Crossref
Similarty Check
Powered by iThenticate

CONSTRUCTION AND COMPARISON OF GENE REGULATORY NETWORKS OF HUMAN HIV-1 VPR MICROARRAY DATASETS BY RADIAL BASIS NEURAL NETWORK APPROACH

Bandana Barman

Department of Electronics and Communication Engineering, Kalyani Government Engineering College, Kalyani- 741235, Dist. Nadia, West Bengal
Email: bandanabarman@gmail.com

Paper received on: Septenber 20, 2016, accepted after revision on: July 22, 2017
DOI:10.21843/reas/2016/61-69/158777

Abstract: Gene Regulatory Network (GRN) construction by using neural network approach is very important and useful approach for analyzing microarray gene expression microarray datasets. The human HIV-1 Vpr mutant microarray time series gene expression value carries the experimentally validated interaction records. Firstly, the subtractive clustering approach is used to cluster the microarray data. Secondly, GRN is constructed within cluster centers of HIV-1 Vpr mutant dataset using Radial Basis Neural Network approach. The optimized output of genetic network is found using genetic algorithm. Then the influence of range of cluster centers of data clusters and also GRN outputs are compared. It is found that proximity of gene expressed values in wild type cell line HIV-1 is higher than other two HIV-1 Vpr mutants. In this paper, the nature of network output functions are also identified.

Keywords: AIDS; HIV-1 Vpr mutants; time series microarray data;subtractive clustering;Genetic Network; Radial Basis Neural Network; Genetic Algorithm.

1. INTRODUCTION

Acquired Immunodeficiency Syndrome (AIDS) is a life-threatening infectious condition in human which may destroy total immunity system. The computational tools are sincerely utilized to predict viral-host interactions. Human Immunodeficiency Virus (HIV) is a RNA-lentivirus which is reversely transcribed to its genomic RNA to DNA [1,2]. This reverse transcription process takes place by the enzyme reverse transcriptase. A portion of gp120 protein's V1 region is responsible to begin the replication cycle of HIV. The interactions of cellular and viral factors activate the HIV expression from latent state. Then HIV-mRNA is translated into proteins. The complexity of HIV-1 is more than other retroviruses. The virus uses CD4 cells to replicate and destroys them in the process. When the number of CD4 cell falls below 200 cells per cubic millimeter of blood (200 cells/mm^3), one considers to have progressed to AIDS. The Vpr i.e., viral protein R plays an important role to suppress the host cellular responses which is one of viral offensive strategies [3,4]. The genes of HIV-1 encode the structural proteins of the virus. Several computational approaches are used previously to analyze the interactions between virus and its host of HIV-1.

In this article, algorithm is developed to analyze the HIV-1 Vpr mutants. To obtain co-expressed genes, clustering of microarray data matrix is performed using subtractive clustering approach [5]. The computation in this clustering approach is linearly proportional to the problem size [6,7]. Classification problem is a vast problem to analyze this type of microarray data because the property of a human HIV-1 Vpr mutant cell is like an infected dendritic cell. By subtractive clustering method and genetic algorithm the derivation of effective and efficient dataset has been analyzed.

Genetic Network construction based on neural network approach is a research field nowadays. Different neural network approaches are used to generate gene regulatory network within microarray data. In this paper, gene regulatory network or genetic network within cluster-centers is constructed by using radial basis function neural network (RBN) based approach. This is a two-layer feed-forward neural network. The hidden layer of this network is a set of radial basis functions (RBF). The output layer of RBN implements the nodes of linear summation functions. The network training process is divided into two stages: In first stage, input layer to hidden layer weights are determined. In second stage, hidden layer to output layer weights are obtained. The training learning procedure is very fast [8,9]. The optimization technique is used to explore large and complex space in such a way that finds the values close to global optimum. Genetic algorithm (GA) is a good technique for optimization. In the many research fields, GAs are applied for optimization. Optimization of output of the gene regulatory network is performed by using genetic algorithm (GA) [10]. Here, the neural network output fitness function is being optimized.

2. MATERIALS

The derived algorithm is applied on microarray datasets of inducible HIV-1Vpr protein on cellular gene expression. The dataset is collected from http://www.ncbi.nlm.nih.gov.in/geodata/. The dataset contain cell lines express wild type (WT) HIV-1 Vpr and two mutant Vprs, F72A/R73A and R80A. Microarray cells were collected at 0, 1, 2, 4, 6, 8, 12, 16 and 24 hours post induction (hpi). The number of RNA present in the cell is 21,794. So, each microarray time series data matrix consists of 21794 rows with its 9 time point values. The algorithm is applied on HIV-1 microarray data after doing its normalization. The normalization is done to get mean and variance of each gene 0 and 1 respectively.

3. METHODS

At first subtractive clustering approach on each of three samples of HIV-1 Vpr microarray time series dataset is applied to get the co-expressed genes from the microarray data matrix. Then, construction of Gene Regulatory Network within cluster centers is discussed in this section. After developing genetic network by using Radial Basis Function Neural Network approach, network output objective functions are optimized by using genetic algorithm (GA). The entire algorithm is coded and implemented in Matlab (R2014a).

Reverse engineering is an knowledge to combine and to develop mathematical model with biological sample data (microarray data) and their other information. It is known that, proteolysis rate is inversely proportional to substrate amount and this rate is represented by ordinary differential equation (ODE) over space and time (Equation 1),

$$\frac{dG}{dt} = -K_p G(t) \tag{1}$$

where, G, $\frac{dG}{dt}$ are state variable and rate of change of state variable respectively. K_p, is constant and $G(t)$ is the change of state variable. In a microarray dataset, gene expression values change only with time but not with space. Ordinary Differential Equation (ODE) for genes present in microarray gene expression data matrix might be represented by partial differential equation (PDE) with time (Equation 2):

$$\frac{\partial G}{\partial t} = -K_p G(t) \tag{2}$$

Where, G is the gene expression in microarray time series data matrix and it can be represented by integrating the mentioned (Equation 2) Partial Differential Equation at any time.

3.1 Subtractive Clustering of HIV-1 Expression Data

The subtractive clustering algorithm is applied to cluster the microarray data matrices. The data matrices are wild type (WT) HIV-1 Vpr and two HIV-1 mutant Vprs. Subtractive clustering algorithm is the basis of fast and robust algorithm to identify fuzzy models. It is an extension of mountain clustering method also. This clustering process tries to obtain a new cluster center to revise potentials and which is amended to ease difficulty in establishing a very sensitive parameter. The difference between fuzzy c-means method (FCM) and subtractive clustering method (SCM) are estimating potential values, influences of neighboring data points. The subtractive clustering method considers each and every data point as potential cluster center and more neighboring data points will have higher opportunity to become cluster center than points with fewer neighboring data. Based on density of surrounding data points, potential value for each data point is calculated. The data values outside of this range has a little influence on potential [5,6].

At first, data point with highest potential is chosen as first cluster center after computing potential of every data point. The data points which are near first cluster center will have greatly reduced potential and therefore are chosen as next cluster center. Thus, cluster center selection process goes on. This process continues until no further cluster center is found. There are two parameters which are involved for chosing a new cluster center with data points. Those parameters are: (1) Accept Ratio (AR) (2) Reject Ratio (RR). Accept Ratio involves influence range and squash factor together with. Reject Ratio sets four criteria for selection of cluster centers in subtractive clustering method [7].

In this article, wild type HIV-1Vpr microarray data matrix is denoted as W(i, j), HIV-1Vpr mutants data (F72A/R73A and R80A) as F(i, j) and R(i, j) respectively. Here, ‘i’ is number of rows i.e., genes presents in the microarray, 21794 and ‘j’ is number of columns i.e. time points, 9. So, the size of all microarray data matrices are same, i.e., 21794 × 9.

The three sample data matrices, viz., W(i, j), F(i, j) and R(i, j) are clustered. The vector variable of entries between 0 and 1 that specifies a cluster center’s range of influence in each of data dimensions is called ‘radii’. The clustering is performed on each data matrix by using subtractive clustering approach in such a way so that total number of clusters for all datasets becomes same. The value of range of influence (‘radii’) for data matrix W(i, j) is 0.0299, for data matrices F(i, j) and R(i, j) are 0.05, and 0.0691 respectively. It is mentioned before that all datasets are HIV-1 infected dendritic cell so in this microarray data, genes’ (RNAs) expressed values are very close to each others. After doing clustering, total 46 clusters for every sample data matrix are obtained. As total time point values in each data matrix is 9, a size of 46 × 9, cluster center matrix for each of microarray time series data matrix is obtained. Three cluster center matrices are named as Cw(i, j), Cf (i, j), Cr(i, j) for wild type HIV-1 (WT), HIV-1 vpr mutants F72A/R73A and R80A respectively.

In next section, construction of three gene regulatory networks within cluster center matrices by using radial basis neural network approach will be discussed.

3.2 Genetic Network using Radial Basis Network (RBN) Approach

The main concept of Radial Basis Neural Network is based on artificial neural network which uses radial basis functions as activating functions. Radial basis neural network is used for function approximation, time series prediction, and to control the used data. The layer concept of RBN is like, first layer with input nodes, one hidden layer with radial basis functions and one layer with output nodes. The characteristic feature of Radial Basis Function (RBF) is that, their responses will increase or decrease with distances from a certain central point. So, parameters of this function are the center point, distance and a precise shape [8,9]. If this function is linear then all parameters are fixed.

A typical Radial Basis Function, R(x), by nature is a Gaussian function and it is defined as (Equation 3):

$$R(x)=e^{\frac{(x-c)^2}{r^2}} \qquad (3)$$

where, c is the center point and r is the width (spread). The three gene regulatory network or genetic networks with cluster center matrices, Cw(i, j), Cf (i, j), Cr(i, j) respectively, are constructed. The algorithm for constructing GRN is mentioned in this section. The neural network transfer function i.e., radial basis transfer function calculates a layer's output from its net input.

Algorithmic Steps:

Input: Cluster center matrices, Cw(i, j) or Cf (i, j) or Cr(i, j) of size 46 × 9

Output: GRN outputs, Y Cw(1,i), Y Cf (1,i) and Y Cr(1,i)

Step1: Select any cluster center matrix from the list, Cw(i, j) or Cf (i, j) or Cr(i, j) of size 46 × 9

Step2: The maximum value (M) and minimum value (m) of selected center matrix is found. It will result in a column matrix.

max(clustercentermatrix) = M.

min(clustercentermatrix) = m.

Step3: Spread or width of Radial Basis Transfer Function are calculated from M and m for selected center matrix.

spread = m : steps : M

Step4: The target of network is set at the first time point value.

Step5: The sum of all time point values is being calculated.

$$Sum = \sum_{1}^{9}\left(alltimepointvalues\right)$$

Step6: The maximum (M1) and minimum (m1) values of summation are calculated as sum is a row matrix.

M1 = max(Sum)

m = min(Sum)

Step7: The new spread of the network is set with the results of step6.

newspresd = m1 :steps : M1

Step8: Simulate and train all radial basis neural networks with new spread.

Step9: The above 1 to 8 steps for all the cluster center matrices are repeated.

Step10: Output of GRNs, Y Cw(1, i), Y Cf (1, i) and Y Cr(1, i) are obtained.

Step11: Then plot and save network outputs.

After constructing gene regulatory networks (GRNs), genetic algorithm (GA) based approach is applied to optimize network outputs.

3.3 Optimization of GRN using Genetic Algorithm

For constructing GRNs by using RBN approach, Gaussian function is considered as radial basis function and output matrices of size 1 × i is obtained as GRN output for each cluster center matrix of microarray datasets. To get polynomial equation for each network outputs, viz., Y Cw(1, i), Y Cf (1, i) and Y Cr(1, i) curve fitting method is applied. For this, linear least squares method is done and robustness of this method is Least Absolute Residuals (LAR).

Then 9th degree polynomial equations of Y Cw(1, i), Y Cf (1, i) and Y Cr(1, i) are optimized by considering them as objective functions for optimization. The number of variables in each equation is 1.

To optimize the objective functions, population type as double vector is considered and the population size is taken as 20. The initial score is zero and initial range is in [0, 1]. In this optimization process, the scaling function is a rank function and selection function is stochastically uniform [10,11]. The elite count, 2 and crossover fraction value, 0.8 are taken for reproduction of new individual with recombination through crossover. The crossover function, by nature is a scatter function. With these selected parameter values the best individual values are obtained [12-14].

For, wild type HIV-1(WT) cell line microarray time series data cluster center matrix genetic network

Table 1: Cluster center values of the Wild Type HIV-1 cell line datasets

Cls No.	t1	t2	t3	t4	t5	t6	t7	t8	t9
Cls1	-6.1	-8.8	-5.7	-9.9	-6.1	-5.5	-3.3	-4.9	-7.3
Cls10	-8.3	-9.9	-9.1	-12.9	-7.4	-8.0	-4.2	-8.5	-9.5
Cls20	-7.2	-13.4	-7.2	-13.4	-9.0	-15.6	-4.2	-5.4	-9.2
Cls30	-2.1	-3.2	-2.3	-4.1	-1.8	-6.1	-0.5	1.0	1.7
Cls40	0.0	2.1	0.1	-3.1	-1.1	-5.7	-1.9	0.6	-3.1
Cls46	-11.6	-26.1	-13.4	-24.7	-16.7	-17.3	-8.8	-14.9	-19.1

Table 2: Cluster center values of the HIV-1 Vpr mutant (F72A/R73A) datasets

Cls No.	t1	t2	t3	t4	t5	t6	t7	t8	t9
Cls1	-6.1	-8.8	-5.7	-9.9	-6.1	-5.5	-3.3	-4.9	-7.3
Cls10	-8.3	-9.9	-9.1	-12.9	-7.4	-8.0	-4.2	-8.5	-9.5
Cls20	-7.2	-13.4	-7.2	-13.4	-9.0	-15.6	-4.2	-5.4	-9.2
Cls30	-2.1	-3.2	-2.3	-4.1	-1.8	-6.1	-0.5	1.0	1.7
Cls40	0.0	2.1	0.1	-3.1	-1.1	-5.7	-1.9	0.6	-3.1
Cls46	-11.6	-26.1	-13.4	-24.7	-16.7	-17.3	-8.8	-14.9	-19.1

Table 3: Cluster center values of the HIV-1 Vpr mutant (R80A)datasets

Cls No.	t1	t2	t3	t4	t5	t6	t7	t8	t9
Cls1	-6.1	-8.8	-5.7	-9.9	-6.1	-5.5	-3.3	-4.9	-7.3
Cls10	-8.3	-9.9	-9.1	-12.9	-7.4	-8.0	-4.2	-8.5	-9.5
Cls20	-7.2	-13.4	-7.2	-13.4	-9.0	-15.6	-4.2	-5.4	-9.2
Cls30	-2.1	-3.2	-2.3	-4.1	-1.8	-6.1	-0.5	1.0	1.7
Cls40	0.0	2.1	0.1	-3.1	-1.1	-5.7	-1.9	0.6	-3.1
Cls46	-11.6	-26.1	-13.4	-24.7	-16.7	-17.3	-8.8	-14.9	-19.1

output, Y Cw(1, i), the best individual value is 0.605 and optimized value of network output is 2.97e + 08.

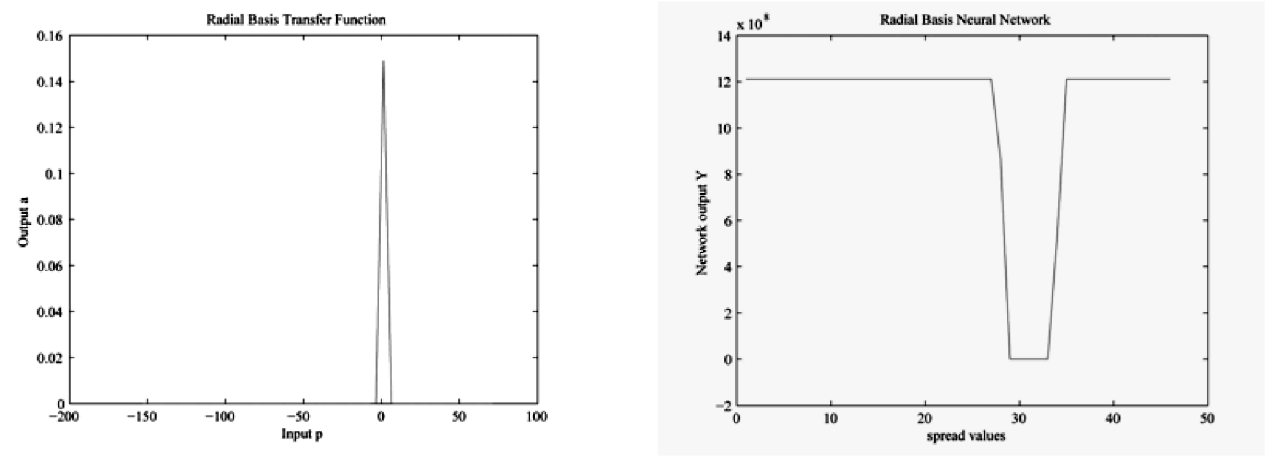
For HIV-1 Vpr mutant (F72A/R73A) microarray time series data cluster center matrix geneticnetwork output, Y Cf (1, i), the best individual value is “20.4800 and network output optimized value is “6.8157e+ 12.

For HIV-1Vprmutant (R80A)microarray time series data cluster center matrix genetic networkoutput, Y Cr(1, i), the best individual value is 34.3950 and network output optimized value is “2.6900e+ 18.

4 RESULTS

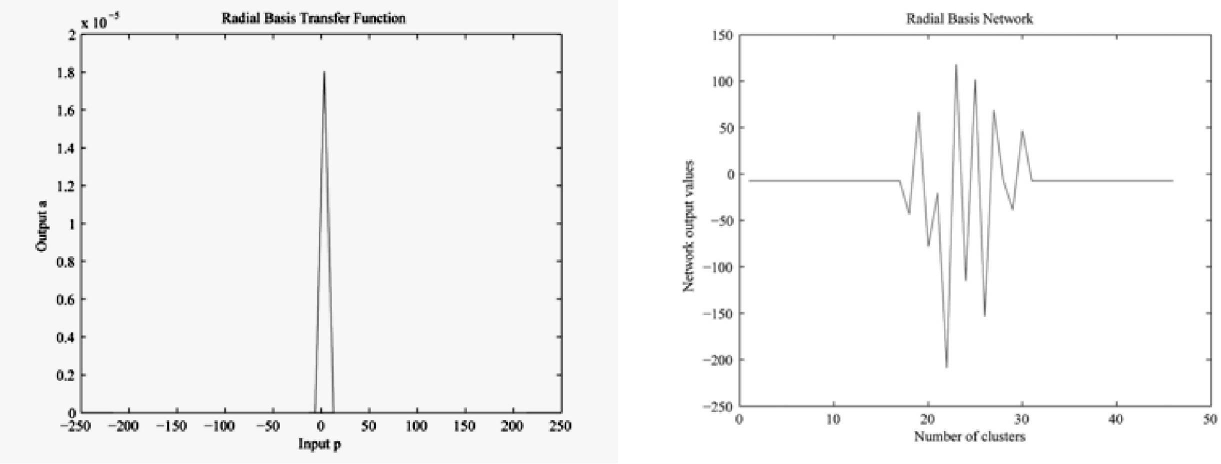
The algorithm is coded and implemented with MATLAB(R2014a). The cluster center matrices of size 46 × 9 of the three sample microarray time series data matrices, W(i, j), F(i, j) and R(i, j) are shown in Tables 1, 2, 3 respectively. In the three tables, only five cluster center values of each cluster center matrices are listed.

After finding radial basis transfer function (RBF) from each cluster center matrix, GRNs are constructed. The graphical plots of all Radial Basis Functions and Gene Regulatory Network outputs, YCw(1, i), YCf(1, i) and YCr(1, i) are shown in Fig.1 through Fig.3 respectively.



(a) (b)

Fig. 1: (a) Plot of Input Vs. Radial Basis transfer Function of wild type HIV-1; (b) Plot of Input Vs. Radial Basis Neural Network of wild type HIV-1



(a) (b)

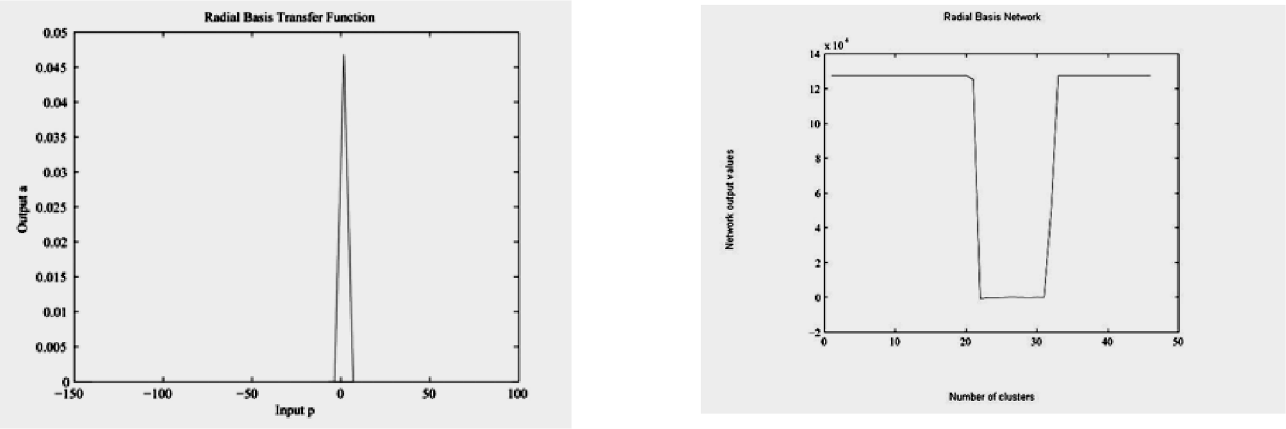
Fig. 2: (a) Plot of Input Vs Radial Basis Transfer Function of HIV-1 mutant F72A/R73A; (b) Plot of Input Vs. Radial Basis Neural Network output of sample, HIV-1 mutant F72A/R73A

The value of each GRN output is a column matrices of size 1 × 46. To get the best fitted value from output matrix, optimization of objective function of each network output is performed. For this, 9th degree polynomial equation is constructed over matrix elements present in output matrix by using curve fitting tools of Matlab (R2014a) software.

After this, the fitting equations are developed as follows (Equation 4, 5, 6):

For Wild type HIV-1, the network output (Y Cw(1, i)) fitting equation,

$$fw(x) = (5.797e + 07)\,x^{9} + (4.037e + 08)x^{8} + (-1.58e + 08)x^{7} + (-2.394e + 09)x^{6} + (-4.872e + 08)x^{5} + (4.475e + 09)x^{4} + (1.81e + 09)x^{3} + (-2.581e + 09)x^{2} + (-1.319e + 09)x + (1.192e + 09) \tag{4}$$



(a) (b)

Fig. 3: (a) Plot of Input Vs Radial Basis Transfer Function of HIV-1 mutant R80A; (b) Plot of Input Vs. Radial Basis Neural Network output of sample, HIV-1 mutant R80A

For HIV-1 Vpr mutant F72A/R73A, the network output (Y Cf (1, i)) fitting equation,

$$ff(x) = (10.32)\,x^{9} + (-12.5)\,x^{8} + (-70.92)\,x^{7} + (77.12)\,x^{6} + (169)\,x^{5} + (-157.2)\,x^{4} + (-161.1)\,x^{3} + (117)\,x^{2} + (49.65)\,x + (-30.32) \qquad (5)$$

For HIV-1 Vpr mutant R80A, the network output (Y Cr(1, i)) fitting equation,

$$fr(x) = (-3.976e + 04)\,x^{9} + (-1.419e + 04)\,x^{8} + (2.854e + 05)\,x^{7} + (1.036e + 05)\,x^{6} + (-7.218e + 05)\,x^{5} + (-2.683e + 05)\,x^{4} + (7.495e + 05)\,x^{3} + (2.883e + 05)\,x^{2} + (-2.633e + 05)\,x + (2.117e + 04) \qquad (6)$$

After optimization of objective functions mentioned in Equation 4, 5, and 6, the optimized values of objective functions are reported in Table 4.

In genetic algorithm (GA), fitness value for optimization of each individual is determined. The function selects individuals from population to reproduce new chromosomes. Crossover takes place within two selected chromosomes and they split. After that, two new individuals mutate. This process is repeated for a certain number of times in a iterative way. In this work, 54 iterations are done. At last, the best individual value and optimized value of network output are obtained. The optimized value of network outputs for cell line express wild type HIV-1 Vpr data matrix, HIV-1 Vpr mutant, F72A/R73A and HIV-1 Vpr mutant R80A are 2.97e+08, -6.8157e+12, and -2.69e+18 respectively. It means that nature of curves for wild type is positively increasing and mutant Vprs are also increasing but mirror function in nature with that of WT.

From Table 4, it is noticed that the best individuals for objective function for Vpr induced microarray data (F72A/R73A) network output, is the lowest among the three microarrays. The mean value and the best fitness value for Vpr induced expression data (R80) is the lowest whereas its current best individual value is the highest. For, wild type expression microarray data, the mean value and the best fitness value is the highest. These are summarized in Table 5.

Table 4: Optimized values of objective functions

Objective function	Current best individuals	Mean value	Best fitness value
fw(x)	0.605	4.115e+08	2.97e+08
ff (x)	-20.48	-6.78e+12	-6.8157e+12
fr(x)	34.3950	-2.68e+18	-2.69e+18

Table 5: Comparison within optimized values of objective functions of network outputs Network

Obj. func.	Best individuals	Mean values	Best fitness values
fw(x)	high	highest	highest
ff (x)	lowest	low	low
fr(x)	highest	lowest	lowest

5. CONCLUSION

Computational analysis of gene expression microarray time series data is very important and an efficient research tool in assisting experimental efforts to analyze microarray data. This paper describes algorithm to cluster data matrices, construction of GA-optimized gene regulatory network by using radial basis neural network approach. The algorithm is coded and implemented on human HIV-1Vpr gene expressed microarray data samples. The predictions will inspect with biological datasets to identify optimum value of Radial Basis Network outputs. From the optimum value the most effective gene from data sample will be found in future. These may considered as the most significant gene because those genes would take the major role for disease progression. By identifying the features of those genes and by considering their amino acid sequences one may identify drug targets which may lead to drug design in future.

REFERENCES

[1] Zhao, R.Y.,Bukrinsky, M. and Elder, R.T., HIV-1 viral protein R (Vpr) and host cellular responses, The Indian Journal of Medical Research, Vol. 121, No.4, pp. 270-286, 2005.

[2] Yao, X.J.,Rougeau, N. and Duisit, G., Analysis of HIV-1 Vpr determinantsresponsible for cell growth arrest in Saccharomyces cerevisiae, Retrovirology, Vol. 1, No.21, 2004.

[3] Kogan, M. and Rappaport, J., HIV-1Accessory ProteinVpr: Relevance in the pathogenesisof HIV and potential for therapeutic intervention, Retrovirology, Vol. 8, No.25, pp. 25-44, 2011.

[4] Sarafianos, S.G., Das, K. and Tantillo, C., Crystal structure of HIV-1 reversetranscriptase in complex with a polypurine tract RNA:DNA, The EMBO Journal, Vol.20, No.6, pp.1449-1461, 2001.

[5] Doan , C.D.,Liong, S.Y. and Karunasinghe, D. S.K., Derivation of effective and efficientdata set with subtractive clustering method and genetic algorithm, Journal of Hydroinformatics,Vol. 7, No.4, pp. 219-233, 2005.

[6] Bataineh, K.M., Najia, M. and Saqera, M., A Comparison Study between VariousFuzzy Clustering Algorithms', Jordan Journal ofMechanical and Industrial Engineering, Vol. 5(4), pp. 335-343, 2011.

[7] Priyono, A.,Ridwan, M. and Alias, A. J., Generation of Fuzzy Rules with Subtractive Clustering, Journal Teknologi, Vol. 43, pp. 143-153, 2007.

[8] Goyal, S. and Goyal, G.K., Radial Basis (Exact Fit) ArtificialNeural Network Techniquefor Estimating Shelf Life of Burfi, Advances in Computer Science and its Applications, Vol. 1, No.2, pp.93-96, 2012.

[9] Orr, M.J.L., Introduction to radial basis function networks, Technical Report,Centre for Cognitive Science, University of Edinburgh, Scotland, 1996.

[10] Fiszelew, A., Britos, P. and Ochoa, A., Finding Optimal Neural Network ArchitectureUsing Genetic Algorithms', Advances in Computer Science and Engineering Research in Computing Science,Vol. 27, pp. 15-24, 2007.

[11] Zhang, B.T. and Muhlenbein, H., Evolving Optimal Neural Networks Using

GeneticAlgorithms with Occams' Razor', Complex Systems, Vol. 7, No.3, pp. 199-220, 1993.

[12] Sug, H., Generating Better Radial Basis Function Network for Large Data Set of Census,International Journal of Software Engineering and Its Applications, Vol. 4 (2), pp. 15-22, 2010.

[13] Barman, B. and Mukhopadhyay, A., Construction of GA-optimized Radial Basis Neural Network from HIV-1 Vpr Mutant Microarray Gene Expression Data, Proceedings of International Conference on Computational Intelligence: Modeling, Techniques and Applications (CIMTA 2013),Kalyani, India, Procedia Technology, Vol. 10, pp. 450-456, 2013.

[14] Barman, B.,Biswas, P. and Mukhopadhyay, A., Comparison of gene regulatory networks usingadaptive neural network and self-organising map approach over Huh7 hepatoma cell microarray datamatrix, International Journal of Bio-Inspired Computation, Vol. 8, No.4, pp. 240-247, 2016.