Scientific Journal Impact Factor (SJIF): 1.711

IJMTER

International Journal of Modern Trends in Engineering and Research

ISSN (ONLINE) : 2349 - 9745

ISSN (PRINT) : 2393 - 8161

A Survey on Chaos Game Representation for Online Signature Verification

Aswathy K.V.
Dept. of CSE, SBCEW

Abstract—This paper presents a survey on a new technology known as Chaos Game Representation (CGR), which has already been implemented in the field of biochemistry. This method has permitted to represent the patterns in sequences. So it was first proposed for DNA sequences. CGRs produce many interesting patterns. Thus it can be a new tool for pattern recognition. This survey also checks whether CGR can be used for signature verification.

Keywords—chaos game representation, chaos theory, DNA nucleotides, proteins, fractals.

I. INTRODUCTION

Chaos Game Representation is a graphical representation of a sequence. The word ‘Chaos’ derived from the Greek word ‘Khaos’ refers to unpredictability. Many researchers focused on the chaos theory, which is used to study dynamical systems. Since the state of the dynamical systems cannot be predicted, this theory is mostly intended to them. The Chaos theory is the field of mathematics. It has many applications in several disciplines such as meteorology, sociology, physics, engineering, economics, biology, and philosophy.

1.1. Chaos Game

The term ‘chaos game’ defined in mathematics was coined by Michael Barnsley and is a method of creating fractals using polygons. A fractal is an object or quantity that displays self-similarity on all scales. Self-similar object is nothing but an object that looks roughly the same on any scale. Fractals are class of self-similar objects. The chaos game is an algorithm which produces the pictures of fractal structures using paper, pencil or computer. The picture produced by the chaos game is known as the attractor.
The Chaos game plays as follows:

- Take three points; color them as red, green and blue.
- Place the tree points in a triangular manner and we can call them as the vertices of the triangle.
- Next, take a die and color its two sides with red, green and blue.
- Now start the game from an initial point called seed point by rolling the die. The seed point can be a point anywhere in the triangle.
- In the first roll of the die, if we get the red colored face, then move the seed point half the distance to the red colored vertex or point and the point that last we plotted will be the next seed point.
- In the second roll of the die, if we get the green colored face, then move the seed point half the distance to the green colored vertex and so on.

The figure given below shows the chaos game:

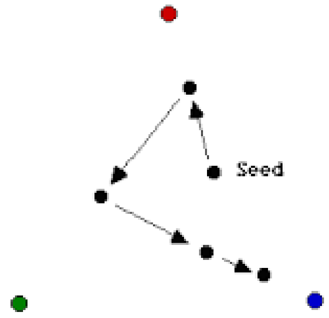

Figure 1. Playing Chaos Game

When the die is rolled hundreds of times, the chaos game produces a pattern of points. So the main goal of this game is to find out what the resulting pattern will be. When this game continues, it will produce a pattern what the mathematician called Sierpinski triangle. The figure 2 shows this triangle in which the upper triangle is in red color, the lower left triangle is in green, and the lower right triangle is in blue.

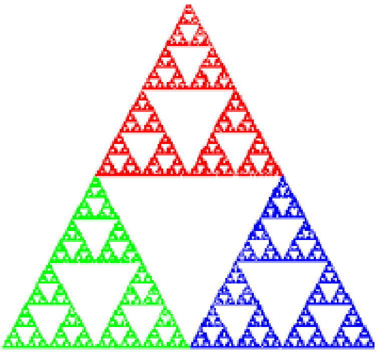

Figure 2. The Sierpinski triangle

II. CHAOS GAME REPRESENTATION OF DNA SEQUENCES

Chaos Game is an algorithm whose input is a sequence of letters and the output is an image, which is CGR of that sequence. The chaos game produces a pattern of any sequences and it is known as the Chaos Game Representation (CGR) of that sequence. Since early 1990's the use of CGR for representing biological sequences such as DNA has been investigated. Usually a genetic sequence has been treated as a string of nucleotides such as A, T, G, and C. To derive the CGR of DNA sequence, first draw a square having desired size and the letters A, T, G and C are plotted to the four corners of the square. The positions of A, T, G and C, are (0, 0), (1, 0), (1, 1) and (0, 1) respectively and the center point, P_0 is at the position (0.5, 0.5) as shown in figure 3. Instead of rolling a 4-sided die, take each next base in the sequence to plot the points. For plotting the CGR of the given sequence, we start from the center point P_0. The processing steps for plotting CGR are given below:

1. Choose the first nucleotide from the given sequence.

2. Calculate the midpoint between the center (x_c, y_c) of the square and the first nucleotide (x_n, y_n). Let (x_i, y_i) be the midpoint to be plotted, then $x_i = (x_c + x_n) = 2$ and $y_i = (y_c + y_n) = 2$.

3. Repeat the following steps until the final nucleotide in the sequence is plotted. Read next nucloetide. Then calculate the midpoint between the last point plotted (x_i, y_i) and the coordinate of the newly read nucleotide. $x_i+1 = (x_i + x_n)=2$ and $y_i+1 = (y_i + y_n) = 2$.

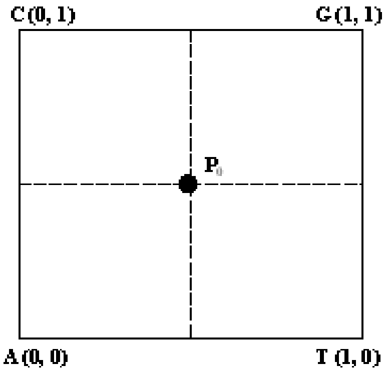

Figure 3. CGR square with each nucleotide assigned to corners

III. LITERATURE REVIEW

This technique was first proposed for DNA sequences by H. Joel Jeffrey in 1990 in his paper chaos game representation of gene structure [1]. In this, he took the sequence of DNA nucleotides (A - Adenine, G - Guanine, C - Cytosine, and T - Thymine or U - Uracil) and are assigned to the four corners of a square as shown in figure 3. Then, instead of rolling a die, he used the next base (a, c, g, t/u) to pick up the next point.

He discovered the CGR pattern of the first 6 bases of the GenBank sequence HUMHBB (Human beta globin region, chromosome 11) ‘gaattc’ and obtained the characteristic pattern as shown in figure 4. The main characteristic of this CGR is the empty area in the upper right quadrant. He also found many interesting CGR patterns of certain groups of genes.

Andras Fiser et al. [2] proposed the chaos game representation for visualizing and analyzing the primary and the 3-dimensional structures of proteins. Zu-Guo Yu et al [3] has proposed the CGR of protein sequences based on the detailed HP model and also made a multifractal and correlation analysis. The protein consists of twenty different kinds of amino acids. The HP model is the well-known model of protein sequences, in which the twenty kinds of amino acids are divided into two groups, hydrophobic (H) (or non-polar) and polar (P) (or hydrophilic). In the detailed HP model, Zu-Guo Yu et al divided into four groups; non-polar, negative polar, uncharged polar, and positive polar.

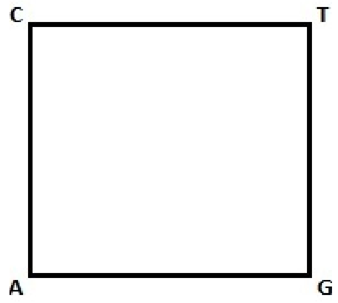

Figure 3. CGR Plot with Each Corner Assigned to Nucleotides

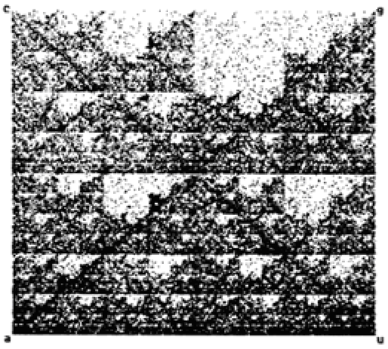

Figure 4: CGR of Human Beta Globin Region on Chromosome

K. Manikandakumar et al. [4] applied the CGR method for the structural analysis of protein sequences by classifying them into five and six groups based on their characteristics such as Pentagon and Hexagon structures. Iman Tavassoly et al. [5] proposed a three dimensional visualization of genomic sequences based on the chaos game representation as shown in figure 5. In this method instead of four points as four bases in nucleotide sequences, eight points are considered as eight corners of a cube. Four different nucleotides in protein coding regions of the sequence (A, C, G and T) and four different nucleotides in noncoding regions (A, C, G and T) are eight corners of the cube. The 3D CGR images are applicable to the analysis of the whole genome.

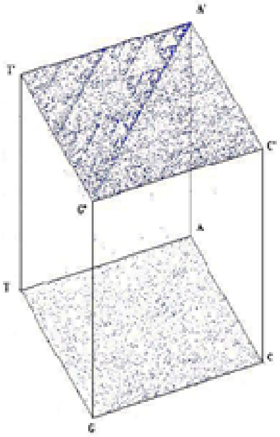

Figure 5. Three dimensional CGR of Arcanobacterium pyogenes plasmid pAP2 genome.

J.M Gutierrez et al [6] presented the application of CGR method in climatology to identify the homogeneous regions according to their climate conditions. They have shown that different climates exhibit characteristic patterns with different fractal exponents and entropies. As an investigation of distance measures in proteins for studying the patterns in the structural classification of protein sequences, a research work has used CGR[7]. The statical analysis of genomic sequences [8] using static measures and techniques can be devised using CGR. For the sequence comparison, a kind of information can be derived from the CGR images and is known as frequency chaos game representation (FCGR). The alignment-free comparisons of sequences based on the frequencies of nucleotides have been introduced in [9].

The CGR has many applications in Bioinformatics as given below:

- Characterization and classification of species.

- Identify and compare different organisms.
- Visualizing and analyzing Genomic sequences.
- To locate specific regions such as Intron, exon, promoter sites, transcription factor binding sites, etc.
- Visualizing and analyzing protein sequences like, Protein Pattern Identification
- The plot of entire genome of an organism and 100Kb of the genome will have the same pattern for the CGR. Because of this property the CGR can be treated as Genome signatures.
- It can be used to identify unknown DNA fragments.
- Horizontal transfers in the genome can easily be identified.

IV. CGR FOR ONLINE SIGNATURE VERIFICATION

The Chaos Game Representation is a tool to visualize the nucleotides and amino acid sequences and to study their fractal properties. It is an effective method for pattern recognition and for visualizing any structural features if they are given as a sequence of elements. In the case of signature verification it can be a new tool for identifying the genuine signatures by representing the features of signatures in Chaos Game Representation. By representing the signature in CGR, it will be possible to track the signature path and thus will obtain a pattern. These patterns of signatures can be compared whether they are matching or not. The CGR on signature verification has not yet been applied.

V. CONCLUSION

This paper has made a review of Chaos Game Representation that was applied in the field of Biochemistry. While going through many of the works, the application of CGR has been utilized in DNA and Protein sequences. Most of the researchers focused on the pattern formation of both the DNA and proteins. Since CGR is providing a pattern of particular sequences of structural elements, it can be recognized as a generalized method for pattern matching. The scope of CGR is not confined to these bio-sequence representations, and is enormous.

REFERENCES

[1] H. Joel Jeffrey, “Chaos Game Representation of gene structure,” Nucleic Acids research, Vol 18, No. 8, pp. 2163-2170.

[2] Andras Fiser, Gabor E. Tusnady, and Istvan Simon, “Chaos Game Representation of Protein structures”,J. Mol. Graphics, Vol. 12, pp. 302-305, Dec. 1994.

[3] Zu-Guo Yu,Vo Anh, and Ka-Sing Lau, “Chaos Game Representation of protein sequences based on the detailed HP model and their multifractal and correlation analysis”, Program in statistics and operations research, Queensland University of technology.

[4] K. Manikandakumar, K. Gokulraj, S. Muthukumaran and R. Srikumar, “Graphical Representation of Protein sequences by CGR: Analysis of Pentagon and Hexagon structures”, Middle-East Journel of scientific research 13(6): 764-771, 2013,ISSN 1990-9233.

[5] Iman Tavassoly, Omid Tavassoly, Mohammad Soltany Rezaee Rad, Negar Mottaghi Dastjerdi, “Three Dimensional Chaos Game Representation of Genomic Sequences”, IEEE comp. Society, 2007.

[6] J. M. Gutierrez,A. Galvan And A. S. Cofino, “Chaos Game Characterization of Temporal Precipitation Variability: Application To Regionalization”, Fractals, Vol. 14, No. 2 (2006) 87- 99

[7] Manju Susan Thomas, "Investigation of Distance Measure in Proteins using Chaos Game Representation ", Centre for Bioinformatics, University of Kerala, March 2006.

[8] Almeida JS, Carrico JA, Maretzek A, Noble PA and Fletcher M, "Analysis of genomic sequences by chaos game representation ", Bioinformatics, 17(5):429-437.

[9] J. Joseph and R. Sasikumar, "Chaos game representation for comparison of whole genomes", BMC Bioinformatics, Vol. 7, No. 243, May 2006.