Texture features from Chaos Game Representation Images of Genomes

Vrinda V. Nair *vrinda66nair@gmail.com*
Professor in ECE
College of Engineering Trivandrum
Thiruvananthapuram, Kerala, India

Nisha N. S. *nish.phy@gmail.com*
B.Tech student, College of Engineering Trivandrum
Thiruvananthapuram, Kerala, India

Vidya S. *vidyavasavan@gmail.com*
B.Tech student, College of Engineering Trivandrum
Thiruvananthapuram, Kerala, India

Y. S. Thushana *ysthushana14@gmail.com*
B.Tech student ,College of Engineering Trivandrum
Thiruvananthapuram, Kerala, India

Abstract

The proposed work investigates the effectiveness of coarse measures of the Chaos Game Representation (CGR) images in differentiating genomes of various organisms. Major work in this area is seen to focus on feature extraction using Frequency Chaos Game Representation (FCGR) matrices. Although it is biologically significant, FCGR matrix has an inherent error which is associated with the insufficient computing as well as the screen resolutions. Hence the CGR image is converted to a texture image and corresponding feature vectors extracted. Features such as the texture properties and the subsequent wavelet coefficients of the texture image are used. Our work suggests that texture features characterize genomes well further; their wavelet coefficients yield better distinguishing capabilities.

Keywords: Chaos Game Representation, Texture analysis, Wavelet decomposition, Support Vector Machines.

1. INTRODUCTION

Chaos Game Representation (CGR) has been used in genomics and proteomics for various applications. Major work focuses on using Frequency Chaos Game Representation (FCGR) matrix for analysis. FCGR values though biologically relevant, has an inherent drawback due to insufficient screen as well as computing resolutions. In this work, attempt is made to investigate the effectiveness of coarse features such as texture properties and wavelet decomposition matrix of the texture image obtained from the gray scale equivalent of the corresponding FCGR. The results show that the coarse features of CGR extracted from the image in the form of texture as well as their wavelet coefficients can characterize genomes effectively.

Chaos game representation (CGR), the method used in this paper, for feature extraction, constructs a 2D image of the sequence data, which offers a visual understanding of the structure of the sequence. The differences between the various categories of sequences are evident from their respective CGR images. The CGR can be mapped into a numeric matrix by obtaining a Frequency CGR (FCGR) [1], [2]. A combined technique for genome classification using one probabilistic technique and two machine learning techniques based on FCGR features was

reported [3]. Hurst CGR a method using Hurst exponent to extract features from CGR is presented in [4].

This work maps the CGR image into its texture equivalent and corresponding properties are taken as features. The wavelet decomposition matrix of the texture image is also used as a feature vector for discriminating genome sequences.

2. MATERIALS AND METHODS

2.1 Dataset

Mitochondrial genomes are considered here. They are the sites of aerobic respiration, and are the major energy production center in eukaryotes. The low mutation rate in metazoan mitochondrial genome sequence makes these genomes useful for scientists assessing genetic relationships of individuals or groups within a species and for the study of evolutionary relationships [5]. Mitochondrial genomes were downloaded from the NCBI Organelle database [5]. Table 1 shows the data used for classification. The number of organisms shown is as listed in NCBI on 01/12/2012

Table 1: Dataset Used for Classification.

Serial number	Name of category	Number of organisms
1	Acoelomata	39
2	Cnidaria	48
3	Fungi	102
4	Plant	63
5	Porifera	44
6	Protostomia	582
7	Pseudocoelomata	63
8	Vertebrata	1729
	Total	2670

2.2 Methodology

2.2.1 Chaos Game Representation. The scope of CGRs as useful signature images of bio-sequences such as DNA has been investigated since early 1990s. CGR of genome sequences was first proposed by H. Joel Jeffrey [6]. To derive a chaos game representation of a genome, a square is first drawn to any desired scale and corners marked A, T, G and C. The first point is plotted halfway between the center of the square and the corner corresponding to the first nucleotide of the sequence, and successive points are plotted halfway between the previous point, and the corner corresponding to the base of each successive nucleotide. Mathematically, co-ordinates of the successive points in the chaos game representation of a DNA sequence is described by an iterated function system defined in Eq. 1 and Eq. 2

$$X_i = 0.5\left(X_{i-1} + g_{ix}\right) \quad (1)$$

$$Y_i = 0.5\left(Y_{i-1} + g_{iy}\right) \quad (2)$$

g_{ix} and g_{iy} are the X and Y co-ordinates respectively of the corners corresponding to the nucleotide at position i in the sequence [7]. The CGR of a random sequence gives a uniformly filled square. The CGR of DNA sequences plotted for various species gives images illustrating the non-randomness of genome sequences, which indeed means that the sequence has a structure, indirectly captured by the signature image. Features of CGRs include marked double scoops, diagonals, varying vertical intensities, absence of diagonals etc. signifying corresponding sequence characteristics. The CGR is thus found to be unique for every species. Hence CGR of

genomic sequences are expected to furnish features of discriminative nature which could subsequently be presented to classifiers.

2.2.2 Texture Analysis. Texture analysis refers to the branch of imaging science that is concerned with the description of characteristic image properties by textural features. Texture analysis provides unique information on the texture, or spatial variation of pixel [8]. In texture analysis, a pixel occurrence probability matrix and a gray co-matrix, both obtained from the grayscale image is considered. The gray scale image is obtained by converting the FCGR matrix values into equivalent gray values. A feature vector is formulated which is an eight element matrix, which are actually eight properties of the image. Out of these eight properties, four - variance, skewness, kurtosis and entropy are obtained from pixel occurrence probability matrix and the other four - contrast, correlation, energy and homogeneity are obtained from gray co-matrix. Thus a feature vector having eight elements corresponding to the eight properties of the texture image characterizing the organism is extracted. Fig. 1 gives the CGR image of NC_000928Echinococcus multilocularis and Fig. 2 the corresponding gray image,

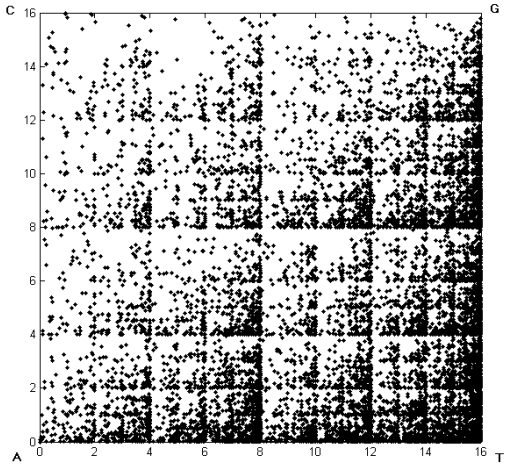

FIGURE 1: CGR Image of NC_000928 Echinococcus Multilocularis.

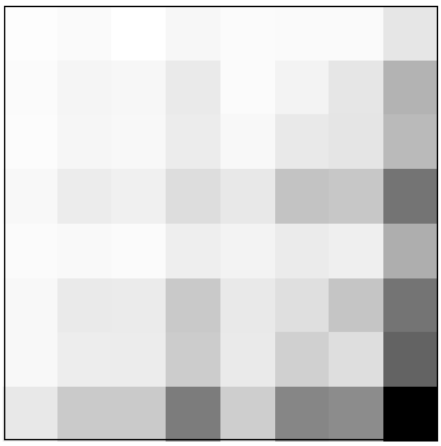

FIGURE 2: Grayscale Equivalent of CGR Image Shown in Figure 1.

2.2.3 Pixel occurrence probability matrix. First-order texture analysis measures use the image histogram, or pixel occurrence probability, to calculate texture. The main advantage of this approach is its simplicity through the use of standard descriptors (e.g. mean and variance) to characterize the data [9].

Assume the image is a function f(x,y) of two space variables x and y, x=0,1…L-1 and y=0, 1... M-1. The function f(x,y) can take discrete values i = 0, 1….G-1, where G is the total number of number of intensity levels in the image. The intensity-level histogram is a function showing (for each intensity level) the number of pixels in the whole image. The histogram contains the first-order statistical information about the image (or its fragment). Dividing the number of pixels having a given intensity value by the total number of pixels in the image gives the approximate probability density of occurrence of the intensity levels [10].If N(i) is the number of pixels with intensity i and M is the total number of pixels in an image, it follows that the histogram, or pixel occurrence probability, is given by

$$P(i)=N(i)/M \qquad (3)$$

Out of several properties of the texture image, 4 were selected which showed maximum variation in values for different organisms under different classes. The four properties are variance, skewness, kurtosis, and entropy. For a random variable X with mean μ and standard deviation σ and expectation value E, the different properties are:

Variance: $\text{var}(X) = E\left[(X-\mu)^2\right]$ (4)

Skewness: $\frac{E(X-\mu)^3}{\sigma^3}$ (5)

Kurtosis: $\frac{E(X-\mu)^4}{\sigma^4}$ (6)

Entropy: $-sum(p.*\log 2(p))$ (7)

p is the histogram counts returned from histogram image.

The variance is a measure of the amount of variation of the values of that variable from its expected value or mean. The skewness is a measure of asymmetry of the data around the sample mean. The skewness is zero if the histogram is symmetrical about the mean, and is otherwise either positive or negative depending whether it has been skewed above or below the mean.. The kurtosis is a measure of flatness of the histogram. For a normal distribution the kurtosis is three and for other cases it will be greater than or less than three. Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image [11]. These four properties constitute the four elements of the feature vector obtained from the texture analysis of the image.

2.2.4 Graycomatrix

The different properties of the graycomatrix are known as the graycoprops. There are four properties for this matrix. The graycomatrix creates a gray-level co-occurrence matrix (GLCM) from an image. Graycomatrix creates the GLCM by calculating how often a pixel with gray-level (grayscale intensity) value i occurs horizontally adjacent to a pixel with the value j. Each element (i, j) in the GLCM specifies, the number of times that the pixel with value i occurred horizontally adjacent to a pixel with value j. The graycomatrix calculates the GLCM from a scaled version of the image. By default, if 'I' is a binary image, graycomatrix scales the image to two gray-levels. If 'I' is an intensity image, graycomatrix scales the image to eight gray-levels [10].

Graycoprops normalizes the gray-level co-occurrence matrix (GLCM) so that the sum of its elements is equal to 1. Each element (r,c) in the normalized GLCM is the joint probability occurrence of pixel pairs with a defined spatial relationship having gray level values r and c in the image. Graycoprops uses the normalized GLCM to calculate properties. The four properties are:

Contrast: It returns a measure of the intensity contrast between a pixel and its neighbor over the whole image.Contrast is 0 for a constant image.

Correlation: It returns a measure of how correlated a pixel is to its neighbor over the whole image. Its range is between -1 and +1.Correlation is 1 or -1 for a perfectly positively or negatively correlated image. Correlation is 'NaN' (not a number) for a constant image.

Energy: It returns the sum of squared elements in the GLCM. Its range is between 0 and 1. Energy is 1 for a constant image.

Homogeneity: It returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal. Its range is between 0 and 1.Homogeneity is 1 for a diagonal GLCM [10], [12].

2.2.5 Wavelet decomposition. The wavelet decomposition of a signal f(x) is performed by a convolution of the signal with a family of basis functions. In the case of two-dimensional images, the wavelet decomposition is obtained with separable filtering along the rows and along the columns of an image. The wavelet analysis can thus be interpreted as image decomposition in a set of independent, spatially oriented frequency channels. The HH sub image represents diagonal

details (high frequencies in both directions – the corners), HL gives horizontal high frequencies (vertical edges), LH gives vertical high frequencies (horizontal edges), and the image LL corresponds to the lowest frequencies. At the subsequent scale of analysis, the image LL undergoes the decomposition using the same g and h filters, having always the lowest frequency component located in the upper left corner of the image [13]. In the case of a 3-scale analysis, 10 frequency channels can be identified. The size of the wavelet representation is the same as the size of the original image. As there is a choice of particular wavelet function for image analysis, symmetric wavelet functions appear superior to non-symmetric one, which is attributed to the linear-property of symmetric filters.

2.2.6 Support Vector Machines. Support Vector Machine classifier. Support Vector Machine was introduced to solve dichotomic classification problems [14] & [15]. Given a training set in a vector space, SVMs find the best decision hyper plane that separates two classes. The quality of a decision hyper plane is determined by the distance between two hyper planes defined by support vectors. The best decision hyper plane is the one that maximizes this margin. SVM extends its applicability on the linearly non-separable data sets by either using soft margin hyper planes or by mapping the original data vectors into a higher dimensional space in which the data points are linearly separable. There are several typical kernel functions. In this work, Support Vector Machine with Radial Basis kernel function and Polynomial kernel functions are used.

3. RESULTS AND DISCUSSION

The work aims to investigate the quality of features derived from the texture analysis and wavelet decomposition of a grey scale image of CGR (Chaos Game Representation) plot of each organism, evaluated through classification. The data set used was mitochondrial DNA sequences. The mitochondrial DNA sequences(DNA in mitochondria of a cell) of 2670 eukaryotic organisms belonging to eight categories of taxonomical hierarchy were obtained from National Centre for Biotechnology Information (NCBI) organelle database.

The total number of organisms in each class is first divided in 1:1 ratio to get two data sets as test and train. The Chaos game representation (CGR) of each sequence is obtained. Subsequently the FCGR matrix is computed and the corresponding gray scale images plotted. The feature vector, which corresponds to different properties of the grey image, is obtained for the different organisms in eight classes. These feature vectors are given as the input to the SVM classifier which is trained using the training set and tested using the test set. Using wavelets, 3 levels of decompositions were considered.

Previous works report using FCGR matrix elements as features for analysis of genomes [1], [2], [3], [7]. For huge sequences, since the screen resolution and computing resolution is limited, there will be error while computing the FCGR matrix. Hence this method is a novel attempt to provide an alternative to FCGR in such cases where huge sequences are involved. The work proves that the texture features as well the wavelet coefficients could be potential elements representing features of genomes. Other image transform coefficients as well as other image features can also be subjected to investigation in future, which may prove to be better representatives of genome sequences.

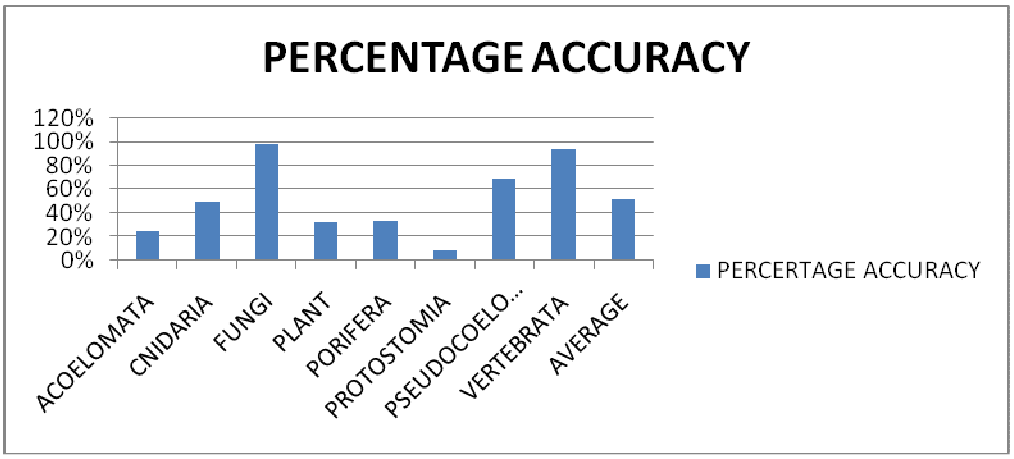

FIGURE 3: Percentage Accuracy of Classification Using Texture Analysis.

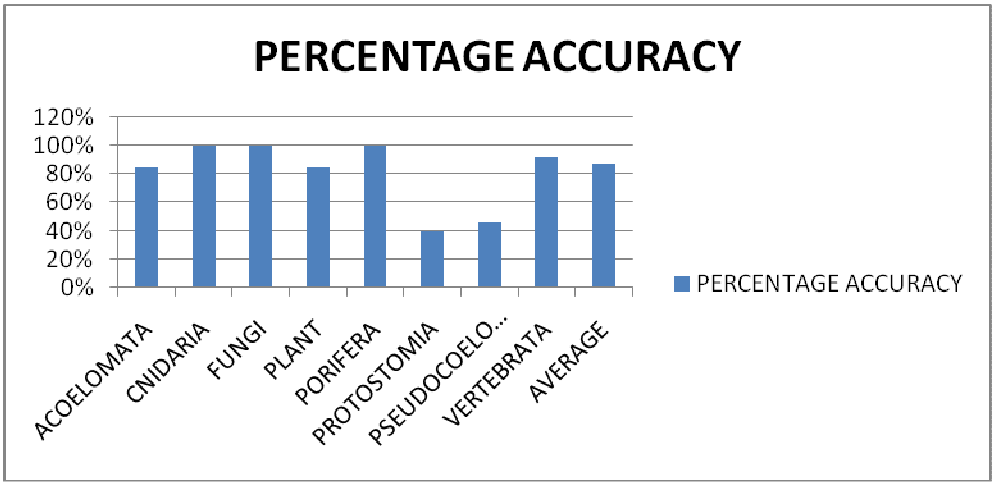

FIGURE 4: Percentage Accuracy of Classification Using 3 Level Wavelet Decomposition.

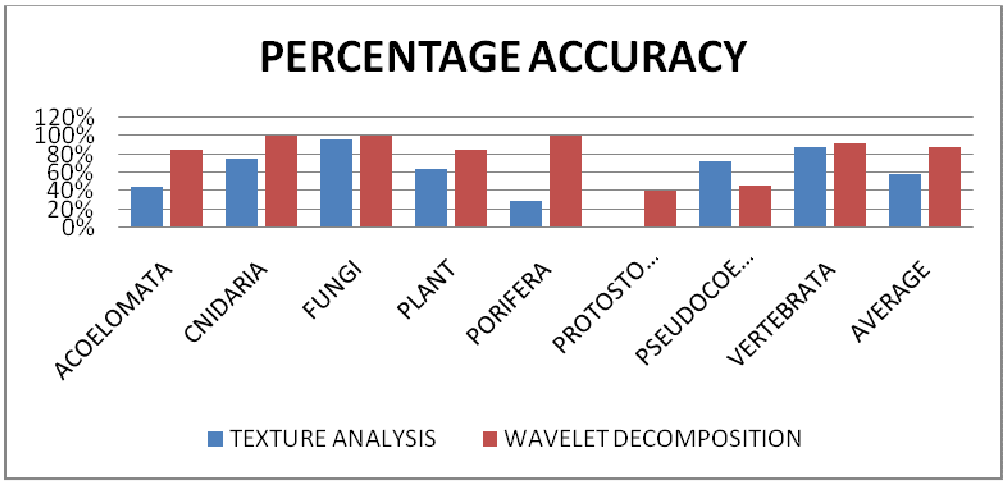

FIGURE 5: Texture Features and Wavelet Coefficient of The Texture Image – A Comparison of Feature Vectors Based on Classification Accuracy.

4. CONCLUSION

This work is an investigation into the quality of features derived from the texture analysis and wavelet decomposition of a grey scale image of CGR(Chaos Game Representation) plot of genomes evaluated through classification of organisms. The feature vectors were used to classify organisms with the help of an SVM classifier. The accuracy of classification stands testimony to the possibility of deriving feature vectors from the texture equivalent of CGR or more precisely – the FCGR matrix thus overcoming the inherent resolution error in FCGR matrices when considered quantitatively. It is thus concluded that coarse features such as texture and wavelet coefficients thereof characterise the genome sequences effectively.

5. REFERENCES

1. J. S Almeida, J. A. Carrico, A. Maretzek, P. A. Noble and M. Fletcher, "Analysis of genomic sequences by chaos game representation." Bioinformatics, vol. 17, (5), pp. 429–437, Jan. 2001.

2. P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot and B. Fertil, "Genomic signature: characterization and classification of species assessed by chaos game representation of sequences." Mol. Biol. Evol., vol 16, pp. 1391–1399, Jun. 1999.

3. V. V. Nair and A. S. Nair, "Combined classifier for unknown genome classification using chaos game representation features" in Proc. International Symposium of Bio Computing, 15-17 February 2010.

4. V. V.Nair, A. Mallya, B. Sebastian, I. Elizabeth, and A. S. Nair, Hurst CGR (HCGR) – A Novel Feature Extraction Method from Chaos Game Representation of Genomes: ACC2011 Springer Proceedings, June 2011.

5. Internet: http://www.ncbi.nlm.nih.gov/Genomes/ ORGANELLES/ organelles.html, [01/03/2012].

6. H. J. Jeffrey, "Chaos game representation of gene structure". Nucleic Acids Res., vol. 18,8, pp. 2163–2170, Mar. 1990.

7. J. Joseph and R. Sasikumar, "Chaos game representation for comparison of whole genomes". BMC Bioinformatics, vol. 7, 243, May 2006.

8. A. Materka, M. Strzelecki, "Texture Analysis Methods – A Review.", University of Lodz, Institute of Electronics, COST B11 report, Brussels 1998.

9. W. H. Nailon, "Texture Analysis Methods for Medical Image Characterization", Department of Oncology Physics, Edinburgh Cancer Centre & School of Engineering, University of Edinburgh, United Kingdom.

10. Manual: MATLAB, R2010a, Image Analysis and Statistics, Texture Analysis, Graycomatrix, The MathWorks Inc.

11. D. Avola, L. Cinque and G. Placidi, "Medical Image Analysis Through A texture based Computer Aided Diagnosis Framework." Int. J Biomet. Bioinf., vol. 6, (5), pp. 144-152, Oct. 2012.

12. S. A. Angadi and M. M. Kodabagi, "A Texture Based Methodology for Text Region Extraction from Low Resolution Natural Scene Images." Int. J Image Proc., vol. 3, (5), pp. 229-245, Nov. 2009.

13. K.P. Soman, K. I. Ramachandran and N. G. Resmi, Insight into Wavelets, PHI Learning Pvt. Ltd., 2010.

14. N. Cristianini and J.S. Taylor, Support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge , 2000.

15. V. N., Vapnik. The Nature of Statistical Learning Theory. Berlin: Springer-Verlag, 1995.