IACSIT International Journal of Engineering and Technology, Vol.2, No.3, June 2010
ISSN: 1793-8236

ANN based Genome Classifier using Frequency Chaos Game Representation

Vrinda V. Nair, Karthika Vijayan, Deepa P. Gopinath and
Achuthsankar S. Nair

***Abstract*—Classification of organisms into different categories using their genomic sequences has found its importance in the study of evolutionary characteristics of organisms and specific identification of previously unknown organisms in bio-diversity studies and related areas. Chaos game representation (CGR) uniquely represents DNA sequence in a visual format and reveals hidden patterns in it. Frequency CGR (FCGR) derived from CGR shows the frequency of sub-sequences present in the DNA sequence. In this paper, a novel method for classification of organisms based on a combination of FCGR and Artificial Neural network (ANN) is proposed. Eight categories, from the taxonomical distribution of Eukaryotic organisms are taken from NCBI and ANN is used for classification. Different con- figurations of ANN are tested and good accuracy is obtained. The way, the fractal nature of CGR helps in classification, is also investigated.**

***Index Terms*—Artificial Neural Network, frequency chaos game representation, genome classification, mitochondrial genomes.**

I. Introduction

Classifying organisms into groups within taxonomical hierarchy has got several applications which include specific identification of any unknown organism, study of evolutionary characteristics, study of mutual relationship existing between organisms etc. Identification of a species from its genome or genomic fragment is a challenging task. This could be done by extracting distinct genomic features of the categories chosen for classification. Here we explore a novel combined technique of chaos game representation (CGR) and Artificial Neural Networks (ANN), the former for genomic feature extraction and the latter for the subsequent species classification.

Chaos Game representation (CGR) is proposed as a unique and scale-independent representation of DNA sequences by H. Joel Jeffrey [1]. The CGR visually reveals both local and global patterns hidden in DNA sequence. By dividing CGR into quadrants, sub-quadrants and sub-sub-quadrants, nucleotide, di-nucleotide and tri-nucleotide frequencies can be estimated [2]. The frequency of nucleotide combinations can be determined by dividing the CGR space with a grid of appropriate size and counting the occurrences in each quadrant. These frequency matrices obtained from CGR are termed as FCGR [3]. In this paper, 8X8 FCGR is used for representing each DNA data.

Species identification based on mitochondrial genomes was implemented and various feature descriptor diagrams constructed [6]. Principal Component Analysis was employed to obtain a unique feature descriptor. It was shown that the feature descriptors were effective representatives of the structural signature of the species. Eubacterial and archeal genomes were classified with an accuracy of 85% using a nave Bayesian Classifier based on di-nucleotide composition bias [4]. Protein classification into domains of life was attempted and the test protein was predicted to be of bacterial or eukaryotic origin with 85% accuracy using a Markov model for compositional bias analysis [5].

Possibility of species classification using CGR images of DNA sequences, by using different distance metrics [7] and by using neural networks [8] has been investigated. A detailed classification problem, using a combination of FCGR and Naive Bayesian approach is explained in Vrinda et al. [9].

ANN have been used for a variety of biological problems including protein sub-cellular localization [10] [11], predicting secondary structure of proteins [12] etc.

In this paper, a classification problem considering 8 classes of Eukaryotes is addressed. CGR is used for uniquely representing the genomic sequences. The FCGR, derived from the CGR image, is used as a representation of mitochondrial genome of organisms and forms the feature vector for classification. ANN is used for classification. The database consists of genomic sequences of organisms which are already classified. The ANN is trained and tested using this database. The classification accuracy obtained is

92.3%. The trained ANN is capable of recognizing the class in which an unknown organism belongs, using its genomic sequence. The way the fractal nature of CGR helps in classification, is also investigated.

II. Materials And Method

A. Tools

CGR is a unique pictorial representation, which reveals the hidden patterns in DNA sequences. The CGR images for DNA sequences are plotted using the tool C-GRex, which is developed by Centre for Bioinformatics, University of Kerala, India. The tool is available at www.cbi.keralauniversity.edu. for free. Conversion of CGR image to FCGR and the classification using ANN, were done

Vrinda V. Nair is with Government Engineering College, Thrissur, Kerala and research scholar at Centre for Bioinformatics, University of Kerala, Thiruvananthapuram, Kerala, INDIA. (e-mail:nairvrinda@rediffmail.com), Karthika Vijayan and Deepa P. Gopinath are with College of Engineering, Thiruvananthapuram, Kerala, INDIA (karthikavijayan@gmail.com, deepapgopinath@gmail.com)and Achuthsankar S. Nair is the Hon. Director, Centre for Bioinformatics, University of Kerala, Thiruvananthapuram, Kerala, INDIA (sankar.achuth@gmail.com)

using MATLAB.

B. Test Data

The mitochondrial DNA sequences of 1476 Eukaryotic organisms belonging to 8 categories of taxonomical hierarchy were obtained from National Centre for Biotechnology Information (NCBI) Organelle database (www.ncbi.nlm.nih.gov). Of these 1476 data, 739 are used for training and 737 are used for testing the neural net- work. The eight classes selected are Plant, Fungi, Cnidaria, Porifera, Acoelomata, Pseudocoelomata, Protostomia and Vertebrata. A few other classes are present in the taxonomical hierarchy which are neglected because enough number of organisms' DNA data is not available from those classes.

C. Methodology

The Chaos game representation for a mitochondrial DNA sequence is obtained. The CGR is plotted using C-GRex. CGR of DNA sequences of organisms show distinct patterns thus revealing the fact that the DNA sequences have specific structures (Fig. 1). The CGR obtained is divided into cells using a 2^n x 2^n grid. The number of black dots in each cell is taken as elements of a matrix which is known as the FCGR matrix. Let A be an n^{th} order FCGR matrix. Then A contains 2^n x 2^n elements. Let $a_{i,j}$ $(1 \le i \le 2^n, 1 \le j \le 2^n)$ denote the elements in the matrix. Normalised FCGR matrix is defined as shown in (1).

$$\bar{A} = \frac{A}{\sum_{I=1}^{2^n} \sum_{j=1}^{2^n} a_{I,j}} \quad (1)$$

Let the element in B be denoted by $b_{i,j}$. Then the sum of the elements of matrix is equal to 1 as shown in (2).

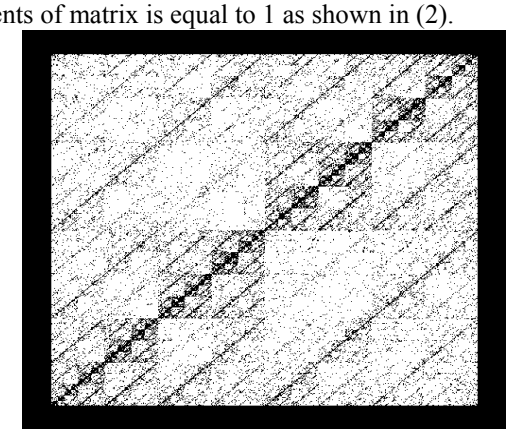

Figure 1: CGR of mitochondrial genome of Saccharomyces cerevisiae (Baker's yeast).

$$\sum_{1}^{2^n} \sum_{1}^{2^n} b_{i,j} = 1 \quad (2)$$

In this study, n is chosen as 3. Subsequently, the black dots in each cell denote the occurrence of a particular tri-nucleotide. Each element in the FCGR matrix, which denotes the tri-nucleotide frequency, is divided by the complete length of the DNA sequence as indicated above, which in turn gives the probability of occurrence of each tri-nucleotide in the genome [13]. Thus the CGR can be viewed as an 8X8 probability matrix, which can be framed as a 64 element vector. The probability set for the baker's yeast is shown in Table I. The same representation is obtained for all the 1476 genomes under consideration.

Artificial Neural networks are used for classification. The total DNA data of 1476 organisms are taken and each one is represented as a 64 element vector. Hence the complete dataset is of size 1476X64. The available DNA data of 1476 organisms are divided in 1:1 ratio to obtain training and testing set for ANN. The distribution of data for all 8 categories into training and testing set is shown in Table II. The training set is of size 739X64 and the size of the test set is 737X64. Using these sets different types of ANN are trained.

To formulate the target set for ANN, the output of the network is assigned to different range of values for each class. Each class is given a label, which is the mid value of the corresponding numerical range assigned to that class. The numerical range and the label assigned to each class is shown in table III. The range indicates that, even though the target value for each class is assigned to a particular label (say 3 for plant), the output obtained in the corresponding range (say 0-6 for plant)will also be classified accordingly. Thus the target matrix consists of the labels which denote the class of the each training vector and it is of size 739X1. A modification of training and test dataset for ANN exploring the fractal nature of CGR, is also attempted.

The fractal nature of CGR discloses the fact that a miniature copy of the whole CGR will be replicated in its single quadrant. So considering single quadrant for classification will be equivalent to classification using complete CGR. Instead of neglecting the other 3 quadrants, all 4 quadrants are taken separately and corresponding elements are averaged.

TABLE I: NORMALIZED AND SCALED FCGR FOR SACCHAROMYCES CEREVISIAE.

6.25	3.10	1.05	3.05	5.15	1.45	5.85	8.50
4.05	6.20	2.85	1.75	2.45	2.30	6.75	4.50
2.85	2.95	3.90	2.45	2.25	5.30	2.15	2.70
7.60	7.30	8.10	9.85	12.10	10.65	8.55	6.90
3.50	2.05	2.75	7.55	5.45	2.70	2.20	7.00
6.65	8.35	11.40	8.30	7.55	7.45	9.30	5.75
6.95	11.45	7.25	9.70	8.50	10.35	9.85	9.40
122.15	72.90	101.95	75.25	70.85	102.70	73.30	48.65

TABLE II: DISTRIBUTION OF AVAILABLE DATA INTO TRAINING AND TESTING SET FOR AN.

Classes	Total	Training	Test
Plant	30	15	15
Fungi	52	26	26
Cnidaria	34	17	17
Porifera	21	11	10
Acoelomata	29	15	14
Pseudocoelomata	30	15	15
Protostomia	256	128	128
Vertebrata	1024	512	512
Total	1476	739	737

TABLE III: THE NUMERICAL RANGE AND LABEL ASSIGNED TO EACH CLASS, AS TARGET.

Class	Range	Label
Plant	0-6	3
Fungi	6-12	9
Cnidaria	12-18	15
Porifera	18-24	21
Acoelomata	24-30	27

Pseudocoelomata	30-36	33
Vertebrata	36-42	39
Protostomia	42-48	45

Thus each genome of 1476 organisms is represented by 16 elements or 4X4 FCGR, which is then converted into a 4X4 probability set. Hence the modified dataset is of size 1476X16. Thus the training and testing set for ANN is redefined and ANN is tested for accuracy of classification.

III. Results And Discussions

Different types of Neural network configurations like Feed forward network with Back propagation training algorithm, Radial basis network, Generalized regression network, Probabilistic neural network etc. are used for classification. The suitability of a network for a particular problem depends upon the nature of the problem.

Back propagation network can be used for problems where the training and target data set are well defined and no discrepancies exist [14]. When large number of training vectors are available, other networks (like Radial Basis Network (RBN), Generalized regression network (GRN), Probabilistic neural network (PNN) etc.) which are capable of producing good results and have much faster training can be used [15]. In the case of back propagation algorithm, the network is created first (number of layers and neurons in each layer are fixed) and then trained using the training set. But other networks are trained while they are created. Each neuron is placed into or removed from the network as the training demands. Spread factor is the only input parameter given to these networks which stipulates the distance between the training vector and the layer weight vector. It is basically a measure of similarity between the weight vector and training vector [15].

The training and testing set, consisting of 64 element vectors representing each genome, are given to the networks. The accuracy of classification using each network is measured using the predefined test dataset. The time taken by ANN to produce the result is also measured, for the comparison between the networks. The processor Intel core2Duo (Intel (R) Core (TM)2Duo CPU @ 2GHz) is used.

A. Feed forward network with Back propagation training algorithm

We tested the feed forward ANN with different back propagation training algorithms, and the results are shown in table IV. Feed forward network with single hidden layer with different types of back propagation learning algorithms, is investigated.

The Levenberg-Marquardt back propagation algorithm gave the best average accuracy of classification. Hence it is chosen for further investigation.

The feed forward ANN with back propagation training algorithm is tested for different number of layers and different number of layer neurons (number of neurons in each layer). Accuracy of classification is noted in each case. Back Propagation training with Levenberg-Marquardt training algorithm and hidden layers with tansigmiod transfer function is used. Results for back propagation ANN with different number of hidden layers and for different number of layer neurons is shown in table V.

Maximum average accuracy obtained from back propagation network with single hidden layer is 55.88%. As the number of layer neurons increases beyond a limit, the average accuracy is seen to reduce.

Back propagation network with 2 hidden layers shows a hike in average accuracy. The maximum accuracy obtained from the network is 73.13%. Here, as network complexity increases (which means as the number of neurons in each layer increases), the accuracy reduces. Thus the back propagation network fails to learn effectively the hidden structures in DNA data.

The results obtained while considering a 3 hidden layer back propagation network is not much different from the above two results. As the network complexity increases, effective learning is hindered. The maximum accuracy produced by the three hidden layer network is slightly lower than that of a two hidden layered network. But training time goes on increasing with network complexity.

TABLE IV: Comparison of performance of different back propagation networks.

Types of networks	Maximum average accuracy (%)
BPN (Levenberg-Marquardt)	55.88
BPN (Batch training)	18.23
BPN (Gradient descent)	15.4
BPN (Gradient descent with momentum)	13.8
BPN (Gradient descent with adaptive)	31.07
BPN (Resilient backpropagation)	38.77
BPN (Fletcher-Powell conjugate gradient)	44.8
BPN (Polak-Ribire conjugate gradient)	40.98
BPN (Powell-Beale conjugate gradient)	42.32
BPN (Scaled conjugate gradient)	37.66
BPN (BFGS quasi-Newton)	43.7
BPN (One step secant)	36.4

TABLE V: Average accuracy of classification obtained from Back Propa- gation network.

No: of hidden layers	No: of neurons in each layer	Average accuracy (%)	Training time (min:sec)
1	20	55.88	0:06
1	40	39.76	0:23
1	60	39.48	1:04
1	80	32.43	1:19
2	[20 10]	64.18	0:13
2	[25 15]	73.13	0:23
2	[30 15]	70.01	0:27
2	[50 25]	55.9	1:15
2	[50 30]	44.09	2:13
3	[20 15 10]	69.2	0:20
3	[25 20 15]	69.3	0:40
3	[30 15 5]	70.15	0:44
3	[30 20 10]	66.3	0:58
3	[30 25 20]	55.8	1:30

B. Radial Basis Network (RBN)

The Radial basis network designed, have two layers with first layer having radial basis function as activation function [14]. The first layer computes the distance between the training vector and weight vector and produces considerable output when the distance between them is low. The second layer boosts the first layers' output to a considerable value.

The spread factor is the measure of similarity between the weight vector and training vector at layer one. It must not be too large which may cause overlapping of two classes. It cannot be too small either, because that may result in less number of right recognitions in a class. Result of radial basis network for different values of spread factor is shown in the table VI. Maximum average accuracy of 72.45% is obtained in a time interval of fraction of a second. Comparing with the back propagation network, the radial basis network produces appreciable accuracy in a very small time interval.

C. Generalized Regression Network (GRN)

Generalized regression network (GRN) is similar to that of RBN. It is a two layer network [14]. The first layer is a radial basis layer as in RBN. But instead of a pure linear layer, the second layer in GRN takes the dot product of layer weights and the first layer output, which in turn is another similarity check.

TABLE VI: AVERAGE ACCURACY OF CLASSIFICATION OBTAINED FROM RADIAL BASIS NETWORK.

Spread factor	Accuracy (%)	Time (sec)
35	69.6	0.4
36	71.2	0.34
38	72.45	0.34
40	71.8	0.414
42	70.69	0.349

TABLE VII: AVERAGE ACCURACY OF CLASSIFICATION OBTAINED FROM GENERALIZED REGRESSION NETWORK.

Spread factor	Accuracy (%)	Time (sec)
3.5	85.07	0.322
4.5	85.7	0.304
5	85.48	0.303
5.5	85.3	0.274
6	84.94	0.318

Thus the GRN ensures better performance than RBN. Result of generalized regression network for different values of spread factor is shown in the table VII. Maximum average accuracy obtained from this network is 85.7%, where that in RBN was 72.45%.

D. Probabilistic Neural Network (PNN)

The Probabilistic neural network (PNN) is another network with similar construction. It has a radial basis layer as first layer, whose output vector shows the similarity of an input vector to a particular training vector [14]. The second layer is a competitive layer, which sums up these contributions for each class to produce a vector of probabilities. And a complete transfer function of this layer will pick the maximum probability and produce a '1' for that class and a '0' for other classes. Result of probabilistic neural network for different values of spread factor is shown in the table VIII. Maximum average accuracy obtained from this network is 86.8%.

E. Comparison of networks

The comparison of performances of different ANNs is shown in table IX. Of all the networks highest accuracy is obtained from the probabilistic neural network (PNN). It requires only fraction of a second to produce the output. Individual class accuracies obtained from PNN are also good. Thus PNN is used for further classification problems. For the comparison between different networks, only 30 data from the class vertebrata was taken. To improve the PNN accuracies further, all the available data from class vertebrata (512) are taken for creating the network. Thus the maximum accuracy obtained is at a spread factor of 1.5 and time taken for this is 1.647seconds. The accuracies obtained for each class is shown in the table X.

TABLE VIII: AVERAGE ACCURACY OF CLASSIFICATION OBTAINED FROM PROBABILIS- TIC NEURAL NETWORK.

Spread factor	Accuracy (%)	Time (sec)
1	85.5	0.305
1.2	86.4	0.269
2	86.8	0.294
3	86.7	0.321
4	86.5	0.553

TABLE IX: COMPARISON OF PERFORMANCE OF DIFFERENT NETWORKS.

Network	Maximum accuracy (%)	Time taken(sec)
BPN	73.13	23
RBN	72.45	0.34
GRN	85.48	0.303
PNN	86.8	0.294

F. Fractal nature of CGR

The fractal nature of CGR can be exploited to reduce the size of feature vectors representing the DNA. Instead of 64 elements, 16 elements are used to represent DNA of organisms. Thus the size of training and test set will be effectively reduced (from 1476X64 to 1476X16). Thus the network complexity of PNN is effectively reduced. Hence the training time for the PNN is also reduced. The results for PNN are shown in the table XI. The following results were obtained with a spread factor of 4.5 and time taken was 0.312 seconds.

Thus the modified training set, which exploits the fractal nature of CGR, reduces the size of training and testing dataset considerably. Also network complexity reduces and hence time taken is lowered. When comparing with the 64 element method, these advantages are obtained without an appreciable reduction in the average accuracy.

The DNA sequences are very long sequences (even up to 3 billion nucleotides/bases). Such long DNA sequences are now effectively represented by vectors of 16 elements, by the exploitation of the fractal nature. By the exploitation of fractal nature, the classification problem is made much simpler. Besides that, a simple and effective representation of long DNA sequence is being found out. Using this representation, classification is done and appreciable accuracy is obtained.

TABLE X: MAXIMUM ACCURACY FOR EACH CLASS PROVIDED BY PNN WITH 64 ELEMENT DATA.

Class	Accuracy obtained (%)
Plant	73.3
Fungi	65.38
Cnidaria	94.12
Porifera	80.0
Acoelomata	85.71
Pseudocoelomata	60.0
Protostomia	86.7
Vertebrata	96.8
Average	92.3

TABLE XI: MAXIMUM ACCURACY FOR EACH CLASS PROVIDED BY PNN WITH 16 ELEMENT DATA.

Class	Accuracy obtained(%)
Plant	66.7
Fungi	57.69
Cnidaria	94.12
Porifera	70.0
Acoelomata	85.71
Pseudocoelomata	60.0
Protostomia	82.81
Vertebrata	95.7
Average	90.1

IV. CONCLUSIONS

Classification of organisms is significant in the study of evolutionary properties of organisms, study of mutual relationships between organisms, specific identification of a previously unknown organism etc. Classification of Eukaryotic organisms into eight categories of taxonomical hierarchy, using the Frequency CGR (FCGR), which is derived from the Chaos game representation (CGR) of their DNA, is discussed in this paper.

The CGR image of mitochondrial DNA sequences was divided using an 8X8 grid, and the number of dots in each cell was counted and entered in an 8X8 matrix (64 elements) to obtain the Frequency-CGR. Thus 64 elements are used to represent a DNA data to formulate test and training set for ANN. Different configurations of ANN were tested and the one which returned highest average accuracy of classification was chosen. The way the fractal nature of CGR helps in classification was also investigated. By exploiting the fractal nature of CGR, 16 elements (4X4 matrix) were used to represent DNA data and thus the test and training set for ANN was redefined.

Different types of artificial neural networks were investigated and the probabilistic neural network (PNN) produced the best classification accuracy (86.8%). Hence that was chosen for further processing. The PNN was improved further by improving the training set and thus an accuracy of 92.3% is obtained.

The fractal property of CGR was used to redefine the training and test set for ANN. 16 element data was used to replace the 64 element data which represents the DNA of organisms. The probabilistic neural network was tested and it returned an accuracy of 90.1%. Comparing with the training set with 64 element vectors, the network complexity and training time reduced a lot.

A novel method for classification of organisms into different categories, using a combination of FCGR and ANN is proposed. Comparing with the previous works in classification, a detailed classification problem is addressed and much higher accuracy of 92.3% is obtained. In comparison with the only previous work reported in the literature using FCGR by Vrinda et. al., two more classes are successfully taken into account for classification and higher average accuracy is obtained. Exploiting the fractal nature of CGR, classification is made much simpler, without an appreciable reduction in accuracy (from 92.3% to 90.1%).

REFERENCES

[1] H. Joel Jeffrey, "Chaos game representation of gene structure", Nucleic Acids Research, Vol. 18, No. 8, pp: 2163-2170,1990.
[2] Nick Goldman, "Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game repre- sentations of DNA sequences", Nucleic Acids Research, 1993, Vol. 21, No. 10, pp: 2487-2491, 1993.
[3] Almeida J S, Joao A. Carrico, Antonio Maretzek, Peter A.Noble and Madilyn Fletcher, "Analysis of genomic sequences by Chaos Game Representation", BIOINFORMATICS Vol. 17, no. 5, Pages 429-437, 2001.
[4] Sandberg R., Winberg G., Branden C. I., Kaske A., Ernberg I. and Coster, "Capturing Whole - Genome characteristics in short sequences using a naive Bayesian classifer", Genome Res., Vol. 11, pp. 1404-09, May 2001.
[5] Zanoguera and de Francesco, "Protein classification into do- mains of life using Markov Chain Models", CSB 2004, The Computational Systems Bioinformatics Conference, IEEE, pp. 517-519, 16-19 Aug. 2004.
[6] Narasimhan S., Sen S. and Konar, "Species identification based on mitochondrial genomes", ICCR 2005, International Conference of Cognition and Recognition, Mysore, India, 22- 23 Dec. 2005.
[7] Deschavanne P., Giron A., Vilain J., Fagot G. and Fertil, "Ge- nomic Signature: Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences", Biol. Evol. 16(10):1391-1399. 1999.
[8] Giron A., Vilain J., Serruys C., Brahmi D., Deschavanne P., and Fertil, "Analysis of parametric images derived from genomic sequences using neural network based approaches", IJCNN'99, International joint conference on neural networks, pp.3604- 3608, 10-16 Jul. 1999.
[9] Vrinda V Nair, Lissy P. Anto and Achuthsankar S. Nair, "Naive Bayesian Classification of unknown sequence fragments based on chaos game representation of mitochondrial genomes", Communications of SIWN, vol 7, pp:27-33, May 2009.
[10] Q. Cui, T. Jiang, B. Liu, and S. Ma, "Esub8: A novel tool to predict protein subcellular localizations in eukaryotic organisms", BMC Bioinformatics, vol 5, 66 May 2004.
[11] G. Schneider, and U. Fechner,"Advances in the prediction of protein targeting signals", Proteomics, vol 4, Apr. 2004, pp.1571-1580.
[12] A. Malekpour, S. Naghizadeh, H. Pezesh, M. Sadeghi, and C. Eslahchi,"Protein secondary structure prediction using three neural networks and a segmental semi Markov model", Math. Biosciences, vol 217, Feb. 2009, pp. 145-150.
[13] W. Fu, Y. Wang, and D. Lu, " Multi-fractal Analysis of Ge- nomic Sequences CGR Images", Proceedings of the 2005 IEE Engineering in Medicine and Biology 27th Annual conference, September 1-4, 2005.
[14] www.statsoft.com
[15] www.mathworks.com