0.029	95.9184%
Bagging	0.971	0.031	0.969	0.029	96.9388%

Detailed Accuracy By Class [bagging]

TPRate FPRate Precision Recall F-Measure ROC Class

0.971 0.031 0.943 0.971 0.957 0.986 soluble

0.969 0.029 0.984 0.969 0.976 0.986 membrane

Weighted Avg.

0.969 0.03 0.97 0.969 0.969 0.986

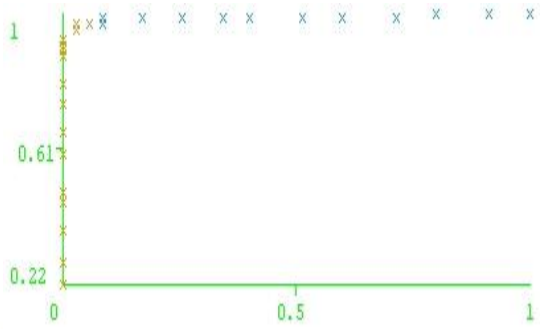
=== Confusion Matrix ===

a b <-- classified as

33 1 | a = soluble

2 62 | b = membrane

Figure 1. ROCof average hydrophobicity Bagging



CASE 2: In number of amino acids bagging has given better result. The result and comparative analysis is shown in table 2.

Table 2 shows better result with bagging.

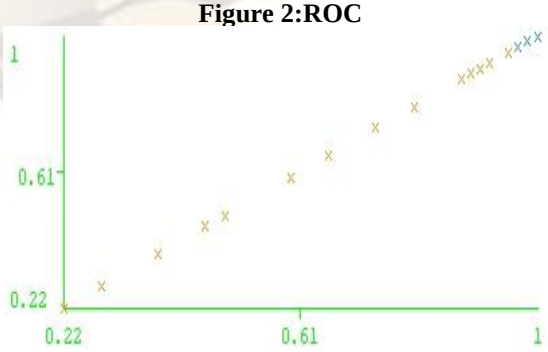
Soluble			Membrane		
Classifier	TP	FP	TP	FP	Accuracy
logistic	0.941	0.031	0.969	0.059	95.9184%
J48	0.971	0.047	0.953	0.029	95.9184%
Bagging	0.971	0.031	0.969	0.029	96.9388%

=== Detailed Accuracy By Class ==
TP Rate FP Rate Precision Recall F-Measure ROCclass
0.971 0.031 0.943 0.971 0.957 0.986 soluble
0.969 0.029 0.984 0.969 0.976 0.986 membrane
Weighted Avg.
0.969 0.03 0.97 0.969 0.969 0.986
=== Confusion Matrix ===

a b <-- classified as

33 1 | a = soluble
2 62 | b = membrane

Figure 2:ROC



CASE 3: In length of transmembrane region, Bagging has given better result. The result and comparative analysis is shown in table 2.

Table 3 shows better result with bagging.

	Soluble		Membrane		
Classifier	TP	FP	TP	FP	Accuracy
logistic	0.941	0.031	0.969	0.059	95.9184 %
J48	0.971	0.047	0.953	0.029	95.9184 %
Bagging	0.971	0.031	0.969	0.029	96.9388 %

Detailed Accuracy By Class

TPRate FPRate Prcision Recall F-Measure ROC Class

0.971 0.031 0.943 0.971 0.957 0.986 soluble
0.969 0.029 0.984 0.969 0.976 0.986 membrane
Weighted Avg.
0.969 0.03 0.97 0.969 0.969 0.986

Confusion Matrix

a b <-- classified as
33 1 | a = soluble
2 62 | b = membrane

Figure 3: ROC

CASE 4: In number of transmembrane segment, Bagging has given better result. The result and comparative analysis is shown in table 4.

Table 4 shows better result with bagging.

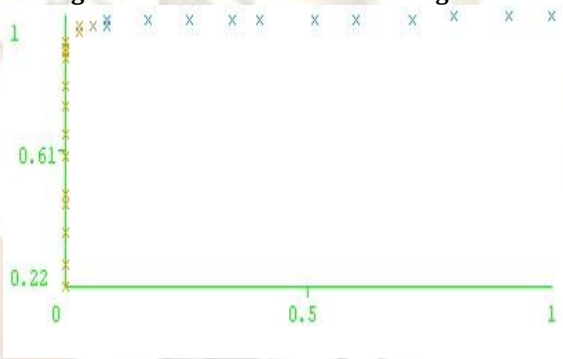
	Soluble		Membrane		
Classifier	TP	FP	TP	FP	Accuracy
logistic	0.941	0.031	0.969	0.059	95.9184 %
J48	0.971	0.047	0.953	0.029	95.9184 %
Bagging	0.971	0.031	0.969	0.029	96.9388 %

Detailed Accuracy By Class

TPRate FPRate Precision Recall F-Measure ROC
0.971 0.031 0.943 0.971 0.957 0.986 soluble
0.969 0.029 0.984 0.969 0.976 0.986 membrane

Weighted Avg.
0.969 0.03 0.97 0.969 0.969 0.986
Confusion Matrix
a b <-- classified as
33 1 | a = soluble
2 62 | b = membrane

Figure 4: ROC transmembrane segment.



Discussion: Bagging has given better result with cases1 to 4.Bagging is used also in the sensitivity analysis procedure. A ROC curve 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 **(**1-5**)**. As ROC depicts the performance, we can refer from the confusion matrix that in case 1,2,3,4, the false positive ratio is 0.031 in soluble protein and 0.029 in membrane protein, which clearly indicates that the true positive ratio is 0.971 in soluble

protein and 0.969 in membrane protein. In all these cases the true positive and false positive values are same, this shows that Bagging is better in all the above mentioned cases. Case5 shows false positives 0 in soluble and 1 in membrane and true positives 0 in soluble and 1 in membrane with Bagging. And J48 has predicted and classified better with type of membrane proteins. It has showed false positive 0.047 in soluble, 0.029 in membrane and true positives 0.971 in soluble and 0.953 in membrane. The accuracy of results for the five cases obtained from all the three classifiers with input as protein sequences as predicted from three different classifier and their comparison is presented in (Tables 1-4). Cases 1 to 4 Bagging has predicted and classified better with accuracy 96.9388%.

CONCLUSION

Among all the three classifiers, the classification of HIV protein sequences on the basis of average hydrophobicity, number of amino acid, length of transmembrane segment, number of transmembrane segments, type of membrane protein as are five cases. So it is concluded that Bagging found suitable for cases 1, 2, 3, 4. As it gives estimates of what variables are important in the classification. J48 predicts better result in case 5 as its speed are good and performs better calculation and has better memory. And a computational model is being developed which accurately classified HIV proteins into soluble proteins and membrane proteins As more proteins have discovered the accuracy of the model is maintained and server is also developed. Database can also be redesigned to provide more scalable system. The challenge now is to organize these data in a way that evolutionary relationships between proteins can be uncovered and used to understand better membrane protein function. The first steps common to the analysis of any large set of data are to group together data points that are similar, and then to identify connections between those elementary groups. These steps are usually performed with classification techniques. Hence structural classification of proteins leads to drug discovery and also helpful to biomedical scientists to develop protocols for identification of HIV.

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. Gerald Karp (2009). Cell and Molecular Biology: Concepts and Experiments. John Wiley and Sons. pp. 128–.ISBN 9780470483374. http://books.google.com/books?id=arRGYE0GxRQC&pg=PA128. Retrieved 13 November 2010.
2. Johnson JE, Cornell RB (1999). "Amphitropic proteins: regulation by reversible membrane interactions (review)". *Mol. Membr. Biol.* **16** (3): 217–235. doi:10.1080/096876899294544. PMID 10503244.
3 "CDC – HIV/AIDS – Resources – HIV Prevention in the United States at a Critical Crossroads". dc.gov.http://www.cdc.gov/hiv/resources/reports/hiv_prev_us.htm.. Retrieved 2010-07-28.
4. "HIV and AIDS among Gay and Bisexual Men" (PDF). http://www.cdc.gov/nchhstp/newsroom/docs/FastFacts-MSM-FINAL508COMP.pdf. Retrieved 2010-07-28.
5.Various (2008) (PDF). HIV Sequence Compendium 2008 Introduction.http://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/2008/frontmatter.pdf. Retrieved 2009-03-31.
6. Frishman, D. and Mewes, H.W., Protein structural classes in five complete genomes, Nature structural biology, 4:626{628, 1997.
[7] Hirokawa, T., Seah, B.-C., and Mitaku, S., SOSUI: classification and secondary structure predic-tion system for membrane proteins, Bioinformatics Application Note, 14, 1988.36
8] WWW.UNIPROT.org
9] http://www.cs.waikato.ac.nz/ml/weka
[10]Pang-Ning, Tan.M.Steinbach, V.Kumar. Introduction to Data Mining, **2008**. (s)
11]http://en.wikipedia.org/wiki/Logistic_regression
12] A.Dubey, B.Pant and Neeru Adlakha,"SVM Model for Amino Acid Composition based Classification of HIV1 Groups". *IEEE digital library* published.
[13] A.Dubey, B.Pant and Usha Chouhan," SVM Model for Classification of Structural and Regulatory Proteins of HIV1 and HIV2 is published in Journal of Advanced Bioinformatics Applications and Research ISSN 0976-2604 Vol 2, Issue 1, 2011, pp 84-88
14] Anubha Dubey, Bhaskar Pant, Usha Chouhan Machine learning model for HIV1 and HIV2 enzyme secondary structure classification, Scholars Research Library J. Comput. Method. Mol. Design, 2011, 1 (2): 1-8
[15] ANUBHA DUBEYAND USHA CHOUHAN SUPPORT VECTOR MACHINE FOR CLASSIFICATION OF HIV, PLANT AND ANIMAL miRNA's International Journal of Bioinformatics Research ISSN: 0975–3087, E-ISSN: 0975–9115, Vol. 3, Issue 2, 2011, pp-202-206