International Conference on Computational Intelligence: Modeling, Techniques and Applications (CIMTA) 2013

Construction of GA-optimized Radial Basis Neural Network from HIV-1 Vpr Mutant Microarray Gene Expression Data

Bandana Barman\textsuperscript{a,}\textsuperscript{*}, Anirban Mukhopadhyay\textsuperscript{b}

\textsuperscript{a}Dept. of Electronics and Commun. Engg., Kalyani Govt. Engg. College, Kalyani - 741235, India
\textsuperscript{b}Dept. of Computer Sc. and Engg., University of Kalyani, Kalyani - 741235, India

Abstract

Finding the most significant gene from microarray time series data is important for designing drugs of particular disease. Construction of Neural Network through protein interactions is a vital and useful approach to develop new drugs target. Some of the computational tools are being utilized for predicting the viral-host interactions. The database of human HIV-1 Vpr mutant gene expression microarray time series expression value contain records of experimentally validated interactions. The main problem to analyze this type of microarray data is classification problem as because human HIV-1 Vpr mutant cell is an infected dendritic cell. We firstly, have clustered the gene microarray time series data using subtractive clustering method then construct Radial Basis Neural Network on cluster of HIV-1 Vpr mutant microarray time series data. The network output is optimized by using Genetic Algorithm and from the optimized value of network output we got a significant gene which lead to drug discovery in future.

© 2013 The Authors. Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
Selection and peer-review under responsibility of the University of Kalyani, Department of Computer Science & Engineering

Keywords: Gene expression time series microarray data; Subtractive Clustering; Gene Regulatory Network; Radial Basis neural network; Genetic Algorithm

1. Introduction

Human immunodeficiency virus (HIV) can lead to acquired immunodeficiency syndrome (AIDS), a condition in humans which may destroy the immunity system leading to life-threatening infections [1]. HIV, an RNA-lentivirus, is a member of retrovirus family whose hallmark is the reverse transcription of its genomic RNA to DNA by the enzyme reverse transcriptase. Replication cycle of HIV begins with gp120 protein via a portion of its V1 region. Activation of HIV expression from the latent state depends on the interaction of a number of cellular and viral factors. Following transcription, HIV mRNA is translated into proteins. HIV-1 has genes that encode the structural proteins of the virus. However HIV-1 is more complex than other retroviruses [2]. Firstly HIV virus binds with the dendritic cells of host which play an important role in the initiation of the viral infection. Likewise, each point in the replication cycle of HIV is a real or potential target for therapeutic intervention.

\textsuperscript{*} Corresponding author. Tel.: +91-33-2582-1309
\textit{E-mail address:} bandanabarman@yahoo.com
Some computational approaches have been developed earlier to analyze the microarray gene expression data of HIV-1 host responses and also to develop the gene regulatory network within the data. The scientists did research to find target of molecular interactions between the virus and its host of HIV-1 [3]. The host cell growth responses with HIV-1 viral protein Vpr [4] is analyzed and the analysis of HIV-1 Vpr is done in Saccharomyces cerevisiae (yeast) cell [5]. The relevance and potential of viral protein Vpr of HIV-1 for therapeutic intervention has been studied [6]. Some The approaches have also studied the regulation of the HIV-1 promoter by Vpr proteins [7].

HIV-1 is one of the species of HIV virus that relies on human host cell proteins in virtually all phases of its life cycles. As HIV diverges from founder to chronically replicating virus, it accumulates N-linked glycosylation sites. We have collected a publicly available HIV-1 Vpr mutant gene microarray time series data. The data is clustered to obtain co-expressed genes using subtractive clustering approach as the computation in this approach is proportional to the problem size instead of the problem dimension. As actual cluster centers are not necessarily located at only one of the data points. However in most of the cases it can be considered as a good approximation, specially with the reduced computation time of the approach [8].

Construction of Gene Regulatory Network (GRN) or Genetic Network with the concept of Neural Network within Gene microarray time series data is an area of research field now a days. Previously, many approaches were applied to develop GRN, like GRN with Sparse network [9], Bayesian Network, Graphical Gaussian Network, Feed Forward Network and Feed Forward Back Propagation Network. Pairwise gene-gene interaction was developed to construct genetic network. As we know only proteins participate to make interactions, protein-DNA interaction are also generated to develop GRN.

We constructed genetic network within the cluster centers, we got after subtractive clustering of our HIV-1 Vpr mutant gene microarray time series data. Our approach is Radial Basis Function Neural Network (RBN) approach [10] which are two-layer feed-forward networks. The hidden nodes implement a set of radial basis functions (e.g. Gaussian functions). The output nodes implement linear summation functions as in an MLP. The network training is divided into two stages: first the weights from the input to hidden layer are determined, and then the weights from the hidden to output layer are obtained. The training learning is very fast and the networks are very good at interpolation.

We optimize the neural network output fitness function constructed with the centers of clusters of microarray data using genetic algorithm (GA). GAs are good at exploring a large and complex space in an intelligent way to find near-global-optimal values [11]. There are so many research field areas where GAs are applied for optimization like Image Processing (Image Gesture Enhancement), Cluster Analysis (Fuzzy Clustering, Subtractive Clustering, Feed Forward Neural Network optimization [12] etc. In this article, the GA is developed with its three main steps selection, crossover, and mutation. To find the most effective gene from the microarray time points data, we find the minimum Euclidean Distance between the Radial Basis Neural Network optimized value and HIV-1 microarray data.

2. Materials

We collected the microarray data sets of inducible HIV-1 Vpr protein on cellular gene expression. Data set is publicly available at http://www.ncbi.nlm.nih.gov/geo/data/. The cell lines express HIV-1 Vpr mutant F72A/R73A. Data cells were collected at 0, 1, 2, 4, 6, 8, 12, 16 and 24 hours post induction. Total number of RNA presents in the cell are 21,794. So our microarray data matrix consists of 21,794 rows and 9 columns corresponding to the time points values. We first normalize the dataset so that each gene has mean 0 and variance 1.

3. Methods

In this section, we first discuss the subtractive clustering algorithm which is applied on HIV microarray data. Then we describe the algorithm to find Gene Regulatory Network based on Radial Basis Function Neural Network approach to find the objective function of the genetic network output. We optimize the output objective function using genetic algorithm. Then to find the most efficient gene from the microarray data we calculate the minimum Euclidean distance with optimized value and microarray data. All of these algorithms are described in brief in this section. We implemented our algorithms in Matlab (R2010a).
We know that the rate of proteolysis is inversely proportional to the amount of substrate and it is represented by ordinary differential equation (ODE) over space and time,

\[ \frac{dG}{dt} = -K_p G(t), \]  

(1)

where, \( G, \frac{dG}{dt} \) are state variable and rate of change of state variable respectively. \( K_p \) is fixed parameter and \( G(t) \) is the change of the state variable.

Gene expression values in a microarray data change only with time. ODE for genes present in the data matrix will be represented by partial differential equation (PDE) over time:

\[ \frac{dG_i}{dt} = -K_p G_i(t), \]  

(2)

where, \( G_i \)s are the genes present in the microarray data matrix. At any time the genes can be represented by integrating the PDE.

3.1. Subtractive Clustering of HIV-1 Expression Data

The objective of subtractive clustering method (SCM) is to identify fuzzy clustering models. The main differences between fuzzy c-means (FCM) and SCM are estimating potential values and the influence of a neighboring data point. The procedure to obtain a new cluster center and to revise potentials was amended to ease the difficulty in establishing a very sensitive parameter. The SCM starts assuming each data point as a potential cluster center. The points having more neighboring data points will have a higher opportunity to become a cluster center. The potential value for each data point is computed based on the density of surrounding data points. Data points that are outside a range has little influence on potential.

After computing potential of each data point, the data point with the highest potential is chosen as the first cluster center and the data points near the first cluster center will have greatly reduced potential. This process continues until no further cluster center is identified. Data points are chosen as new cluster centers based on two parameters: (1) Accept Ratio (AR) (2) Reject Ratio (RR). Hence the influence range and squash factor along with and RR, set the four criteria for identification of cluster centers in SCM.

Let the microarray matrix to be clustered is denoted by \( X(i, j) \)xc. Each row of \( X(i, j) \) denotes a gene and each column denotes a time point value of the corresponding gene. The ‘radii’ is a vector variable of entries between 0 and 1 that specifies a cluster center’s range of influence in each of the data dimensions.

Size of the data matrix is \( 21794 \times 9 \). To apply subtracting clustering, we select value of influence range i.e. ‘radii’ as 0.05 on this data as the data is of HIV-infected dendritic cells. It means genes are very closely expressed. We get 46 clusters from our data and get a cluster center matrix, \( S(k, j) \) of size \( 46 \times 9 \).

3.2. GRN with Radial Basis Network within cluster centers

Radial basis network (RBN) is mainly used in time series prediction, and control. Our data is a time series data of 9 time point values. We constructed RBN on cluster centers. The concepts behind RBN are described as follows:

Characteristic feature of Radial Basis Function is that their responses decrease or increase with distance from a central point. The parameters of the model are the center, the distance, and the precise shape. All of these parameters are fixed if it is linear.

A typical Radial Basis Function is Gaussian in nature and it is defined as:

\[ H(x) = e^{-\gamma x^2}, \]  

(3)

where, \( c \) is the center and \( r \) is the width (spread).

The following is the algorithm for construction of RBN with cluster center matrix. Transfer functions calculate a layer’s output from its net input.

**Algorithmic Steps:**

- **Algorithmic Steps:**
Step 1: Input: Cluster center matrix $S(k, j)$ of size $46 \times 9$

Step 2: Find the maximum and minimum value of each time point column of matrix $S$.

Step 3: Find spread or width of the Radial Basis Transfer Function with maximum and minimum value.

Step 4: First set the first time points value as the target of network.

Step 5: Find the sum of all time point values.

Step 6: Find maximum and minimum value of summation.

Step 7: Now set the spread of the network with the values of step 6.

Step 8: Simulate and train the network with new spread.

Step 9: Output: The network, $Y$ of size $1 \times 46$.

Step 10: Plot and Save the output of the network.

After developing the genetic network we use genetic algorithm based approach to optimize the network.

3.3. Optimization of GRN with Genetic Algorithm

After constructing genetic network we got output matrix of size $1 \times 9$. We fit the 9th degree polynomial equation of $Y$ using curve fitting tool. Here we consider fitting method as linear least squares and robustness is LAR.

We optimized this equation considering it as objective function. Here number of variables is 1. The genetic algorithm [12] solver could solve unconstrained optimization problems before, but now it has ability to solve general nonlinear optimization problems [13] with linear constraints, bounds, and nonlinear constraints. We optimize the objective function using optimization tool. We take the chromosome type a real-valued vector. The population size is fixed at 20. The initial population and score both are zero and initial range is $[0, 1]$. Selection function is stochastically uniform. For reproduction of new individual with recombination through crossover, we select elite count as 2 and crossover fraction as 0.8. The crossover function is scatter in nature. By selecting the parameters as above we found the best individual value $-24.6$. Now we put this individual value into the objective function and then we get the optimized value of radial basis neural network.

3.4. Identification of Most Significant Gene

We tried to find the most significant gene from human HIV-1 Vpr mutant gene microarray time series data. To get that, the minimum Euclidean distance $d$ is calculated between optimized value of RBN output i.e. $Y_{(t-1)}$ and human HIV-1 Vpr mutant data. Then we got actual significant gene with expression value that maintained minimum Euclidean distance. The identified gene is significant one because it is responsible to make the genetic network.

4. Results

After coding the algorithm with MATLAB(R2010a) we find the cluster center matrix of size $46 \times 9$ which is showing in Table 1.

Among the 46 center values, we have shown only five cluster centers with nine time point values. After finding the radial basis function and neural network we plot the RBF and RBN output which are shown in Fig. 1.
Table 1. Cluster center values of the HIV-1 Vpr mutant datasets

| Cls no. | t1   | t2   | t3   | t4   | t5   | t6   | t7   | t8   | t9   |
|---------|------|------|------|------|------|------|------|------|------|
| Cls1   | 1.42 | -0.59| -3.79| -2.64| -1.09| -1.92| -2.03| -8.59| -2.92|
| Cls10  | 26.47| 14.23| 3.75 | 16.72| 12.38| 4.19 | 0.26 | 14.94| 6.64 |
| Cls20  | 22.71| 17.47| 7.84 | 23.28| 27.91| 10.09| 9.05 | 13.16| 9.33 |
| Cls30  | -44.97| -10.57| -8.18| -11.95| -19.73| -6.54| -8.30| -45.37| -8.29|
| Cls40  | 29.50| 28.25| 12.23| 40.25| 29.36| 10.86| 6.71 | 22.33| 8.41 |
| Cls46  | 13.41| -5.10| -1.30| -1.50| -1.47| 0.75 | 2.90 | -10.82| 8.56 |

![Fig. 1. (a) Plot of Input Vs. Radial Basis Function; (b) Plot of Input Vs. Radial Basis Neural Network](image)

The network output $Y$ is a matrix of size $1 \times 46$. To optimize the network output values we first find the $9^{th}$ degree polynomial equation from the values by curve fitting over that. The fitting equation is as follows:

$$f(x) = (1.869e-015) * x^9 + (-4.223e-013) * x^8 + (3.989e-011) * x^7 + (-2.036e-009) * x^6 + (6.063e-008) * x^5 + (-1.06e-006) * x^4 + (1.047e-005) * x^3 + (-5.369e-005) * x^2 + (0.0001206) * x + (-7.549)$$

Now we plot the objective function input values and also analyze it for the function (Fig. 2).

![Fig. 2. (a) 9th degree Polynomial fit of RBN; (b) Polynomial fit of Objective function with 95 percent confidence bounds](image)

We describe the techniques to optimize the network output. We plot the generation Vs. fitness value and number of variables Vs. best fitting value. These are shown in Fig. 3.

We got the best value of objective function $-9.26$, mean value $-9.252$ and best individual value is $-24.6$. We then calculate the minimum Euclidean distance between best objective function value and the HIV-1 Vpr mutant gene microarray time series data. We got the minimum distance value $8.9343$ and found the corresponding gene ID reference number $1175727 - 1$. The name of gene having this ID reference number is GTP binding protein 2 (GTPBP2), mRNA.
5. Analysis and Discussion

In this section we analyze the implemented algorithm and the results we got after experiments with human HIV-1 Vpr mutant microarray time series data. Reverse transcriptase (RT) [14] is an enzyme which converts the RNA to DNA and it is unique. Nowadays it has become very important to learn more about reverse transcriptase functions. The occurrence of resistance will help researchers to develop drugs that can inhibit RT. It is known that AIDS is a global pandemic that is caused by HIV-1 and HIV-2. HIV-1 is the major cause of AIDS [2]. At the time of infection of host cells by HIV-1, host-viral interactions [1] take place and these interaction processes are time dependent. These interactions determine the efficiency of viral infection and subsequent disease progression. As our data is a large data to analyze, we first cluster it so that the coexpressed genes are grouped. The viral offensive strategies has also developed tactics to suppress host cellular responses where viral protein R (Vpr) plays a particularly active role [4]. We clustered the data using subtractive algorithm because this algorithmic process extracts the most representative data from a raw data set. Subtractive Clustering Method (SCM) (a) removes outliers and (b) discard unnecessary or superfluous points. Our HIV-1 data is a raw data and by using SCM we extract the most representative data then we form Radial Basis Network on result.

Radial Basis Function (RBF), has simple layered structure and it does not change its Gaussian function and it is exponential in nature. The data set is time dependent and the network is constructed in an entire time of 24 hours duration. This approach is a choice as hidden unit activation in an RBF uses a distance to a prototype followed by local transformation. RBF uses localized basis functions and typically only a few hidden units have significant activations. The training of this network completes in two stages, the first of which is unsupervised and the second is a linear supervised method. We got the network output in a matrix because all the cluster center values in each time point make interactions. So that the time point-wise distance matrices with their weights are generated. Then to optimize the network output we used Genetic Algorithm approach [13].

The genetic algorithm works in three steps: Selection, Crossover and Mutation. Some function determines fitness of each individual and the other function selects individual from the population to reproduce. Then crossover takes place within two selected chromosomes and two offspring solutions are produced. Next, the two new individuals mutate and the process is then repeated for a certain number of times. This process is an iterative process. In our coding we used 100 iterations (generations). As a result we get the best individual value and optimized output of network.

We consider that the gene expressed value which keeps the minimum distance from the optimized value of network, is strongly participated in the network. So, we found Euclidean Distance within HIV-1 data and RBN optimized value and it results a single gene ID-reference with its time point.
6. Conclusion

Computational analysis is a very important research tool in assisting experimental efforts to identify protein or protein pairs within a single organism. This paper proposes methods to predict a single gene from human HIV-1 Vpr mutant gene expressed data that is potentially responsible for disease progression. Our knowledge of human drug target is integrated in a supervised learning framework. Our prediction tries to identify the most effective gene. By identifying the features with amino acid sequences we may get drug target and this may lead to drug design in future.

References

[1] Zhou, Y., Ratner, L.. Phosphorylation of human immunodeficiency virus type 1 Vpr regulates cell cycle arrest. Journal of Virology 2000;74(14):6520–6527.
[2] Sarafianos, S.G., Marchand, B., Das, K., Himmel, D.M., Parniak, M.A., Hughes, S.H., et al. Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition. Journal of Molecular Biology 2009;385(3):693–713.
[3] Bannwarth, S., Gatignol, A.. HIV-1 TAR RNA: the target of molecular interactions between the virus and its host. Current HIV Research 2005;3(1):61–71.
[4] Zhao, R., Bukrinsky, M., Elder, R.. HIV-1 viral protein R (Vpr) and host cellular responses. The Indian Journal of Medical Research 2005;121(4):270–286.
[5] Yao, X.J., Rougeau, N., Duisit, G., Lemay, J., Cohen, E.A.. Analysis of HIV-1 Vpr determinants responsible for cell growth arrest in Saccharomyces cerevisiae. Retrovirology 2004;1(21).
[6] Kogan, M., Rappaport, J.. HIV-1 accessory protein Vpr: relevance in the pathogenesis of HIV and potential for therapeutic intervention. Retrovirology 2011;8(25):25–44.
[7] Deshmene, S.L., Amini, S., Sen, S., Khalili, K., Sawaya, B.E.. Regulation of the HIV-1 promoter by HIF-1a and Vpr proteins. Virology Journal 2011;8(477).
[8] Priyono, A., Ridwan, M., Ali, A.J., Rahmat, R.O.K., Hassan, A., Ali, M.A.M.. Generation of fuzzy rules with subtractive clustering. Journal Teknologi 2007;43.:143–153.
[9] Sarder, P., Schierding, W., Cobb, J.P., Nehorai, A.. Estimating sparse gene regulatory networks using a bayesian linear regression. IEEE Transactions on Nanobioscience 2010;9(2):121–131.
[10] Orr, M.J.L.. Introduction to radial basis function networks. Tech. Rep.; Centre for Cognitive Science, University of Edinburgh; Scotland; 1996.
[11] Branke, J.. Evolutionary algorithms for neural network design and training. In: Proc. First Nordic Workshop on Genetic Algorithms and its Applications. 1995, p. 145–163.
[12] Fiszewlew, A., Britos, P., Ochoa, A., Merlino, H., Fernandez, E., Garcia-Martinez, R.. Finding optimal neural network architecture using genetic algorithms. Advances in Computer Science and Engineering Research in Computing Science 2007;27:15–24.
[13] Zhang, B.T., Muhlenbein, H.. Evolving optimal neural networks using genetic algorithms with occams’ razor. Complex Systems 1993;7(3):199–220.
[14] Sarafianos, S.G., Das, K., Tantillo, C., Clark, A.D., Ding, J., Whitcomb, J.M., et al. Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract ma:dna. The EMBO Journal 2001;20(6):1449–1461.