

## Matrix Frequency Analysis of *Oryza Sativa* (japonica cultivar-group) Complete Genomes

K. Manikandakumar<sup>1\*</sup>, S. Muthu Kumaran<sup>2</sup>, R. Srikumar<sup>3</sup>

<sup>1</sup>Department of Physics, Bharathidasan University College (W), Orathanadu – 614 625,  
Tanjavore District, Tamil Nadu, India

<sup>2</sup>Department of Physics, Nehru Memorial College, Puthanampatti – 621 007,  
Tiruchirappalli District, Tamil Nadu, India

<sup>3</sup>Department of Microbiology, Bharathidasan University College (W), Orathanadu – 614 625,  
Tanjavore District, Tamil Nadu, India

\*Corresponding author: K. Manikandakumar, Department of Physics,  
Bharathidasan University College (W), Orathanadu – 614 625, Tanjavore District,  
Tamil Nadu, India, Tel: 9787383327; E-mail: [bioinfokm@gmail.com](mailto:bioinfokm@gmail.com)

Received January 28, 2009; Accepted April 20, 2009; Published April 22, 2009

**Citation:** Manikandakumar K, Kumaran MS, Srikumar R (2009) Matrix Frequency Analysis of *Oryza Sativa* (japonica cultivar-group) Complete Genomes. J Comput Sci Syst Biol 2: 159-166. doi:10.4172/jcsb.1000027

**Copyright:** © 2009 Manikandakumar K, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

### Abstract

The genome sequence information is essential to understand the function of extensive arrangements of genes. It is significant to combine all sequence information in a precise database to provide an efficient manner of sequence similarity search. The complete genome analysis, which is one of the essential steps to know their characteristics, is very important. Complete genome analysis is depends on matrix frequency of sequence residue calculation and CGR analysis. In this study, we select rice as the specimen for complete genome analysis. Rice is one of the most essential cereal crops providing food for more than half of the world's population. *Oryza sativa* (japonica cultivar-group) species is an important cereal and model monocot. We have generated a matrix frequency for genetic code analysis, which helps in the study of complete genome residues. Here we report the duplets and triplets codon for genetic code analysis of *O. sativa* chromosomes. We illustrate a new method of Chaos Game Representation, which produces the objects possessing self-similar structure. As per our findings, the average matrix frequency of stop codons is similar to the matrix frequency of start codon. This average is seems to be similar in the complete genome sequences of every *Oryza sativa* (japonica cultivar-group) chromosomes.

**Keywords:** *Oryza sativa* (japonica cultivar-group); Chaos game representation; Chromosome; Matrix frequency; Fractal structure

### Introduction

DNA is a double anti-parallel helix built by concatenating nucleotide blocks. Several physicochemical properties of DNA depends on the interactions between consecutive bases, thus, the classification of patterns from nearest neighbor bases could help in the description of nucleotide sequences (Mohanty and Narayana Rao, 2000). However, (Almeida et al., 2001) have followed the scale independence of CGR of genetic sequence method to investigate local and global homology. The two patterns iden-

tified from the analysis of whole genomes and the number of different dinucleotides are unequal frequencies of manifestation of some asymmetric pairs and preferences of certain nucleotides with specific nearest neighbors over equivalent dinucleotides (Nussinov, 1980; Nussinov, 1981).

Small plant chromosomes, such as those in rice, often show irregular condensation at mitotic prometaphase. Thus, the condensation pattern appearing at prometaphase was

only a morphological landmark to divide the rice chromosomes into sub-regions. Characteristics of each rice chromosome with uneven condensation have quantitatively been analyzed by using image analysis methods. (Fukui, 1985; Iijima and Fukui, 1991) developed a method for identifying rice chromosomes based on a flow chart that consists of 11 discriminates, which classify specific chromosome groups. All rice chromosomes have identified and numbered by comparing the categories given by discriminates, one after another. The chromosomal spread is worth analyzing if, chromosomes 4, 11, and 12 are distinguishable by visual inspection and if chromosomes 1, 2, and 3 are completely recognized. If these six chromosomes are identified using discriminates 1 to 6 in order, then there is a great possibility of identifying all 12 chromosomes within the particular spread. The relevance of accessing the frequency of non-integer genomic sequences may not be apparent at first given that (Almeida et al., 2001) physically make all sequences of integer number of nucleotides.

The genome sequence information is indispensable in understanding the function of the wide array of genes that constitute the rice plant. Therefore, it is important to consolidate all sequence information in a specified database to provide an efficient method of sequence similarity search that eliminates artifactual matches be analysed by (Yoshiaki et al., 2003). We have generated a matrix frequency for genetic code analysis with all available rice genome for *Oryza sativa* japonica-cultivar group. (Gao et al., 2005) analysed that DNA shuffling is a direct evolution process which generates genetic diversity through the recombination of parental sequences in order to evaluate which pair of sequences could potentially produce the best result. *Oryza sativa* (japonica cultivar-group), a subspecies of rice, is an important cereal and model monocot (35884299 bp). The rice genome sequence provides a foundation for the improvement of cereals, our most important crops (Stephen et al., 2002). Experiments of direct evolution have successfully used to improve specific biological functions. (Fuentes et al., 2005) analyses Genetic diversity of rice varieties (*Oryza sativa* L.) based on morphological, pedigree and DNA polymorphism data, Plant Genetic Resources and phenotypic, genealogical, RAPD and AFLP diversity groups.

Mathematical characterization of DNA sequences could help in the understanding of structural relationships among different whole genomes along the chromosomes. The de-generated translation of trinucleotide codons encode for 20 amino acids, and remaining three nonsense codons signal for the end of transcription. Base concentrations, stretches and patches are the main factors explaining the variability observed among sequences (Deschavanne et al., 1999). The genomic signature as expressed in terms of short nucle-

otide usage extends and generalizes the genomic signature and it takes advantages of whole genome data reveals genome wide trends (Karlín and Burge, 1995). The measure of similarity using CGR can be the basis of a new set of algorithms to align sequences with considerable advantages over the conventional scoring methods (Almeida et al., 2001). The CGR is a formalism that bridges between sequence of discrete units and numeric coordinates in a continuous space. Consequently, basic statistic measures and techniques have applied to sequences and a wide range of new tools have devised for statistical analysis. Here we report the genetic code analysis of complete genome of all chromosomes of *O. sativa*. We have generated a matrix frequency of the rice genome. We describe a new method of Chaos Game Representation applied to *O. Sativa* (japonica cultivar-group) species sequences, which produces fractal objects possessing self-similar structure.

## Material and Methods

The *Oryza sativa* (japonica cultivar-group) Eukaryote complete genomes have downloaded from the GOLD (<http://www.genomesonline.org/>) database. The species have been totally 12 chromosomes. The details of the chromosomes are giving below.

[<http://www.genomesonline.org/gold.cgi?want=Published+Complete+Genomes>]

Ch. No.	Web link
1	<a href="http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008394">http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008394</a>
2	<a href="http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008395">http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008395</a>
3	<a href="http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008396">http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008396</a>
4	<a href="http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008397">http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008397</a>
5	<a href="http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008398">http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008398</a>
6	<a href="http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008399">http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008399</a>
7	<a href="http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008400">http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008400</a>
8	<a href="http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008401">http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008401</a>
9	<a href="http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008402">http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008402</a>
10	<a href="http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008403">http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008403</a>
11	<a href="http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008404">http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008404</a>
12	<a href="http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008405">http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&amp;val=NC_008405</a>

A simple model, which permits the simulation of these features of nucleotide residues, is discrete time Markov Chain (Goldman, 1993). In this model, a 4 X 4 matrix, P defines the probabilities with which subsequent bases follow the current base in a nucleotide residue. If the base labels A, T, G, and C are equated with the numbers 1, 2, 3 and 4; then  $P_{ij}$  is the  $j^{\text{th}}$  element of the  $i^{\text{th}}$  row of P which defines the probability that base j follows base i. The row sums of P must equal 1. Using this matrix, a simulated nucleotide residue may be obtained by selecting a first base randomly according to the frequencies of the bases in the nucleotide residue under study. If the base is i, then the probabilities will be  $P_{i1}$ ,  $P_{i2}$ ,  $P_{i3}$  and  $P_{i4}$ . These probabilities are used to select the next base, and so on until the simulated sequence is the same length as the original nucleotide residues.

This first-order Markov Chain model is in which successive bases in a residue depend only on the preceding base. The probabilities in the matrix P may be estimated by direct calculation from the residues dinucleotide frequencies. If the dinucleotide XY is observed  $n_{xy}$  times in the sequence, then probability  $P_{xy}$  is estimated by  $n_{xy} / (n_{xA} + n_{xT} + n_{xG} + n_{xC})$ . This permits a protein sequence to be simulated with both individual base frequencies and digroup frequencies matching those of the original sequence. Dinucleotide frequencies ( $n_{xy}$ ) and Markov Chain probabilities ( $P_{xy}$ ) for the *Oryza sativa* (japonica cultivar-group) genomes are given in Table - 2.

The first-order Markov Chain model successfully recreates other genomes. The lack of banding suggests approximate equality of the frequencies of the bases A, C, G, and T, confirmed by direct calculation from the residues. The first-order Markov Chain model will not give the observed patterns, but a more complex second-order Markov Chain, in which each base depends on the previous two, does. Second-order Markov Chains have been used to describe both structure and with-in-structure of nucleotide residues.  $P_{XYZ}$ , the probability that base Z follows the trigroup XYZ, is estimated directly from the nucleotide residues trigroup frequencies  $n_{XYZ}$  using the formula  $P_{XYZ} = n_{XYZ} / (n_{XYA} + n_{XYT} + n_{XYG} + n_{XYC})$ . Trinucleotide frequencies ( $n_{XYZ}$ ) and Markov Chain probabilities ( $P_{XYZ}$ ) for the *Oryza sativa* (japonica cultivar-group) genomes are given in Table - 3.

We apply the CGR method to *Oryza sativa* (japonica cultivar-group) species by considering the four different nucleotide residues into four groups namely Adenine, Thymine, Guanine and Cytosine. Using this distinctive way of CGR technique, the *Oryza sativa* (japonica cultivar-group) species produces the intrinsic fractal structure. The percentage values of nucleotide residues of the four groups available

in the species under consideration has also computed and used for analysis. We find that some of the species of *Oryza sativa* (japonica cultivar-group) nucleotide sequences produce the similar kind of self-similar fractal structure. The CGR shows the characteristics of the *Oryza sativa* (japonica cultivar-group) genome.

To begin with, let us generate the typical fractal object namely the 'Square' possessing the self-similar structure using the Chaos Game Representation (CGR). Let us start with three vertices located at (0,0), (1,0), (0,1) and (1,1) labeled as A, T, C and G respectively. Now random sequences of 1, 2, 3 and 4 are obtained using a random number generator available in typical C compiler. In generating CGR, the  $n^{\text{th}}$  point of the attractor is simply the mid-point between the  $(n-1)^{\text{th}}$  point and the vertex corresponding to the  $n^{\text{th}}$  value. Similarly, the successive application of this procedure for 100,000 points produces the 'Square' as shown in Fig. 1, which is a typical fractal object possessing self-similar structure.

We calculate the nucleotide contents of the above species into grouping of four types name as A, T, G and C. Used in computer algorithm the nucleotide contents are differentiating to each group. Then we calculate the A+G and T+C ratios of the above species. Finally, the average ratio of the each chromosome is calculating by the method A+G/T+C. All the results are given by percentage values. The Table-1 is given by all the chromosome details of the *Oryza sativa* (japonica cultivar-group) species. The CGR plots are drawn using Gnu plot method.

## Pictorial Representation

Chaos Game Representation (CGR) for gene (or DNA) sequences was introduced by (Jeffrey, 1990; Jeffrey, 1992) and the essential structures of genome sequences of a few model organisms were obtained using CGR plots. Each chromosome has been taken in above 200,000 base pairs. Therefore, we did not represent the whole genomes. We have taken only first 100,000 base pairs nucleotide sequences for the above representation.

## Results and Discussions

The structure of DNA is specific to each species and undergoes only slight variations along the whole genome (Deschavanne et al., 1999). Diversity among species is considerable and is primarily a consequence of base concentration, stretches of bases with unusual frequencies. The frequencies of occurrence are to point out the basis of the genome (Deschavanne et al., 1999). In our analysis is giving the matrix frequency calculation of every chromo-

some complete genome sequences. We analysed every chromosome and given in the Table - 1 is shown by the individual nucleotide contents percentage. The table - 2 has shown by the first order Markov chain matrix frequency of all chromosomes and it is representing in dinucleotide codons. The Table - 3 is show by the second order Markov chain matrix frequency of all chromosomes and it is representing in trinucleotide codons. The Table - 4 is shown, the classification of triplets to occurring in which regions. The Table - 5 is shown by the relations of the start and stop codons of all chromosomes.

We analyze the complete genome of *Oryza sativa* (japonica cultivar-group) species chromosomes nucleotide contents and the calculation is giving in Table - 1. From the table-1, the chromosome 4 has been largest residues (17028043 base pairs). The lowest residues have been chromosome 11 (298736 base pairs). Chromosome no. 1 is having the lowest Adenine residues (26.99%) and chromosome 10 is having the highest adenine residues (28.84%). The range of Thymine residues is 27.81% got the chromosome no. 9 and 28.73% of thymine residues are having chromosome no.10. The range of Guanine ratio is 21.36% (chromosome 5) and 22.70 (chromosome 1). The range of Cytosine ratio is 21.19% (chromosome 10) and 22.16 (chromosome 1). The range of A+G content ratio is 49.55% (chromosome 3) and 50.18% (chromosome 2). The range of T+C content ratio is 49.82% (chromosome 2) and 50.45% (chromosome 3). The chromosome 8 has been representing in same nucleotide content ratio in Guanine and Cytosine residues (21.57%).

We generate and analyse the first order Markov chain matrix frequency for 16 (4 x 4) nucleotide doublet codons for *Oryza sativa* (japonica cultivar-group) species complete genome chromosomes and the matrix frequency for each doublet codon is given in Table-2. The Table-2 has shown, the AA codon minimum frequency range is 0.299 for chromosome-1 and the maximum frequency range is 0.321 for chromosome-10. The CA codon minimum frequency range is 0.289 for chromosome-1 and the maximum frequency range is 0.317 for chromosome-5. The GA codon minimum frequency range is 0.271 for chromosome-3 and the maximum frequency range is 0.285 for chromosome-5. The TA codon minimum frequency range is 0.223 for chromosome-1 and the maximum frequency range is 0.245 for chromosome-7. The AC codon minimum frequency range is 0.179 for chromosome-10 & 11 and the maximum frequency range is 0.189 for chromosome-3. The CC codon minimum frequency range is 0.232 for chromosome-2 and the maximum frequency range is 0.249 for chromosome-4. The GC codon minimum frequency range is 0.229 for chromo-

some-10 and the maximum frequency range is 0.248 for chromosome-3. The TC codon minimum frequency range is 0.223 for chromosome-1 and the maximum frequency range is 0.207 for chromosome-11. The AG codon minimum frequency range is 0.204 for chromosome-10 and the maximum frequency range is 0.224 for chromosome-1. The CG codon minimum frequency range is 0.160 for chromosome-5 and the maximum frequency range is 0.192 for chromosome-4 & 6. The GG codon minimum frequency range is 0.233 for chromosome-12 and the maximum frequency range is 0.250 for chromosome-9. The TG codon minimum frequency range is 0.221 for chromosome-10 and the maximum frequency range is 0.248 for chromosome-1. The AT codon minimum frequency range is 0.283 for chromosome-6 & 9 and the maximum frequency range is 0.296 for chromosome-10. The CT codon minimum frequency range is 0.266 for chromosome-4 and the maximum frequency range is 0.289 for chromosome-11. The GT codon minimum frequency range is 0.235 for chromosome-9 and the maximum frequency range is 0.245 for chromosome-10. The TT codon minimum frequency range is 0.306 for chromosome-1 and the maximum frequency range is 0.320 for chromosome-8.

We have generate and analyse the second order Markov chain matrix frequency for 64 (4 x 4 x 4) nucleotide triplet codons for *Oryza sativa* (japonica cultivar-group) species complete genome chromosomes and the matrix frequency for each triplet codon is given in Table-3. From Table-3, most of the highest nucleotide triplet codon is representing in AAA and TTT. The most of the lowest nucleotide triplets are AGC and TGC. All the above chromosomes, we identify the low sparseness regions are mostly played in species of triplets as AC-A,T; GCA, TC-A,T; AG-A,C, CG-A,T; GGC, TGA and TGC respectively. The highest sparse-ness regions are mostly played in species of triplets as AA-A,C,G; CA-A,C; GAA, ATT, CTT, GTT and TT-A,T respectively.

Table-4 has shown the frequency triplet codons have separated by four regions. The regions are classifying the frequency range of 0.125-0.199, 0.200-0.249, 0.250-0.299, and above 0.300 matrix frequency of triplet codons. This table is easy to analyse and to study, how many triplets are coming under particular range of frequency.

We have analysed the relations between the start codon and stop codon frequencies and it has given in Table-5. The genetic code is show in three types of stop codons. But the start codon is only one. Therefore, we tried to show one stop codon for every sequence. Our analysis has not succeeded. Nevertheless, the average of two-stop codon value

is nearly equal to the start codon. This Table-4 describes, separated and shown in start and stop codon for each chromosome in *Oryza sativa* (japonica cultivar-group). This table shows every start codon frequency is equal to the average of two-stop codon frequency. So this analysis is used to finding and expressed the start codon is equal to stop codon for every *Oryza sativa* (japonica cultivar-group) chromosome complete genome sequences.

The figure 1, is shown the CGR plot for the first 100,000 base pairs of 12 chromosomes of *Oryza sativa* (japonica cultivar-group) species, we identify the genomes are cross overlapping in A, G and T, C. Four triangles are connecting in the mid point of 0.5, 0.5, 0.5, and 0.5 respectively. The A-T region is keeping in more numbers of residues.

## Analysis of Individual Chromosome

### Chromosome 1

The chromosome 1 is, total of 301936 base pairs. From the Table-1, the highest percentage value is Thymine residue (28.15%). The lowest percentage value is Cytosine residue (22.16%). The Adenine and Guanine residues are 26.99% and 22.70%. The highest combination of nucleotide residues of T+C percentage is 50.31%. The ratio of A+G & T+C is 1.0. The triplet of chromosome 1, the highest tri-nucleotide is TTT (0.348%). The lowest tri-nucleotide is AGC (0.163%). From the Table-3, the matrix frequency of chromosome 1, high frequency of tri-nucleotide sequence has been representing in AAA, AAC, TTA, and TTT (above 0.300%). Tri-nucleotide sequence of 0.250% to 0.299% has been represented in AAG, AAT, CA-A,C,G; GA-A,C,G; TA-A,C,G; CC-G,T; GC-C,G; TC-C,G; AGT, CGG, GG-A,T; AT-A,C,T; CT-A,C,T; GT-A,C,T; TT-C,G. Tri-nucleotide sequence of 0.200% to 0.249% has been represented in CAT, GAT, TAT, AC-C,G,T; CC-A,C; GC-G,T; TC-G,T; AG-A,G; CG-A,C,T; GGG, TG-A,G,T; ATG,CTG and GGG. Tri-nucleotide sequence of 0.150% to 0.199% is representing in ACA, GCA, TCA, AGC, GGC, and TGC.

### Chromosome 2

The chromosome 2 is total of 662387 base pairs. The highest percentage value is Adenine residue (28.39%). The lowest percentage value is Guanine residue (21.79%). The Thymine and Cytosine residues are 27.98% and 21.84%. The highest combination of nucleotide residues of A+G percentage is 50.18%. The ratio of A+G & T+C content is 1.0. The triplet codon of chromosome 2, the highest triplet is AAA (0.351%). The lowest triplet codon is TGC (0.157%).

The matrix frequency of chromosome 2, the frequency of tri-nucleotide sequence has been represented in AA-A,C,G; CA-A,C; GAC, TAC, TT-A,T (above 0.300%). Tri-nucleotide sequence of 0.250% to 0.299% has been represented in AAT, CAG, GA-A,G; TA-A,G; CC-G,T; GCG, CGG, GG-A,T; AT-A,C,T; CT-A,C,T; GT-A,C,T; TT-C,G. Tri-nucleotide sequence of 0.200% to 0.249% has been represented in CAT, GAT, TAT, AC-C,G,T; CCC, GC-C,T; TC-C,G,T; AG-A,G,T; CG-A,T; GGG, TG-G,T; ATG,CTG and GTG. Tri-nucleotide sequence of 0.150% to 0.199% has been representing in ACA, CCA, GCA, TCA, AGC, CGC, GGC, TGA, and TGC.

### Chromosome 3

The chromosome 3 is total of 831805 base pairs. The highest percentage value is Thymine residue (28.34%). The lowest percentage value is Guanine residue (22.05%). The Adenine and Cytosine residues are 27.50% and 22.11%. The highest combination of nucleotide residues of T+C percentage is 50.45%. The ratio of A+G & T+C content is 1.0. The triplet codon of chromosome 3, the highest triplet is TTT (0.355%). The lowest triplet codon is TGC (0.162%).

The matrix frequency of chromosome 3, the frequency of tri-nucleotide sequence has been represented in AA-A,C; GTT, TT-A,C,T (above 0.300%). Tri-nucleotide sequence of 0.250% to 0.299% has been represented in AA-G,T, CA-A,C, GA-A,C,G; TA-A,C,G; CC-G,T; GC-C,G, CGG, GG-A,T; AT-A,C,T; CT-A,C,T; GT-A,C; TTG. Tri-nucleotide sequence of 0.200% to 0.249% has been represented in CA-G,T; GAT, TAT, AC-C,G,T; CC-A,C; GCT, TC-C,G,T; AG-A,G,T; CG-A,C,T; GGG, TG-G,T; ATG,CTG and GTG. Tri-nucleotide sequence of 0.150% to 0.199% has been representing in ACA, GCA, TCA, AGC, GGC, TGA and TGC.

### Chromosome 4

The chromosome 4 is total of 17028043 base pairs. The highest percentage value is Thymine residue (27.95%). The lowest percentage value is Cytosine residue (22.08%). The Adenine and Guanine residues are 27.88% and 22.10%. The highest combination of nucleotide residues of T+C percentage is 50.03%. The ratio of A+G & T+C is 1.0. The triplet codon of chromosome 4, the highest triplet is AAA (0.347%). The lowest triplet codon is AGC (0.169%). The matrix frequency of chromosome 4, the frequency of tri-nucleotide sequence has been represented in AA-A,C; CAA, CTT, TT-A,T (above 0.300%). Tri-nucleotide sequence of 0.250% to 0.299% has been represented in AA-G,T, CA-C,G; GA-A,C,G; TA-A,C,G; CCT, GC-G,T; TCC, CGG, GG-A,T; TGG, AT-A,C,T; CT-A,C; GT-A,C,T; TT-C,G. Tri-

nucleotide sequence of 0.200% to 0.249% has been represented in CAT, GAT, TAT, AC-C,G,T; CC-A,C,G; GCT, TC-G,T; AG-A,G,T; CG-C,T; GG-C,G, TG-A,T; ATG, CTG and GTG. Tri-nucleotide sequence of 0.150% to 0.199% has been representing in ACA, GCA, TCA, AGC, CGA, and TGC.

## Chromosome 5

The chromosome 5 is total of 476423 base pairs. The highest percentage value is Adenine residue (28.61%). The lowest percentage value is Guanine residue (21.36%). The Thymine and Cytosine residues are 28.47% and 21.55%. The highest combination of nucleotide residues of T+C percentage is 50.02%. The ratio of A+G & T+C content is 1.0. The triplet codon of chromosome 5, the highest triplets is AAA and TTT (0.354%). The lowest triplet codon is AGC (0.143%). The matrix frequency of chromosome 5, the frequency of tri-nucleotide sequence has been represented in AA-C,G; CA-A,C; GA-A,C(same ratio); TAC & TTC are the same ratio. CTT, GTT, TT-A; AAA and TTT are the same ratio of nucleotides (above 0.300%). Tri-nucleotide sequence of 0.250% to 0.299% has been represented in AAT, CAG; GAG; TA-A,G; CC-G,T, GCG=TCC; CGG, GG-A,T; AT-A,C,T; CT-A,C; GT-A,C; TTG. Tri-nucleotide sequence of 0.200% to 0.249% has been represented in CAT, GAT, TAT, AC-C,G,T; CCC, GC-C,T, TC-G,T; AG-A,G,T; CG-A,T; GGG, TG-G,T; ATG, CTG and GTG. Tri-nucleotide sequence of 0.150% to 0.199% has been representing in ACA, CCA, GCA, TCA, AGC, CGC, GGC and TGC.

## Chromosome 6

The chromosome 6 is total of 1949261 base pairs. The highest percentage value is Adenine residue (28.03%). The lowest percentage value is Cytosine residue (21.88%). The Thymine and Guanine residues are 27.96% and 22.12%. The highest combination of nucleotide residues of A+G percentage is 50.15%. The ratio of A+G & T+C content is 1.0. The triplet codon of chromosome 6, the range of triplets is TTT (0.359%) and TGC (0.172%). The matrix frequency of chromosome 6, the frequency of tri-nucleotide sequence has been represented in AA-A,C,G; CAA, GAA, GTT, TT-A,T of nucleotides (above 0.300%). Tri-nucleotide sequence of 0.250% to 0.299% has been represented in AAT, CA-C,G; GA-C,G; TA-A,C,G; CC-G,T, GC-C,G; CGG, GG-A,T; AT-A,C,T; CT-A,C,T; GT-A,C; TT-C,G. Tri-nucleotide sequence of 0.200% to 0.249% has been represented in CAT, GAT, TAT, AC-C,G,T; CC-A,C, GCT, TC-C,G,T; AG-G,T; CG-A,C,T; GG-C,G, TG-A,G,T; ATG, CTG and GTG. Tri-nucleotide sequence of 0.125% to 0.199% has been represented in ACA, GCA, TCA, AG-A,C and TGC.

## Chromosome 7

The chromosome 7 is total of 993326 base pairs. The highest percentage value is Adenine residue (28.67%). The lowest percentage value is Guanine residue (21.47%). The Thymine and Cytosine residues are 28.24% and 21.63%. The highest combination of nucleotide residues of A+G percentage is 50.14%. The ratio of A+G & T+C content is 1.0. The triplet codon of chromosome 7, the frequency of triplets is AAA (0.361%) and AGC (0.161%). The matrix frequency of chromosome 7, the frequency of tri-nucleotide sequence has been represented in AA-A,C,G; CA-A,C; TT-A,T of nucleotides (above 0.300%). Tri-nucleotide sequence of 0.250% to 0.299% has been represented in AAT, CAG, GA-A,C,G; TA-A,C,G; CC-G,T, GC-C,G; CGG, GG-A,T; AT-A,C,T; CT-A,C,T; GT-A,C,T; TT-C,G. Tri-nucleotide sequence of 0.200% to 0.249% has been represented in CAT, GAT, TAT, AC-C,G; CC-A,C, GCT, TC-C,G; AG-G,T; CG-C,T; GGG, TG-G,T; ATG, CTG and GTG. Tri-nucleotide sequence of 0.125% to 0.199% has been represented in AC-A,T; GCA, TC-A,T; AG-A,C, CGA, GGC, TGA and TGC.

## Chromosome 8

The chromosome 8 is total of 8367279 base pairs. The highest percentage value is Adenine residue (28.49%). The lowest percentage values are represented in Guanine and Cytosine residues (21.57%). The Thymine residue is 28.37%. The highest combination of nucleotide residues of A+G percentage is 50.06%. The ratio of A+G & T+C content is 1.0. The triplet codon of chromosome 8, the range of triplets is TTT (0.360%) and AGC (0.164%). The matrix frequency of chromosome 8, the frequency of tri-nucleotide sequence has been represented in AA-A,C,G; CA-A,C; GAA, ATT, CTT, GTT, TT-A,T of nucleotides (above 0.300%). Tri-nucleotide sequence of 0.250% to 0.299% has been represented in AAT, CAG, GA-C,G; TA-A,C,G; CC-G,T, GC-C,G; CGG, GGT, AT-A,C; CT-A,C; GT-A,C; TT-C,G. Tri-nucleotide sequence of 0.200% to 0.249% has been represented in CAT, GAT, TAT, AC-C,G,T; CC-A,C, GCT, TC-C,G,T; AG-G,T; CG-C,T; GG-A,G, TG-G,T; ATG, CTG and GTG. Tri-nucleotide sequence of 0.125% to 0.199% has been represented in ACA, GCA, TCA, AG-A,C; CGA, GGC, TGA and TGC.

## Chromosome 9

The chromosome 9 is total of 2439243 base pairs. The highest percentage value is Adenine residue (28.06%). The lowest percentage value is Guanine residue (22.06%). The Thymine and Cytosine residues are 27.81% and 22.07%. The highest combination of nucleotide residues of A+G

percentage is 50.12%. The ratio of A+T & G+C is 1.0. The highest triplets are AAA and TTT (0.346%). The lowest triplet codon is AGC (0.162%). The matrix frequency of chromosome 9, the frequency of tri-nucleotide sequence has been represented in AA-A,C; CA-A,C; CTT, GTT, TT-A,T (above 0.300%). Tri-nucleotide sequence of 0.250% to 0.299% has been represented in AA-G,T, CAG, GA-A,C,G; TA-A,C,G; CCT, GC-C,G; TCC, CGG, GG-A,T; AT-A,C,T; CT-A,C; GT-A,C; TT-C,G. Tri-nucleotide sequence of 0.200% to 0.249% has been represented in CAT, GAT, TAT, AC-C,G; CC-A,C; GC-C,G; TC-C,G; AG-A,G,T; CG-A,T; GGG, TG-A,G,T; ATG, CTG, GTG. Tri-nucleotide sequence of 0.150% to 0.199% has represented in ACA, GCA, TCA, AGC, CGA, GGC and TGC.

### Chromosome 10

The chromosome 10 is total of 306812 base pairs. The highest percentage value is Adenine residue (28.84%). The lowest percentage value is Cytosine residue (21.19%). The Thymine and Guanine residues are 28.73% and 21.24%. The highest combination of nucleotide residues of A+G percentage is 50.08%. The ratio of A+G & T+C content is 1.0. The highest triplet is AAA (0.360%). The lowest triplet codon is ACA (0.165%). The matrix frequency of chromosome 10, the frequency of tri-nucleotide sequence has been represented in AA-A,C,G; CA-A,C; GAA, ATT, CTT, TT-A,T (above 0.300%). Tri-nucleotide sequence of 0.250% to 0.299% has been represented in AAT, CAG, GA-C,G; TA-A,C,G; CCT, GCC, CGG, GG-A,T; AT-A,C; CT-A,C; GT-A,C,T; TT-C,G. Tri-nucleotide sequence of 0.200% to 0.249% has been represented in CAT, GAT, TAT, AC-C,G; CC-A,C,G; GC-G,T, TC-C,G; AG-G,T; CGC, GGG, TG-G,T; ATG, CTG, GTG. Tri-nucleotide sequence of 0.150% to 0.199% has been represented in AC-A,T; GCA, TC-A,T; AG-A,C, CG-A,T; GGC, TGA and TGC.

### Chromosome 11

The chromosome 11 is total of 298736 base pairs. The highest percentage value of Thymine residue is 28.66%. The lowest percentage value of Cytosine residue is 21.21%. The Adenine and Guanine residues are 28.50% and 21.63%. The highest combination of nucleotide residues of A+G percentage is 50.13%. The ratio of A+G & T+C content is 1.0. The highest triplet codon frequency (TTT) is 0.363%. The lowest triplet codon frequency AGC is 0.146%. The matrix frequency of chromosome 11, the frequency of tri-nucleotide sequence has been represented in AA-A,C,G; CA-A,C; GA-C,G, TAC, GTT, TT-A,C,T (above 0.300%). Tri-nucleotide sequence of 0.250% to 0.299% has been represented in AAT, CAG, GAA, TA-A,G; CC-G,T, CGG, GG-A,T; AT-

A,C,T; CT-A,C,T; GT-A,C; TTG. Tri-nucleotide sequence of 0.200% to 0.249% has been represented in CAT, GAT, TAT, AC-C,G; CC-A,C; GC-C,G; TC-C,G; AG-A,G,T; CG-A,T; GGG, TG-A,G,T; ATG, CTG, GTG. Tri-nucleotide sequence of 0.150% to 0.199% has been represented in AC-A,T; GC-A,T; TC-A,T; AGC, CGC, GGC and TGC.

### Chromosome 12

The chromosome 12 is total of 2229048 base pairs. The highest percentage value of Adenine residue is 28.54%. The lowest percentage value of Guanine residue is 21.60%. The Thymine and Cytosine residues are 28.21% and 21.65%. The highest combination of nucleotide residues of A+G percentage is 50.14%. The ratio of A+G & T+C content is 1.0. The highest triplet codon frequency of TTT is 0.353%. The lowest triplet codon frequency of AGC is 0.154%. The matrix frequency of chromosome 12, the frequency of tri-nucleotide sequence has been represented in AA-A,C,G; CA-A,C; GA-A,C; TAC, TT-A,T (above 0.300%). Tri-nucleotide sequence of 0.250% to 0.299% has been represented in AAT, CAG, GAG; TA-A,G; CC-G,T, GCC, CGG, GG-A,T; AT-A,C,T; CT-A,C,T; GT-A,C,T; TT-C,G. Tri-nucleotide sequence of 0.200% to 0.249% has been represented in CAT, GAT, TAT, AC-C,G; CC-A,C; GC-C,T; TC-C,G,T; AG-A,G,T; CG-A,T; GGG, TG-A,G,T; ATG, CTG, GTG. Tri-nucleotide sequence of 0.150% to 0.199% has represented in ACA, GCA, TCA, AGC, CGC, GGC and TGC.

### Conclusion

The new techniques just described identifying the rice chromosome and specifying the region using the first order Markov chain, second order Markov chain and CGR methods of the DNA sequence analysis. The probabilities defining these models can be calculated directly and easily from the raw DNA sequences, implying that the CGR gives no further insight into the structure of the DNA sequence than is given by the dinucleotide and trinucleotide frequencies. In this paper, we have shown that simple Markov Chain models based solely on dinucleotide and trinucleotide frequencies can account for the complex patterns exhibited in CGR of *Oryza sativa* (japonica cultivar-group) chromosome sequences. The *Oryza sativa* (japonica cultivar-group) species chromosome sequences are more similar to each other. However, our analysis is visible of sequence pattern is similar in each other. Some high matrix frequency value (0.300) of tri-nucleotide codon is having by small number of trinucleotides, but the matrix values are different in each other. The low-resolution codon frequencies are having by small number of tri-nucleotides (0.125-0.199), but the matrix value

is different. We observed the above results the low-resolution tri-nucleotides are very low. The chromosome 1 is having four high frequency (above 0.300%) tri-nucleotides. The chromosome 8 has been representing in same nucleotide content ratio in Guanine and Cytosine residues (21.57%). The frequency matrix values 0.200% to 0.299% are highly responsible for the *Oryza sativa* (japonica cultivar-group) species. It is representing more number of trinucleotide codons. Finally, we observed the frequency of start codon is equal to average of two stop codon frequencies. The *in silico* analysis of the matrix frequency study is used in *invitro/invivo* studies for reassembling the particular repaired codon region which is modified by the gene tinkering (codon replacing) methods. The future analysis can integrate these procedures into one logical, individual gene.

## References

- Almeida JS, Carrico JA, Marezek A, Noble PA, Fletcher M (2001) Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics* 17: 429-437. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
- Deschavanne PJ, Giron A, Villain J, Fagot G, Fertil B (1999) Genomic Signature: Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences. *Mol Biol Evol* 16: 1391-1399. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
- Fukui K (1985) Identification of plant chromosome by the image analysis method. *The Cell* 17: 145-149.
- Fuentes JL, Cornide MT, Alvarez A, Suarez E, Borges E (2005) Genetic diversity analysis of rice varieties (*Oryza sativa* L.) based on morphological, pedigree and DNA polymorphism data, *Plant Genetic Resources: Characterization and Utilization*. *Plant Genet Resour* 3: 353-359.
- Gao LZ, Zhang CH, Chang LP, Jia JZ, Qiu ZE et al. (2005) Microsatellite diversity within *Oryza sativa* with emphasis on indica-japonica divergence. *Genet Res* 85: 1-14. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
- Goff SA, Ricke D, Lan TH, Presting G, Wang R, et al. (2002) A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. japonica). *Science* 296: 92-100. » [CrossRef](#) » [Google Scholar](#)
- Goldman N (1993) Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representation of DNA sequences. *Nucleic Acids Res* 21: 2487-2491. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
- Iijima K, Fukui K (1991) Clarification of the conditions for the image analysis of plant chromosomes. *Bull Natl Inst Agrobiol Resour* 6: 1-58.» [CrossRef](#) » [Google Scholar](#)
- Jeffrey HJ (1990) Chaos game representation of gene structure. *Nucleic Acids Res* 18: 2163-2170.» [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
- Jeffrey HJ (1992) Chaos game visualization of sequences. *Computer Graphics* 16: 25-34. » [CrossRef](#) » [Google Scholar](#)
- Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* 11: 283-290. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
- Mohanty AK, Narayana Rao AV (2000) Factorial Moments Analyses Show a Characteristic Length Scale in DNA Sequences. *Phys Rev Lett* 84: 1832-1835. » [Pubmed](#) » [Google Scholar](#)
- Nussinov R (1980) Some rules in the ordering of nucleotides in the DNA. *Nucleic Acids Res* 8: 4545-4562. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
- Nussinov R (1981) Nearest neighbor nucleotide patterns: Structural and biological implications. *J Biol Chem* 256: 8458-8462. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
- Yoshiaki NA, Baltazar A, Hisataka N, Ikuo H, Manabu A, et al. (2003) A Comprehensive Homology Search for Rice Specific Sequences. *Genome Inform* 14: 533-534.