

PREDICTION OF BIOLOGICAL ACTIVITIES OF HIV-1 PROTEASE INHIBITOR COMPOUNDS WITH INTEGRATED SOM-FUZZY ARTMAP

Latha Parthiban¹, Rangasamy Parthiban²

¹SSN College of Engineering

²Sri Venkateswara College of Engineering

Email: ¹lathap@ssn.edu.in, ²lathaparthiban@yahoo.com

ABSTRACT

In this paper, the biological activities of HIV-1 protease inhibitor compounds are predicted with Standard Fuzzy ARTMAP (FAM) and GA-FAMR, from the attributes describing the molecular descriptor of the compounds. Self-organized maps (SOM) have been applied to analyze the similarities of chemical compounds and to select from a given pool of molecular descriptors the smallest and more relevant subset needed to build robust QSAR (Quantitative Structure-Activity Relationship) models based on FAM. FAM is provided with 196 sets of data, out of which 176 are used for training and the remaining 20 are used for testing. The data are normalized and fed to the neuro-fuzzy network and the output indicates whether the compound is a suitable inhibitor for the HIV-1 virus. The analysis is done with and without Genetic Algorithm (GA) for the small dataset and GA-FAMR algorithm is used to optimize the relevance's assigned to the training data. The performance of the integrated SOM-FAM was evaluated and the prediction accuracy obtained is 93.09%.

Key words: Fuzzy ARTMAP, protease inhibitor, Self Organizing Map, Genetic Algorithm, QSAR

I. INTRODUCTION

Currently human experts are seriously investigating a possible treatment for the HIV-1 virus. In particular, the use of HIV protease inhibitors has revolutionized AIDS care but the real problems associated with the currently marketed protease inhibitors are low bioavailability, significant side effects and drug interactions, inconvenient dosing schedules, and high cost of drug [1]. More ominous is the emergence of HIV strains which are resistant to current therapies. Therefore, the search for a new generation of inhibitors which are not cross resistant to the current agents continues unabated. The classification of compounds can greatly help in contributing to find a suitable compound to resist the action of the virus. Since the prediction is being automated, the diagnosis time and cost reliance on human experts can be reduced. The classification abilities and resilience against noise exhibited by neural networks are characteristics desired for predicting the biological activities of newly designed potential HIV-1 protease inhibitors [15].

Neural models have been used to predict properties of chemical compounds such as inhibition of HIV-1 reverse transcriptase [1], lipophilicity and aqueous solubility [2], intestinal absorption [3], and site of protease cleavage [4]. Several neural architectures were successful for such tasks; among these are

backpropagation [1], [5], [6], associative neural networks [2], probabilistic neural networks [7], generalized regression neural networks [8], radial basis function networks [4], [8], cascade correlation [9], neural networks trained via evolutionary algorithms [10], fuzzy ARTMAP [11]–[13], and support vector machines [14].

Due to lack of availability of inhibitor compounds [17] only a small molecular data set has been obtained which is not sufficient enough to train a neural net so as to obtain high efficiency of the system. For this purpose the neuro-fuzzy-genetic hybrid is designed for better performance and efficiency. The FAM neural networks have several advantages due to their ability to classify and analyze noisy information with fuzzy logic, and to avoid the plasticity-stability dilemma of other neural architectures [16].

GA-FAMR algorithm is used to optimize the relevances assigned to the training data. When the relevance of the training data is optimized, the FAMR system is trained with the training data. The system eventually learns while going through all the input data. When the compounds with similar molecular descriptors occur in the training set, the net will usually give the same result.

The paper is organized as follows. Sec. 2 explains the prediction of HIV-1 inhibitor input, Sec 3 explains the neuro-fuzzy GA, Sec.4 & 5 provides the

implementation of FAM and GA-FAMR, Sec. 6 the performance measures and Sec. 7 the conclusions.

II. IC₅₀ PREDICTION OF HIV-1 PROTEASE INHIBITORS INPUT

The current clinical approach for the therapy of HIV/AIDS utilizes the co-administration of two reverse transcriptase inhibitors with one protease inhibitor (usually referred to as combination therapy). Combination therapy [22] reduces viremia to very low levels; however, in 30- 50% of patients, antiviral therapy is ineffective due to resistance development. Furthermore, in many patients, side effects associated with these drugs pose serious problems. Due to current drug resistance and toxicity, there is an urgent need for the development of more efficient drugs with different resistance profiles, decreased toxicity and more than one type of inhibitory action. Thus, new methods for designing enzyme inhibitors, and for predicting their properties, are expected to have great value in drug discovery [21].

In our study, 30 molecular descriptors were selected using SOM which analyzes the similarities of chemical compounds and select from a given pool of descriptors the smallest and more relevant subset needed to build robust QSAR models. First, the category maps for each molecular descriptor and for the target activity variable were created with SOM and then classified based on topology and nonlinear distribution. The best subset of descriptors was obtained by choosing from each cluster the index with the highest correlation with the target variable and then in order of decreasing correlation. This process was terminated when a dissimilarity measure increased, indicating that the inclusion of more molecular indices would not add supplementary information. The optimal subset of descriptors was used as input to a FAM and the descriptors are displayed in Table I.

SYBYL 3 molecular modeling software was used to create the data files that provided descriptors. Descriptors were then normalized according to the formula $(x - y)/z$, where x is the actual value for the descriptor of the inhibitor of interest, y is the minimum descriptor value for the data set, and z is the range (maximum-minimum) value for the data set. As a result of normalization, descriptors for all the molecules in the data set fell in the range [0, 1].

The data set for training and testing consisted of 151 known compounds with experimentally determined

IC_{50} values. The IC_{50} value represents the concentration of inhibitors that is required to reduce enzyme activity by 50% [21]. A total of 26 novel compounds were designed via a 'combinatorial' approach using only known compounds with low IC_{50} and high TI values 4. Side chains were then modeled on all core structures, excluding the one from which they came from, into their original binding pockets. More details on data description can be had from [21].

Table 1. Molecular Descriptors

Number	Descriptor
D1	Total Number of Atoms
D2	Number of Bromine Atoms
D3	Number of Carbon Atoms
D4	Number of Chlorine Atoms
D5	Number of Fluorine Atoms
D6	Number of Hydrogen Atoms
D7	Number of Nitrogen Atoms
D8	Number of Oxygen Atoms
D9	Number of Sulfur Atoms
D10	Molecular Volume
D11	Index of Hydrogen Deficiency
D12	Total Number of Bonds
D13	Number of Single Bonds
D14	Number of Double Bonds
D15	Number of Triple Bonds
D16	Number of Aromatic Rings
D17	Number of Amide Bonds
D18	Molecular Weight
D19	Total Charge
D20	Bond Stretching Energy
D21	Angle Bending Energy
D22	Torsional Energy
D23	Out of Plane Bending Energy
D24	One to Four Van Der Waals Energy
D25	Van Der Waals Energy
D26	One to Four Electrostatic Energy
D27	Electrostatic Energy
D28	Van Der Waals Electrostatic Pairs
D29	One to Four Van Der Waals Electrostatic Pairs
D30	Scaled Van Der Waals Electrostatic Pairs

Artificial neural networks are particularly well-suited for QSPR (Quantitative Structure-Property Relationship) and QSAR because of their ability to extract both linear and nonlinear information present in the mapping of physio-chemical descriptors to biological activity[18]

III. NEURO-FUZZY GENETIC ALGORITHM

The FNN was implemented [21] according to a modification of the Min-Max Fuzzy Inference Network (MMFIN) The first and second layers are connected by fuzzy membership functions. The second and third layers are connected by a weight matrix, w_2 . During the training process, the final output is compared to the target value, which is the known IC_{50} value of the training legend. If the output differs from this known value by more than the established error tolerance, the membership functions and w_2 weight matrix are adjusted and this legend must be "re-learned." If it is not possible to "re-learn" this molecule so that it falls within the ranges of the already established prototypes (hidden layer neurons), then a new neuron is added to the hidden layer. Training continues until all training molecules are learned and produce output values within the acceptable error tolerance. During training, the calculated output is compared to the target value plus or minus the error tolerance. A high error tolerance thus results in a very compact structure, which has a low predictive ability, but generalizes well. Therefore, the generalization degree of the network is controlled by the error tolerance and FNN quickly learns all training patterns.

Since the FNN architecture is determined only by the error tolerance, the challenge is how to find the best value for this parameter. The goal of optimizing the structure of the FNN with a specified subset of descriptors is to produce the most compact network possible, which still maintains a high prediction and generalization ability. These optimization criteria are measured by the fitness function f , which incorporates both the degree of compactness (the number of hidden nodes) and the predictive ability of the FNN and the fitness is evaluated.

GA is used [19] to optimize the parameter of the FNN. The objective of the GA optimization is to find the optimum balance between the extremes of a highly compact network which is unable to predict accurately, and a very loose, over-fitted network which is not able

to generalize. In the implementation, the same GA in two different instances: to determine the optimal subset of features and to determine the optimal error tolerance for the FNN. Hence, the system is used to compute the IC_{50} value of compounds with an error tolerance of 13.6%.

A. Issues with the existing system

Obtaining satisfactory results with neural networks depends on the availability of large data samples. With small training sets, performance may be reduced, or learning task may not be accomplished which limits ANN severely. The main reason small datasets cannot provide enough information is that there exist gaps between samples and complete coverage of domain by samples cannot be ensured. For a small training set, even a simple neural network can have a complexity (e.g., number of connections/parameters) that is comparable to, or exceeds, the size of the training set.

One of the problems with the system during cross-validation is that the training subsets overlap. This overlap may prevent obtaining a good estimate of the amount of variation that would be observed if each training subset were completely independent of previous training subsets. The GA-FAMR algorithm can be improved so as to increase the efficiency of the system with limited number of available data sets. This can be achieved by:

- retuning of the FAM parameter.
- modifying the training sets.

Improving GA can make the generalization capability of the relevance data set more sufficient [20].

The GA is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. A typical genetic algorithm requires

- (i) genetic representation of the solution domain.
- (ii) fitness function to evaluate the solution domain.

The input data is a mol file which can be viewed with the help of the software known as SYBYL-X 1.1. The figure 1 shows a sample test molecule file which is viewed with the help of the SYBYL software

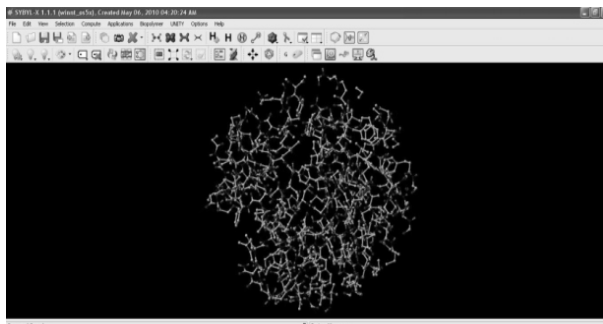


Fig. 1. Sample Molecule file

The four main molecular descriptors used in our system are given in Table 2

Table 2. Molecular Descriptors Used

	Mean	Deviation	Min	Max	Variance
Molecular Weight	567.89	140.92	24.74	312.4	822.99
H - bond Donors	3.070	2.09	67.78	1.00	8.00
H - bond Acceptor	7.05	3.44	48.64	3.00	14.00
Clog P	6.38	1.42	2.42	9.96	22.2

Using SYBYL, the molecule file is processed to obtain the mean, deviation, min, max and variance value for the four descriptors mentioned. The resultant data is normalized to obtain an integer value in the range [0,1]. It is arranged in matrix format and stored

in the input vector one for both the training set and test set data.

IV. IMPLEMENTATION OF THE FAM SYSTEM

The FAM system (figure 2) includes a pair of ART modules (ARTa and ARTb) that create stable recognition categories in response to arbitrary sequences of input patterns. These modules are linked by an inter-ART module called Map field whose purpose is to determine whether the correct mapping has been established from inputs to outputs or not. The ARTa and ARTb vigilance parameters a , respectively b , control the matching mechanism inside the modules[15].

The main parameters required for computation in standard FAM system are

- Choice parameter $\alpha > 0$
- Learning parameter $\beta \in [0,1]$
- Vigilance parameter $\rho \in [0,1]$

The resultant IC_{50} values for the twenty test set compounds using the Standard FAM system is given in the figure 3. The output values are stored in a vector variable class.

The number of FAMR categories is not influenced by the relevance optimization, but only by the FAMR parameters, which are kept constant. Therefore, the number of ARTa categories is 13 and the number of

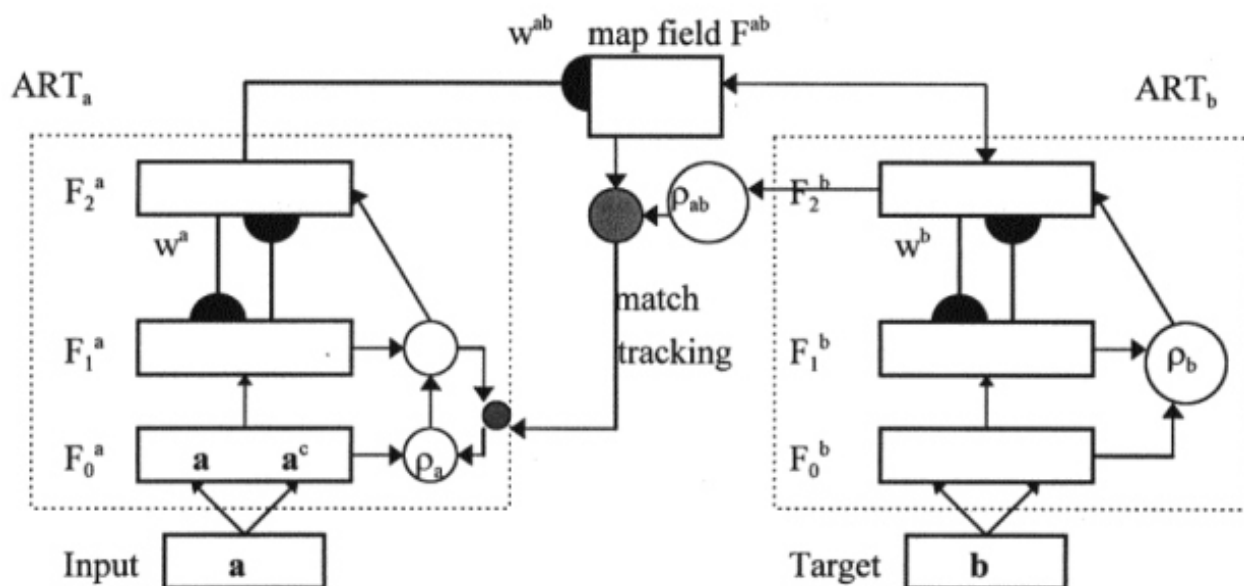


Fig. 2. Fuzzy ARTMAP architecture [15]

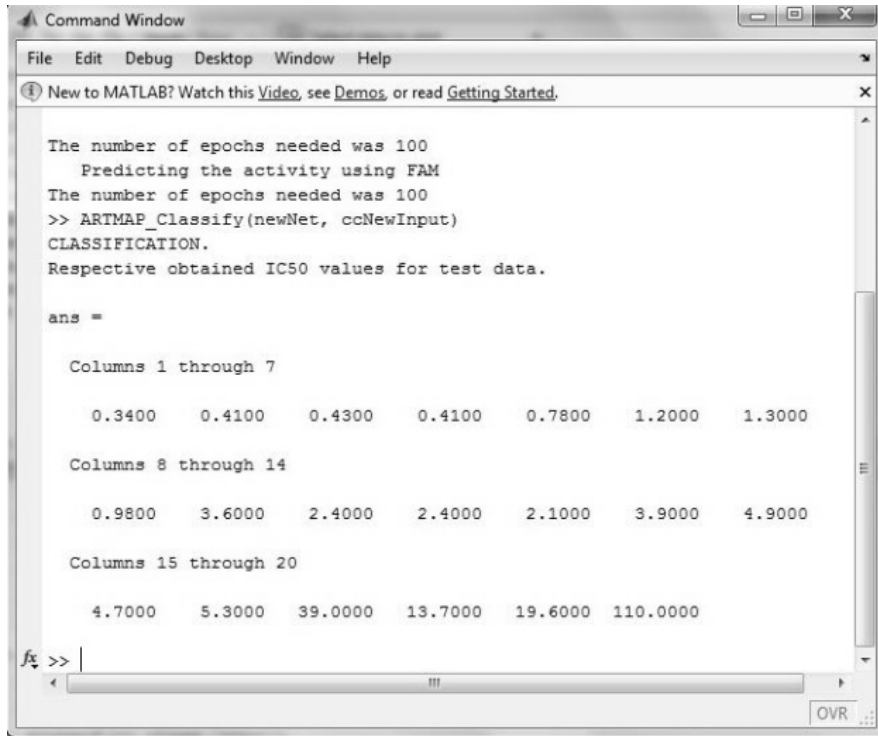


Fig. 3. Output of FAM system

ARTb categories is 8. Before optimization, all relevances are equal; after optimization they vary in the range (0, 10).

V. IMPLEMENTATION OF GA-FAMR SYSTEM

The GA-FAMR (figure 4) operates on an initial population of relevance vectors. Each relevance vector

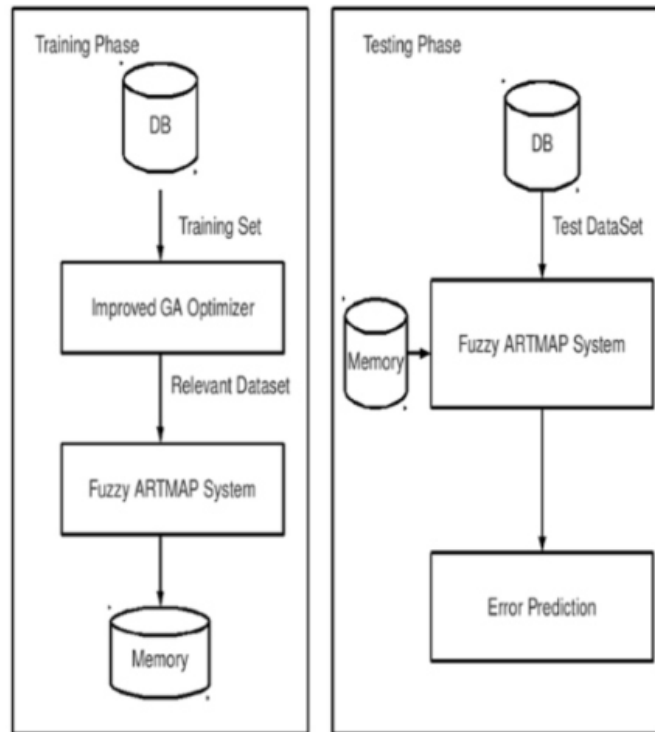


Fig. 4. GA-FAMR System Architecture

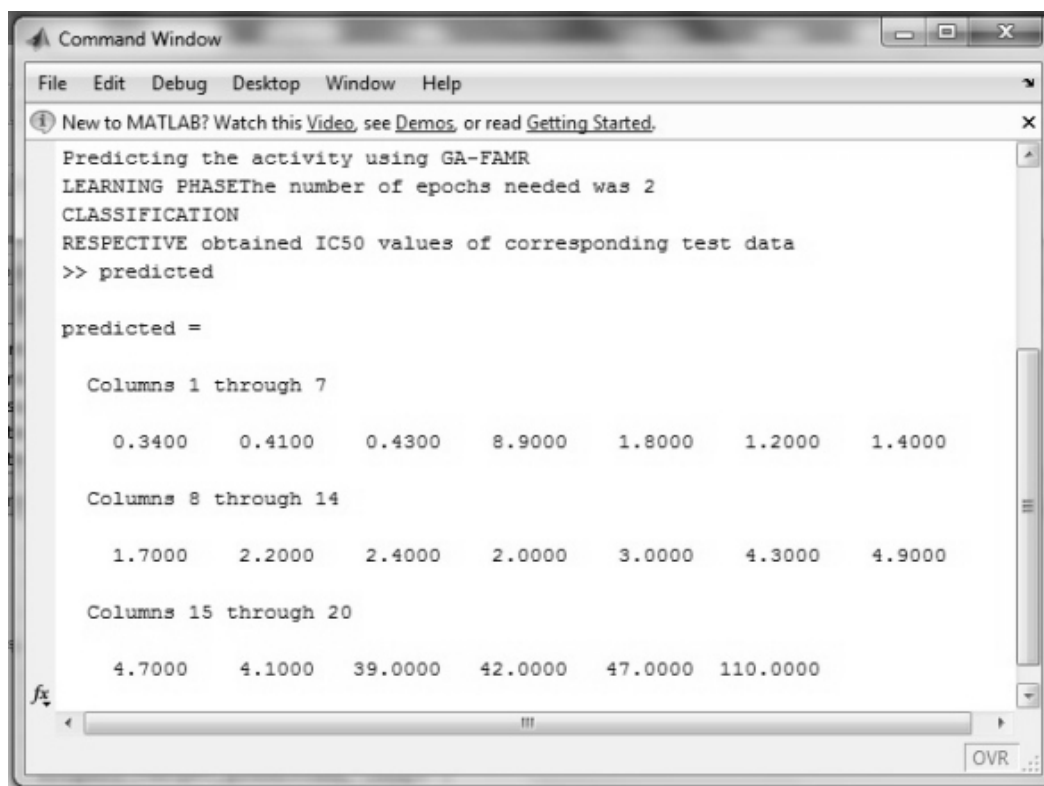


Fig. 5. Output using GA-FAMR system

has a single relevance associated with a specific training data in accordance with the FAMR. Because the relevance of specific data is not known beforehand, this population must be optimized using the following GA

Initially the relevance is set to 1 during the initial learning phase of the system. In the GA optimizer the relevance factor is optimized by a set of 1000 generations. The relevance factor is assigned to each sample pair, proportional to the importance of the respective pair during the learning phase. The resultant IC_{50} values for the twenty test set compounds using the GA-FAMR system is given in the figure 5. The output values are stored in a Vector variable class.

VI. PERFORMANCE MEASURES

In our experiments, training is done with same set of 176 molecules and the percentage error in the values of IC_{50} are computed and calculated for both systems, namely standard FAM (Figure 6) and GA-FAMR (Figure 7). The percentage of error is computed by using the Symmetric Mean Absolute Percentage Error (sMAPE). using the actual target and the predicted target.

. MAPE is computed using the formula,

$$200/k \times \sum |d - y| / (d + y)$$

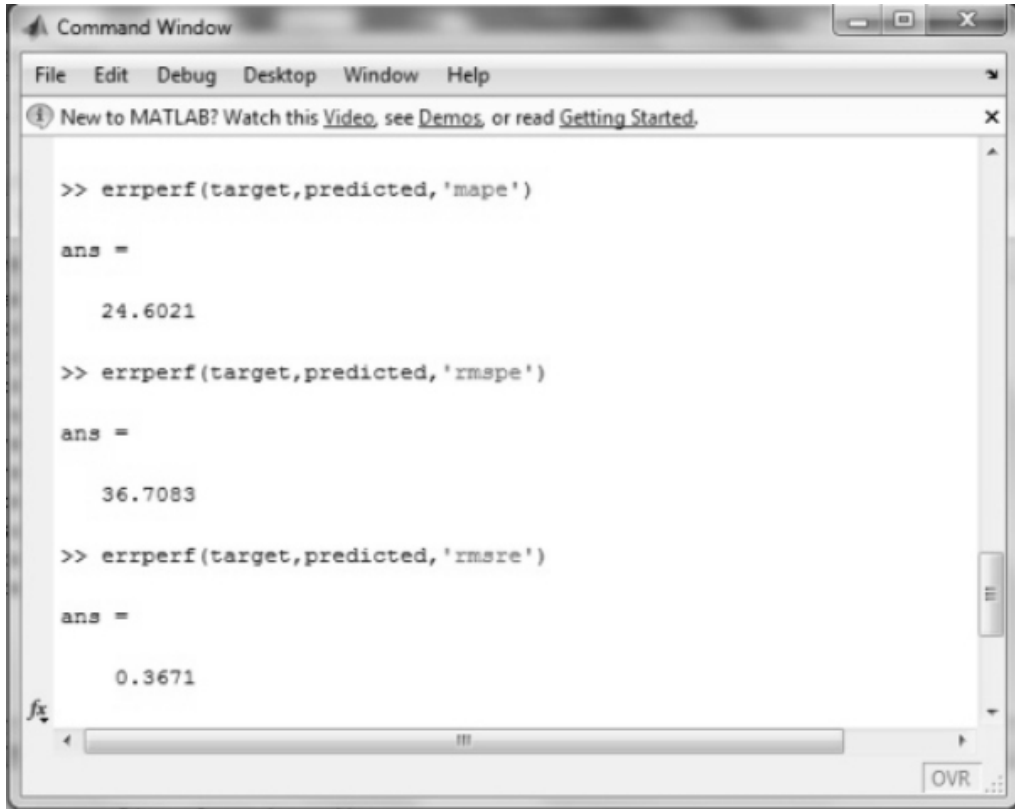
where k is the number of samples, d is the target output and y is the predicted output. Errors computed with other metrics are given in Table 3.

Table 3. Error Computed with Various Metrics

Algorithm	MAPE	RMPSE	RMSRE
Standard FAM	24.6	462.86	4.63
GA FAMR	6.91	10.91	0.11

VII. CONCLUSION

In this paper, we compare the percentage of error on two computational techniques and prove that the GA-FAMR system is much more efficient and accurate than standard FAM system for prediction of biological activities of HIV-1 inhibitors. We optimized not only the relevances, but also the parameters of the FAMR network (ρa , ρb , βa , and βb) using the GA approach and the prediction performance is improved. The optimizations, performed in the Standard FAM and the GA-FAMR, ameliorate the problem of insufficient data.



```
Command Window
File Edit Debug Desktop Window Help
New to MATLAB? Watch this Video, see Demos, or read Getting Started.

>> errperf(target,predicted,'mape')

ans =

    24.6021

>> errperf(target,predicted,'rmspe')

ans =

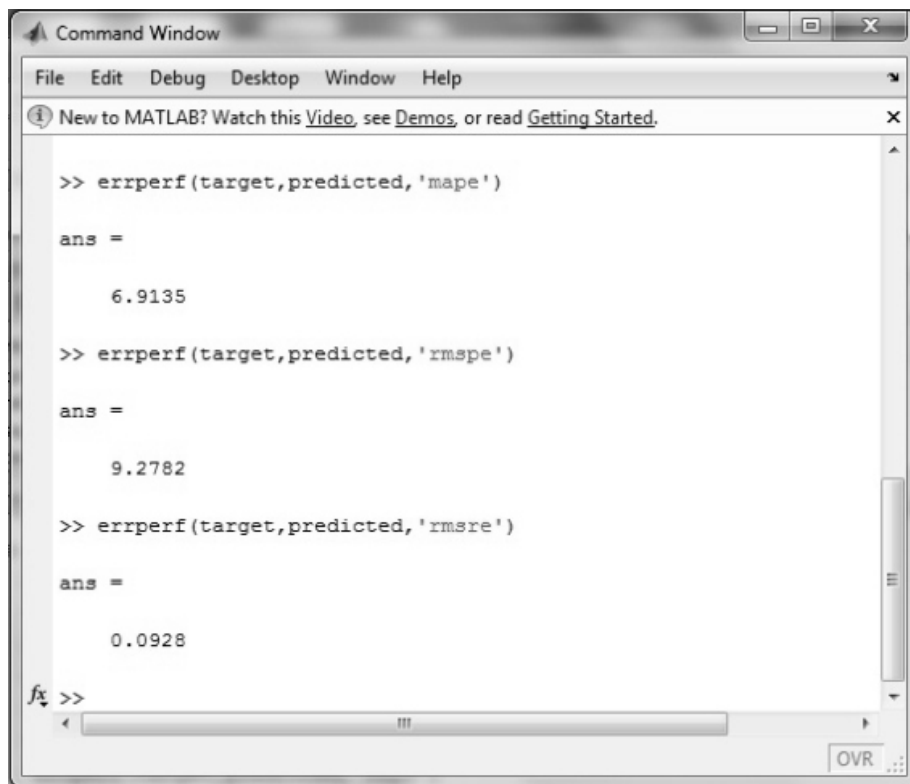
    36.7083

>> errperf(target,predicted,'rmsre')

ans =

    0.3671
```

Fig. 6. Error measurement of FAM system using various metrics



```
Command Window
File Edit Debug Desktop Window Help
New to MATLAB? Watch this Video, see Demos, or read Getting Started.

>> errperf(target,predicted,'mape')

ans =

    6.9135

>> errperf(target,predicted,'rmspe')

ans =

    9.2782

>> errperf(target,predicted,'rmsre')

ans =

    0.0928

fx >>
```

Fig. 7. Error measurement of GA-FAMR system using various metrics

So, the most useful predictive tools for the medicinal chemist appear to be the GA-FAMR

ACKNOWLEDGEMENT

We acknowledge National Cancer Institute (NCI) for providing publicly available chemical compound repository for our research.

REFERENCES

- [1] Tetko I. V., Tanchuk V. Y., and Luik, A. I. 1994, Evaluation of potential HIV-1 reverse transcriptase inhibitors by artificial neural networks, Proceedings of the Seventh Annual IEEE Symposium on Computer-Based Medical Systems, pp. 311–316.
- [2] Tetko I. V. and Tanchuk, V. Y., 2002, Application of associative neural networks for prediction of lipophilicity in ALOPS 2.1 program, Journal of Chemical Information and Computer Sciences, vol. 42, pp. 1136–1145.
- [3] Niwa T., 2003, Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures, Journal of Chemical Information and Computer Sciences, vol. 43, pp. 113–119.
- [4] Yang Z. R. and Thomson R., 2005, Bio-basis function neural network for prediction of protease cleavage sites in proteins, IEEE Transactions on Neural Networks, vol. 16, pp. 263–274.
- [5] Tetko I. V., Luik A. I., and Poda G. I., 1993, Application of neural networks in structure-activity relationships of a small number of molecules, Journal of Medicinal Chemistry, vol. 36, pp. 811–814.
- [6] Devillers J., 1996, Designing molecules with specific properties from intercommunicating hybrid systems, Journal of Chemical Information and Computer Sciences, vol. 36, pp. 1061–1066.
- [7] Niwa T., 2004, Prediction of biological targets using probabilistic neural networks and atom-type descriptors, Journal of Medicinal Chemistry, vol. 47, pp. 2645–2650.
- [8] Potocnik P., Grabec I., Setinc M., and Levec J., 2000, Hybrid modelling of kinetics for methanol synthesis, in Soft Computing Approaches in Chemistry, H. Cartwright and L. M. Sztandera, Eds. Heidelberg:Springer-Verlag.
- [9] Bianucci A. M., Micheli A., Sperduti A., and Starita A., 2000, Application of cascade correlation networks for structures to chemistry,” Applied Intelligence, vol. 12, pp. 117–147.
- [10] Weekes D and Fogel G. B, 2003, Evolutionary optimization, backpropagation, and data preparation issues in qsar modeling of HIV inhibition by hept derivatives,” Bio Systems, vol. 72, pp. 149–158.
- [11] Yaffe D., Cohen Y., Espinosa G., Arenas A., and Giralt F., 2002, Fuzzy ARTMAP and back-propagation neural networks based quantitative structure-property (QSPRs) for octanol-water partition coefficient of organic compounds,” Journal of Chemical and Information Sciences., vol. 42, pp. 162–183.
- [12] Espinosa G., Yaffe D and Arenas A., 2001, A fuzzy ARTMAP based qualitative structure-property relationship (QSPR) for predicting physical properties of organic compounds, Journal of Chemical and Information Sciences, vol. 40, pp. 2757–2766.
- [13] Espinosa G., Arenas A., and Giralt F., 2002, An integrated som-fuzzy ARTMAP neural system for the evaluation of toxicity, Journal of Chemical and Information Sciences, vol. 42, pp. 343–359.
- [14] Yao X. J., Panaye A., Doucet J. P, Zhang R. S., Chen H. F., Liu M. C., Hu Z. D., and Fan B. T., 2004, Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression, Journal of Chemical Information and Computer Sciences., vol. 44, pp. 1257–1266.
- [15] Andonie, R., Fabry-Asztalos L., Abdul-Wahid C. B, Abdul-Wahid S., Barker G. I and Magill L. C., 2011, Fuzzy ARTMAP prediction of biological activities for potential HIV-1 protease inhibitors using a small molecular dataset”, IEEE/ACM Transactions on Computational Biology and Bio-informatics, vol. 8, pp. 80-93.
- [16] Andonie R., Fabry-Asztalos L., Abdul-Wahid C. B., Abdul-Wahid S., 2009, Fuzzy ARTMAP Rule Extraction in Computational Chemistry, Proceedings of international joint conference on neural networks, USA pp.157-163.
- [17] Andonie R. and Sasu L., 2006, Fuzzy ARTMAP with input relevance, IEEE Transactions on Neural Networks, vol. 17, pp. 929 -941.
- [18] Andonie R., Fabry-Asztalos L., Abdul-Wahid S., Collar, C. and Salim N., 2006, An integrated soft computing approach for predicting biological activity of potential HIV- 1 protease inhibitors” Proceedings of the IEEE International Joint Conference on Neural Networks, Canada, pp. 7495-7502.
- [19] Fabry-Asztalos L., Andonie R., Collar C., Abdul-Wahid, S. and Salim N., 2008, A genetic algorithm optimized fuzzy neural network analysis of the affinity of inhibitors for HIV-1 protease”, Bio-organic and Medicinal Chemistry, vol. 16, pp. 2903-2911.

- [20] Andonie R., Fabry-Asztalos L., Magill L., and Abdul-Wahid S., 2007, A new Fuzzy ARTMAP approach for predicting biological activity of potential HIV-1 protease inhibitors, Proceedings of the IEEE International Conference on Bio-informatics and Biomedicine, I. C. S. Press, Ed., San Jose, CA, pp. 56-61.
- [21] Andonie R., Fabry-Asztalos L. Collar C., Abdul-Wahid, S. and Salim N., 2005, A neuro-fuzzy prediction of biological activity and rule extraction of HIV-I protease inhibitors, Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB'05), San Diego, November 14-15, pp.113-120.
- [22] <http://www.hiv.lanl.gov/content/index>



Dr. Latha Parthiban obtained her B.E from Madras University, M.E from Anna University and PhD from Pondicherry central university. She has published 5 books and 9 international journal papers.



Dr. Rangasamy Parthiban completed his B.Tech from CIT M.Tech from REC and PhD from Anna University. He has published 7 books and over 28 international/national publications.