International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:10, 2009

Optimal Multilayer Perceptron Structure For Classification of HIV Sub-Type Viruses

Zeyneb Kurt, Oguzhan Yavuz

***Abstract*—**The feature of HIV genome is in a wide range because of it is highly heterogeneous. Hence, the infection ability of the virus changes related with different chemokine receptors. From this point, R5 and X4 HIV viruses use CCR5 and CXCR5 coreceptors respectively while R5X4 viruses can utilize both coreceptors. Recently, in Bioinformatics, R5X4 viruses have been studied to classify by using the coreceptors of HIV genome.

The aim of this study is to develop the optimal Multilayer Perceptron (MLP) for high classification accuracy of HIV sub-type viruses. To accomplish this purpose, the unit number in hidden layer was incremented one by one, from one to a particular number. The statistical data of R5X4, R5 and X4 viruses was preprocessed by the signal processing methods. Accessible residues of these virus sequences were extracted and modeled by Auto-Regressive Model (AR) due to the dimension of residues is large and different from each other. Finally the pre-processed dataset was used to evolve MLP with various number of hidden units to determine R5X4 viruses. Furthermore, ROC analysis was used to figure out the optimal MLP structure.

***Keywords*—**Multilayer Perceptron, Auto-Regressive Model, HIV, ROC Analysis.

I. Introduction

THE aim of this work is to determine MLP structure with the optimal hidden units for high classification accuracy of the preprocessed HIV dataset. The statistical data of HIV genome was modeled and analyzed by the signal processing approach [1].

In many applications on biomedical and bioinformatics, structures were modeled to determine HIV sub-type viruses. For instance, the infection effects of the chemokine coreceptor and virus entry were introduced by Berger et al. [2]

The Artificial Neural Networks (ANNs) structures with high classification ability were evolved by using bioinformatics models [3-7]. Lamers et al. [8, 9] used HIV-Base software to extract statistical data of R5X4, R5 and X4 viruses and classified these viruses with the ANNs. The performance of the ANNs in those works could not illustrate the actual performance of the system since the training accuracy was given as the classification result.

Z. K is with Computer Engineering Department, Yildiz Technical University, 34349, Istanbul, Turkey (e:mail: zeyneb@ce.yildiz.edu.tr).

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

AR model was used to process gene datasets in many applications. H. Zhou and H.Yan proposed that method in spectral analyses of short tandem in DNA sequences [10]. M. Akhtar et al. determined period -3 behaviors by using AR model [11]. G. Rosen also reduced the dimension of gene sequence via using AR model [12].

In this work, accessible residues of gene sequences were obtained by using Bioinformatics Toolbox in MATLAB 7.1. Since the dimension of gene sequences is large and different from each others, their dimension was reduced and equalized by AR model. This pre-processed dataset was used to train and test the MLP with changeable numbers of hidden units without using any toolboxes.

This paper consists of 3 sections. In Section 2, dataset, AR model, cross validation, MLP and ROC analysis are described. Also, the experimental results are given in this Section. In Section 3, conclusion and future work are mentioned.

II. Materials and Methods

A. Data Mining

77 R5 sequences, 31 R5X4 sequences and 40 X4 sequences [9] were downloaded from 148 Los Alamos National Laboratory HIV Sequence Database including 148 data in total (www.hiv.lanl.gov/content/hiv-db/main-page.html).

These sequences were converted into numeric data by using accessible residues as shown in Table 1. This dataset could not be used in MLP because of dimension of gene sequences is different from each others. That can be remedied by AR model in this study [1].

B. AR Model

AR model was chosen to model HIV data, since that model represents energy of signals successfully which 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,

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

where x_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[13].

In this work, 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 commonly used in examining the robustness of classifiers. The original dataset is split into two groups and one is used to design the classifier while the hold-out group is used for testing purposes [14]. 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 the error criterion interests you.

In this study, a 3-fold cross validation was used. There are 117 R5 or X4 samples and 31 R5X4 samples in the dataset. Due to one of the classes has much more instances than other, instances of less crowded class were cloned. Thus, both of the classes had 117 samples. Each class of HIV gene was partitioned into three pieces which consists of 39 R5 or X4 data and 39 R5X4 data respectively. One of the data set was used for testing MLP, while the remaining was used for training. The training and test sets consisted of 156 and 78 data, respectively.

D. Evolving MLP Structures

Accessible residues of HIV sequences were modeled by 10-th AR coefficients as mentioned. Thus, size of the accessible residues of HIV sequences was reduced and unnecessary details of the signals were eliminated. In the next step, this pre-processed dataset was used for training and testing MLP with various numbers of hidden units to classify R5X4, R5 and X4.

MLP system was shown in Fig. 1. a_i and y represent input and output of MLP, respectively.

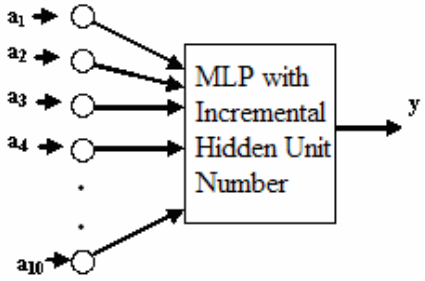

Fig. 1 The Proposed MLP structure

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

MLP had *10* input and one output neurons. The gradient-based back-propagation learning rule was used to determine the optimal hidden unit number. The unit number in the hidden layer began *1*; incremented by *1*; until *30*. The performance of MLP depends on initial conditions. Hence, the training and testing processes were repeated *20* times. The results were obtained by averaging experimental results of 3-fold CV dataset.

The accuracy of training and test steps according to hidden unit numbers (H) are shown in Fig. 2 and 3, respectively.

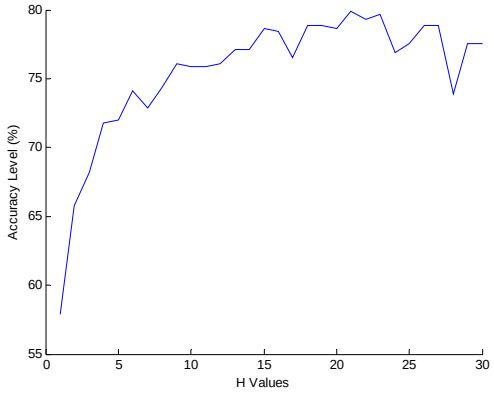

Fig. 2 The classification accuracy of training step according to hidden unit number

The best result of training and test steps were obtained when the hidden unit number (H) was chosen as 21 and 20, respectively. Besides, the mean square error (MSE) for training and test steps according to hidden unit numbers are given in Fig.4 and Fig. 5 respectively.

As shown in Fig 4 and Fig 5, the minimum MSE errors for training and test processes were acquired when the hidden unit number (H) was 22 and 18, respectively.

However, the classification accuracy and MSE are not enough to analyze the performance of MLP. Therefore, ROC analysis was applied to these results.

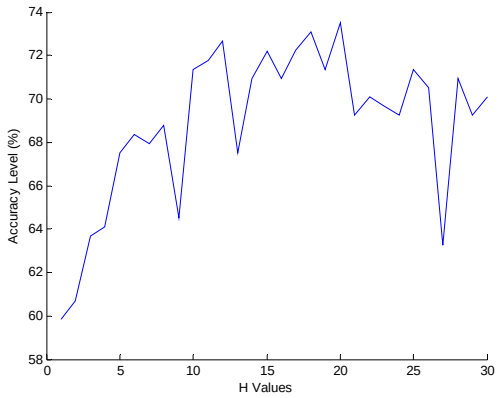

Fig. 3 The classification accuracy of test step according to hidden unit number

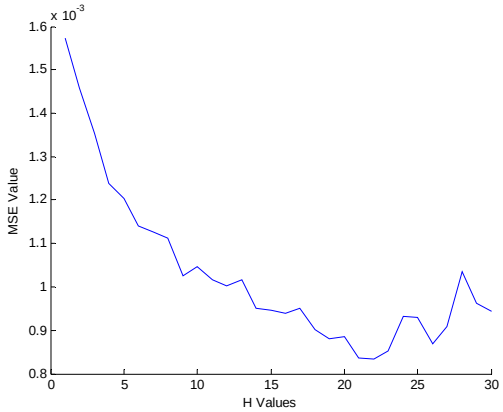

Fig. 4 The MSE for training process according to hidden unit number

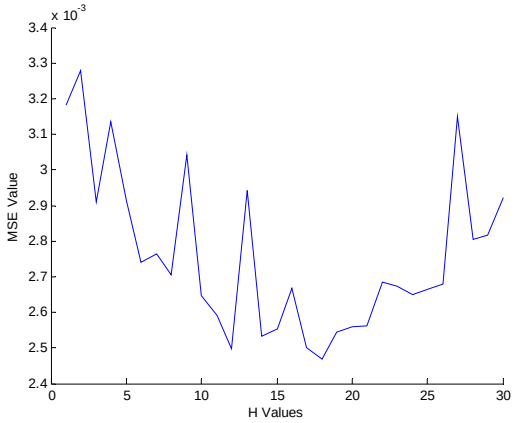

Fig. 5 The MSE for test process according to hidden unit number

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 [1,15].

TABLE II
ROC BLOCK DIAGRAM

		Actual	
		T	F
Predicted	T	True Positives (TP)	False Positives (FP)
	F	False Negatives (FN)	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 II.

The sensitivity and specificity can be written as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Specifity = \frac{TN}{TN + FP} \tag{4}$$

In this study, "sensitivity" and "specificity" of MLP could be defined as follows:

$$Sensitivity = \frac{R5X4_{True}}{R5X4_{True} + (R5\ or\ X4)_{False}} \tag{5}$$

$$Specificity = \frac{(R5\ or\ X4)_{True}}{(R5\ or\ X4)_{True} + R5X4_{False}} \tag{6}$$

"Sensitivity" and "specificity" of the most successful ANN should be approximated to *1*. The results of ROC analysis of MLP for training and test processes according to hidden unit number are given in Fig 6, 7, 8 and 9 respectively.

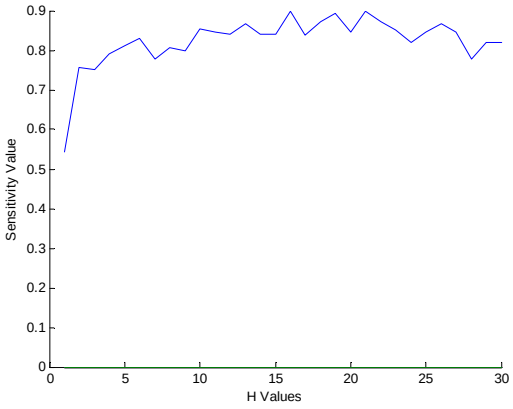

Fig. 6 The training sensitivity according to hidden unit number

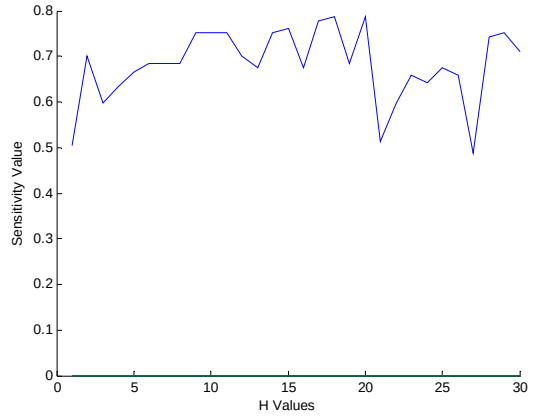

Fig. 7 The test sensitivity according to hidden unit number

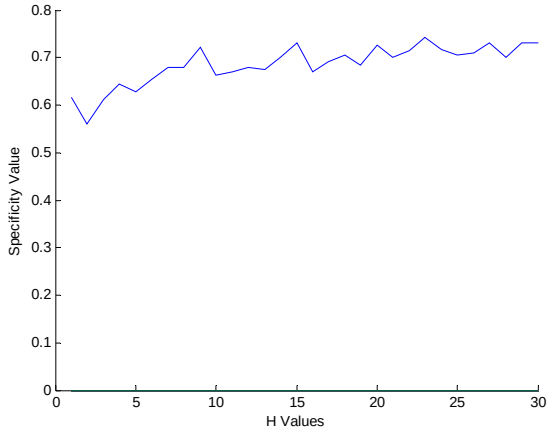

Fig. 8 The training specificity according to hidden unit number

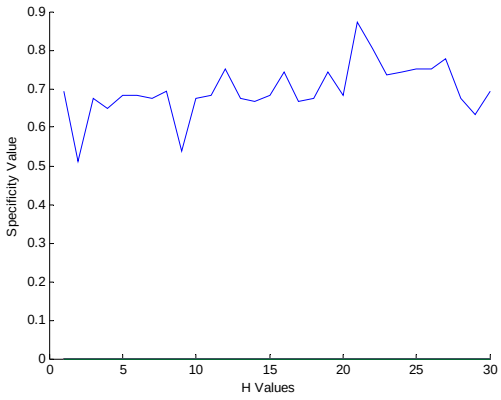

Fig. 9 The test specificity according to hidden unit number

The best result of sensitivity for training and test processes are shown while the hidden units are 16 and 20, respectively. Besides, according to the experimental results, the hidden units for training and test processes are chosen 23 and 21 respectively to obtain maximum specificity values.

F. Pseudo Code of the System

```
for experimentNumber = 1 to 20
   // Weight initializations for MLP
   for crossValidationOption = 1 to 3
      // Training and test data assignment for chosen cross
      //validation option.
      for hiddenUnitNumber = 1 to 30
       // Add a new hidden unit into hidden layer of MLP.
         // Initialize new weights for new MLP structure.
      while (not MLP converges)
            // Train the MLP architecture, whose number of the
            //hidden units equals to hiddenUnitNumber variable.
            // Calculate the training dataset error and make ROC
            //analysis for training data.
         end;
         // Calculate the test dataset error and make ROC
         //analysis for test data.
      end;
   end;
end;
```

II. Conclusion and Future Work

In this study, since gene sequences have different and large dimensions, the statistical data of HIV subtype genes were modeled and analyzed by AR model to reduce the dimension of gene sequences. By using this pre-processed data; the MLP structure with optimal hidden unit number was evolved to classify HIV subtype viruses successfully. The training and test dataset were obtained by using 3-fold cross-validation and these datasets were used to train and test MLP. The best training and testing accuracy were obtained from 21 and 20 hidden units. Also while hidden units were 22 and 18 respectively, the minimum MSE of training and test processes were extracted. Due to the classification accuracy is not enough to analyze the performance of MLP, ROC analysis was applied to these results. The sensitivity and specificity were approximated to *1*, when hidden units were 16 and 23 for training process, 20 and 21 for test process, respectively.

The experimental results and all performed analysis show that the optimal hidden unit number is about 20 to classify HIV gene dataset robustly.

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

References

[1] O. Yavuz, and L. Ozyilmaz, " Analysis and Clasification of HIV-1 Sub-Type Viruses by AR Model through Artificial Neural Networks," *Proc. of World Academy of Science, Engineering and Technology*, vol. 49, pp. 826-831, January 2009.

[2] 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.

[3] 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.

[4] 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.
[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] 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
[10] 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.
[11] 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
[12] G. Rosen. " Comparison of Autoregressive Measures for DNA Sequence Similarity " IEEE Genomic Signal Processing and Statistics Workshop (GENSIPS) pp. 13-17 2007.
[13] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, New Jersey, 2002.
[14] 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.
[15] 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