International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:1, 2009

Analysis and Classification of Hiv-1 Sub-Type Viruses by AR Model through Artificial Neural Networks

O. Yavuz, and L. Ozyilmaz

***Abstract*—** HIV-1 genome is highly heterogeneous. Due to this variation, features of HIV-I genome is in a wide range. For this reason, the ability to infection of the virus changes depending on different chemokine receptors. From this point of view, R5 HIV viruses use CCR5 coreceptor while X4 viruses use CXCR5 and R5X4 viruses can utilize both coreceptors. Recently, in Bioinformatics, R5X4 viruses have been studied to classify by using the experiments on HIV-1 genome.

In this study, R5X4 type of HIV viruses were classified using Auto Regressive (AR) model through Artificial Neural Networks (ANNs). The statistical data of R5X4, R5 and X4 viruses was analyzed by using signal processing methods and ANNs. Accessible residues of these virus sequences were obtained and modeled by AR model since the dimension of residues is large and different from each other. Finally the pre-processed data was used to evolve various ANN structures for determining R5X4 viruses. Furthermore ROC analysis was applied to ANNs to show their real performances. The results indicate that R5X4 viruses successfully classified with high sensitivity and specificity values training and testing ROC analysis for RBF, which gives the best performance among ANN structures.

***Keywords*—** Auto-Regressive Model, HIV, Neural Networks, ROC Analysis.

I. Introduction

THE aim of this work is that, the statistical data of HIV-1 genome is modeled and analyzed by AR model, and then, the structures of ANNs are evolved to classify HIV-1 viruses successfully. Thus, in the future, the evolved viruses could be determined by using these intelligent structures.

In many biomedical and bioinformatics applications, structures which are modeled by using gene sequences, are used to determine HIV-1 sub-type viruses. For example, the infection effects of the chemokine coreceptor and virus entry are introduced [1]. By using these models and statistical data, the ANNs structures, which have high ability to classify, are evolved [2-6].

O. Y is with the Electronics and Communications Engineering Department, Yildiz Technical University, 34349, Istanbul, Turkey (e:mail: ogyavuz@yildiz.edu.tr).

L. O is with the Electronics and Communications Engineering Department, Yildiz Technical University, 34349, Istanbul, Turkey (e:mail: ozyilmaz@yildiz.edu.tr).

Lamers et al. used HIV-Base software to obtain statistical data of R5X4, R5 and X4 viruses and evolved the ANNs to classify these viruses by using this data. Since the classification results in that study is the training accuracy, it is hard to determine the real performance of the evolved ANNs. For determining the real performance of ANNs, the data which is different training data and unknown, is used to test ANNs [7], [8].

E. Sitbon and S. Pietrokovski obtained the accessible residues of gene sequences for describing protein identity [9]. Kong et al. analyzed accessible residues of HIV-1 genome to design peptide [10].

The many bioinformatics researchers use AR model to process gene data. H. Zhou and H.Yan proposed AR model in spectral analyses of short tandem in DNA sequences [11]. M. Akhtar et al. determined period -3 behaviors by using AR model [12]. G. Rosen reduced the dimension of gene sequence using AR model [13].

In this work, accessible residues of gene sequences are obtained. Since the dimension of gene sequences is large and different than each other, their dimension is reduced and equalized by AR model. This pre-processed data is used to train and test the ANNs

This paper consists of 3 sections. In Section II, data sets, AR model, cross validation, ANNs, ROC analysis and methods are described. Also, the simulation results are given in this Section. In Section III, conclusion and future work are mentioned.

II. Materials and Methods

A. Data Mining

77 R5 sequences, 31 R5X4 sequences and 40 X4 sequences [8] were taken from 148 Los Alamos National Laboratory HIV Sequence Database including 148 data in total (www.hiv.lanl.gov/content/hiv-db/main-page.html).

148 sequences were converted into numeric data by using accessible residues as shown in Table I. However, since the dimension of residues is large, they can not be used to evolve ANNs. For solving this problem, residues have to be pre-processed. In this study, the residues were compressed by AR model.

B. AR Model

Since AR model represents energy of signals, it is chosen to model accessible residues. In many genetic applications, especially, for signals with low Signal to Noise Ratio (SNR), AR model has high performance. AR model is defined by all pole filters as follows;

$$H(z) = \frac{G}{1 + \sum_{k=1}^{N} a_k z^{-k}} \quad (1)$$

where, N is the dimension of AR model. Eq. 1 could be written in time domain as,

$$y_k = \sum_{i=1}^{M} a_i x_{k-i} + w_k \quad (2)$$

where y_k is the estimated signal, a_i is the AR coefficient, w_k is the computational error, and M is the number of AR coefficients[14].

In this work, the obtained residues of the gene sequences are in different dimensions and can not be used for ANNs. Therefore, the dimension of residues was reduced using 10-th AR model.

TABLE I
AMINO ACID COMPOSITION OF THE INSIDE AND SURFACE

Residue	Buried	Accessible
Leu	11.7	4.8
Val	12.9	4.5
Ile	8.6	2.8
Phe	5.1	2.4
Cys	4.1	0.9
Met	1.9	1
Ala	11.2	6.6
Gly	11.8	6.7
Trp	2.2	1.4
Ser	8	9.4
Thr	4.9	7
His	2	2.5
Tyr	2.6	5.1
Pro	2.7	4.8
Asn	2.9	6.7
Asp	2.9	7.7
Gln	1.6	5.2
Glu	1.8	5.7
Arg	0.5	4.5
Lys	0.5	10.3

C. Cross Validation

Cross-validation methods are used in examining the robustness of classifiers. The simplest of these methods is the single training and testing scheme that is often employed in the medical literature. The original data set is split into two groups and one is used for designing the classifier while the hold-out sample is used for testing purposes [15]. In k-fold cross-validation, the data is divided into k subsets of approximately equal size. The net is trained k times, each time leaving out one of the subsets from training, but using only the omitted subset to compute whatever error criterion interests you. If k equals the sample size, this is called "leave-one-out" cross-validation. "Leave-v-out" is a more elaborate and expensive version of cross-validation that involves leaving out all possible subsets of v cases.

In this study a 3-fold cross validation was used. Each class of HIV-1 Sequence which belongs to R5 and X4 or R5X4 viruses, was partitioned into three pieces which consists of 39 R5 or X4 data and 10 R5X4 data respectively. One of the data set was used for testing ANNs, while remaining was used for training. The training and test sets consist of 99 and 49 data, respectively. Thus, three test results have been obtained using different subsets 1, 2 and 3 in simulations.

D. Evolving Neural Network Structures

Accessible residues of HIV-1 Sequences were modeled by 10-th AR coefficients. Thus, size of the accessible residues of HIV-1 Sequences was reduced and unnecessary details of the signals were eliminated. In the next step, these pre-processed data of HIV-1 sequences are used for training and testing of alternative ANNs to classify R5X4, R5 and X4.

For the classification the HIV-1 Sequences, three typesofANNs were chosen. These are, Multilayer Perception (MLP) with all learning rules in Matlab, Radial Basis Function (RBF) Neural Network and General Regression Neural Networks (GRNN). Neural network systems were shown in Fig. 1. a_i and y represent input and output of ANNs, respectively.

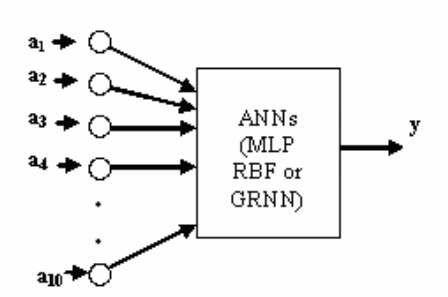

Fig. 3 General ANN structure.

The desired outputs (d) of R5X4 and other genes (R5 or X4) were chosen 0 and 1 respectively.

MLP has 10 input, 20 hidden and 1 output neurons. All learning algorithms in Matlab were examined for determining the most successful ones. After this process, Levenberg-Marquardt backpropagation learning rule (trainlm) and Scaled conjugate gradient backpropagation (trainscg) were chosen as the best the learning rules. The performance of MLP depends on initial conditions. Hence, training and testing processes were repeated 10 times and the best results obtained were given in Table II.

RBF and GRNN have 10 input and 1 output neurons. The spread parameter is chosen 70 for RBF, 1 for GRNN, respectively. The classification accuracy of RBF and GRNN is given in Table III and IV, respectively.

TABLE II
CLASSIFICATION ACCURACY OF MLP

The Learning Rule:Trainlm		
Subset	Training Accuracy	Testing Accuracy
1	94.94%	63.30%
2	100.00%	61.22%
3	94.94%	55.10%
The Learning Rule:Trainscg		
Subset	Training Accuracy	Testing Accuracy
1	91.92%	65.30%
2	91.92%	61.22%
3	94.94%	61.22%

TABLE III
CLASSIFICATION ACCURACY OF RBF

Subset	Training Accuracy	Testing Accuracy
1	98.98%	57.14%
2	94.00%	51.00%
3	97.98%	63.27%

TABLE IV
CLASSIFICATION ACCURACY OF GRNN

Subset	Training Accuracy	Testing Accuracy
1	95.95%	51.00%
2	94.95%	40.14%
3	97.98%	59.18%

Besides, the classification accuracy of MLP, RBF and GRNN were taken average to compared others ANNs as shown in 5 respectively.

Train and test results show that RBF and MLP with trainscg learning rules give the higher performance than the others. However, the classification accuracy is not enough for the analysis performance of ANNs. Therefore, ROC analysis was applied to these results.

TABLE V
AVERAGE OF CLASSIFICATION ACCURACY OF ANNS

ANNs	**Training**	**Testing**
MLP(Trainlm)	96.63%	59.87%
MLP(Trainscg)	92.63%	62.58%
RBF	96.98%	57.13%
GRNN	96.29%	50.10%

E. ROC Analysis

Receiver Operating Characteristic Analysis (ROC Analysis) is related in a direct and natural way to cost/benefit analysis of diagnostic decision making . It is originated from signal detection theory, as a model of how well a receiver is able to detect a signal in the presence of noise. Its key feature is the distinction between hit rate (or true positive rate) and false alarm rate (or false positive rate) as two separate performance measures. ROC analysis has also widely been used in medical data analysis to study the effect of varying the threshold on the numerical outcome of a diagnostic test [16].

TABLE VII
ROC BLOCK DIAGRAM

		Actual	
		T	F
Predicted	T	True Positives (TP)	False Positives (FN)
	F	True Negatives (TN)	True Negatives (TN)

The limitations of diagnostic accuracy as a measure of decision performance require introduction of the concepts of the "sensitivity" and "specificity" of a diagnostic test as shown in Table VI.

The sensitivity and specificity can be written as follows;

$$Sensitivity = \frac{\#true\ positives}{\#true\ positives + \#\ false\ positives} \quad (3)$$

$$Specificity = \frac{\#true\ negatives}{\#true\ negatives + \#\ false\ negatives} \quad (4)$$

In this study, "sensitivity" and "specificity" of ANNs could be defined as follows;

$$Sensitivity = \frac{R5X4_{True}}{R5X4_{True} + R5X4_{False}} \quad (5)$$

$$Specificity = \frac{(R5\ or\ X4)_{True}}{(R5\ or\ X4)_{True} + (R5\ or\ X4)_{False}} \quad (6)$$

"Sensitivity" and "specificity" of the most successful ANNs should be approximated to 1. ROC analysis of RBF, GRNN and MLP are given in Table VII, VIII and IX, respectively.

Sensitivity and specificity of RBF, GRNN and MLP are shown in Table X and XI. Table XII shows that, average of sensitivity and specificity values for all ANNs.

TABLE VII
THE RESULTS OF ROC ANALYSIS FOR RBF

Subset 1							
Training				Testing			
	Actual				Actual		
Predicted		R5X4	R5 or X4	Predicted		R5X4	R5 or X4
	R5X4	21	0		R5X4	7	3
	R5 or X4	1	77		R5 or X4	19	20
Subset 2							
Training				Testing			
	Actual				Actual		
Predicted		R5X4	R5 or X4	Predicted		R5X4	R5 or X4
	R5X4	19	2		R5X4	6	4
	R5 or X4	2	76		R5 or X4	20	19
Subset 3							
Training				Testing			
	Actual				Actual		
Predicted		R5X4	R5 or X4	Predicted		R5X4	R5 or X4
	R5X4	20	1		R5X4	7	4
	R5 or X4	1	77		R5 or X4	14	25

TABLE VIII
THE RESULTS OF ROC ANALYSIS FOR GRNN

Subset 1							
Training				Testing			
	Actual				Actual		
Predicted		R5X4	R5 or X4	Predicted		R5X4	R5 or X4
	R5X4	21	0		R5X4	3	7
	R5 or X4	4	74		R5 or X4	8	31
Subset 2							
Training				Testing			
	Actual				Actual		
Predicted		R5X4	R5 or X4	Predicted		R5X4	R5 or X4
	R5X4	16	5		R5X4	0	10
	R5 or X4	0	78		R5 or X4	19	20
Subset 3							
Training				Testing			
	Actual				Actual		
Predicted		R5X4	R5 or X4	Predicted		R5X4	R5 or X4
	R5X4	20	0		R5X4	5	6
	R5 or X4	1	77		R5 or X4	14	25

TABLE IX
THE RESULTS OF ROC ANALYSIS FOR MLP

The Learning Rule:Trainlm							
Subset 1							
Training				Testing			
	Actual				Actual		
		R5X4	R5 or X4			R5X4	R5 or X4
Predicted	R5X4	19	2	Predicted	R5X4	3	7
	R5 or X4	0	78		R5 or X4	11	28
Subset 2							
Training				Testing			
	Actual				Actual		
		R5X4	R5 or X4			R5X4	R5 or X4
Predicted	R5X4	21	0	Predicted	R5X4	4	6
	R5 or X4	0	78		R5 or X4	12	26
Subset 3							
Training				Testing			
	Actual				Actual		
		R5X4	R5 or X4			R5X4	R5 or X4
Predicted	R5X4	16	4	Predicted	R5X4	5	6
	R5 or X4	0	78		R5 or X4	17	22
The Learning Rule:Trainscg							
Subset 1							
Training				Testing			
	Actual				Actual		
		R5X4	R5 or X4			R5X4	R5 or X4
Predicted	R5X4	15	6	Predicted	R5X4	2	8
	R5 or X4	2	76		R5 or X4	9	30
Subset 2							
Training				Testing			
	Actual				Actual		
		R5X4	R5 or X4			R5X4	R5 or X4
Predicted	R5X4	15	6	Predicted	R5X4	9	1
	R5 or X4	2	76		R5 or X4	18	21
Subset 3							
Training				Testing			
	Actual				Actual		
		R5X4	R5 or X4			R5X4	R5 or X4
Predicted	R5X4	16	4	Predicted	R5X4	8	3
	R5 or X4	0	78		R5 or X4	17	21

TABLE X
SENSITIVITY AND SPECIFICITY OF RBF AND GRNN

RBF				
	Training		Testing	
Subset	Sensitivity	Specificity	Sensitivity	Specificity
1	1	0.9872	0.7	0.5128
2	0.9048	0.9744	0.6	0.4872
3	0.9524	0.9872	0.6364	0.641
GRNN				
	Training		Testing	
Subset	Sensitivity	Specificity	Sensitivity	Specificity
1	1	0.9487	0.3	0.7949
2	0.7619	1	0	0.5128
3	1	0.9872	0.4545	0.641

TABLE XI
SENSITIVITY AND SPECIFICITY OF MLP

The Learning Rule: Trainlm				
	Training		Testing	
Subset	Sensitivity	Specificity	Sensitivity	Specificity
1	0.9048	1	0.3	0.718
2	1	1	0.4	0.667
3	0.8	1	0.455	0.5641
The Learning Rule: Trainscg				
	Training		Testing	
Subset	Sensitivity	Specificity	Sensitivity	Specificity
1	0.7143	0.9744	0.2	0.7692
2	0.7143	0.9744	0.9	0.5385
3	0.8	1	0.7273	0.5526

TABLE XII
AVERAGE OF SENSITIVITY AND SPECIFICITY OF MLP, RBF AND GRNN

	Training		Testing	
ANNs	Sensitivity	Specificity	Sensitivity	Specificity
MLP(Trainlm)	0.9016	1	0.385	0.6497
MLP(Trainscg)	0.7429	0.9829	0.6091	0.62
RBF	0.9524	0.9829	0.6455	0.62
GRNN	0.9206	0.9786	0.2515	0.6496

The results showed that the RBF structure has given the best sensitivity and specificity values as shown in Table XII.

III. CONCLUSION AND FUTURE WORK

In this study, since gene sequences have different and large dimensions, the statistical data of HIV-1 subtype genome was modeled and analyzed by 10-th AR model to reduce the dimension of gene sequences. By using this pre-processed data; the optimum structures of ANNs, which have limited number of input and hidden neurons, were evolved to classify HIV-1 subtype viruses successfully. Thus, the classification process takes the minimum processing time. The training and test data were obtained by using 3-fold cross-validation and these data sets were used to train and test the ANNs (MLP, RBF and GRNN). The best training and testing accuracy were obtained from RBF and MLP (trainscg). The results of ANNs were analyzed thoroughly by ROC analysis to decide the best ANN structures. ROC analysis shows that the most successful ANN is RBF.

In future work, other statistical features of genome such as buried residues will be obtained and the performance of the evolved ANNs in this work will be tried to develop by using these features.

REFERENCES

[1] E.A. Berger, P.M. Murphy, and J.M. Farber, "Chemokine Receptors as HIV-1 Coreceptors: Roles in Viral Entry, Tropism, and Disease," *Ann. Rev. Immunology*, vol. 17, pp. 675-700, 1999.

[2] W. Resch, N. Hoffman, and R. Swanstrom, "Improved Success of Phenotype Prediction of the Human Immunodeficiency Virus Type 1 from Envelope Variable Loop 3 Sequence Using Neural Networks," *J. Virology*, vol. 76, pp. 3852-3864, 2001.

[3] J.A. Loannidis, T.A. Trikalinos, and M. Law, "HIV Lipodystrophy Case Definition Using Artificial Neural Network Modeling," *Antiviral Therapy*, vol. 8, pp. 435-441, 2003.

[4] D. Wang and B. Larder, "Enhanced Prediction of Lopinavir Resistance from Genotype by Use of Artificial Neural Networks," *J. Infectious Diseases*, vol. 188, pp. 653-660, 2003.

[5] Z.L. Brumme, W.W.Y. Dong, B. Yip, B. Wynhoven, N.G. Hoffman, R. Swanstrom, M.A. Jensen, J.I. Mullins, R.S. Hogg, J.S.G. Montaner, and P.R. Harrigan, "Clinical and Immunological Impact of HIV Envelope V3 Sequence Variation after Starting Initial Triple Antiretroviral Therapy," *AIDS*, vol. 18, pp. F1-F9, 2004.

[6] L. Milich, B. Margolin, and R. Swanstrom, "V3 Loop of the Human Immunodeficiency Virus Type 1 Env Protein: Interpreting Sequence Variability," *J. Virology*, vol. 67, no. 9, pp. 5623-5634, 1993.

[7] S. Lamers, S. Beason, L. Dunlap, R. Compton, and M. Salemi, "HIVbase: A PC/Windows-Based Software Offering Storage and Querying Power for Locally Held HIV-1 Genetic, Experimental and Clinical Data," *Bioinformatics*, vol. 20, pp. 436-438, 2002.

[8] S. Lamers, L. Susanna, M. Salemi, M. S. McGrath and G. B. Fogel, "Prediction of R5, X4, and R5X4 HIV-1 Coreceptor Usage with Evolved Neural Networks, " *Trans. On Computational Biology and Bioinformatics*, Vol. 5, pp. 291-300, 2008

[9] E. Sitbon, and S. Pietrokovski, "Occurrence of protein structure elements in conserved sequence regions," *BMC Structural Biology*, vol. 7, 2007.

[10] R. Kong, C. X. Wang, X. H. Ma, J. H. Liu, and W. Z. Chen, "Peptides Design Based on the Interfacial Helix of Integrase Dimer, " *27th Annual Int. Conf. of the Engineering in Medicine and Biology Society*, pp. 4743-4746 2005.

[11] H. Zhou, and H.Yan, "Autoregressive Models for Spectral Analysis of Short Tandem Repeats in DNA Sequences," *IEEE Int. Conf. on Systems, Man and Cybernetics*, vol. 2, pp. 1286-1290, 2006.

[12] M. Akhtar, E. Ambikairajah, and J. Epps, "Detection of period-3 behavior in genomic sequences using singular value decomposition," *Proc. of International Conference on Emerging Technologies*, pp. 13-17, 2007

[13] G. Rosen. "Comparison of Autoregressive Measures for DNA Sequence Similarity" IEEE Genomic Signal Processing and Statistics Workshop (GENSIPS) pp. 13-17 2007.

[14] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, New Jersey, 2002.

[15] A. Sboner, C. Eccher, E. Blanzieri, P. Bauer, M. Cristofolini, G. Zumiani, and S. Forti, "A multiple classifier system for early melanoma diagnosis," AI in Medicine, Vol. 27, pp. 29-44, 2003.

[16] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection". *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* vol. 2 pp. 1137–1143, 1995.