World Applied Sciences Journal 11 (3): 315-320, 2010
ISSN 1818-4952
© IDOSI Publications, 2010

Matrix Frequency Analysis of Genome Sequences: Pattern Identification of *Turfgrass* Species

[1]*K. Manikandakumar,* [2]*K. Gokulraj,* [3]*R. Srikumar and* [4]*S. Muthukumaran*

[1]Department of Physics, Bharathidasan University College (W),
Orathanadu, 614 625, Tanjavore, Tamil Nadu, India
[2]Department of Computer Science,
Jamal Mohamed College, Tiruchirappalli, 620020, Tamil Nadu, India
[3]Department of Microbiology, Bharathidasan University College (W),
Orathanadu, 614 625, Tanjavore, Tamil Nadu, India
[4]Department of Physics, The H.H. Raja's College, Pudukkottai, 622 001, Tamil Nadu, India

Abstract: The complete genome analysis, which is one of the essential steps to know their characteristics, is very important. 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. Complete genome analysis is depending on matrix frequency of sequence residue calculation. In this study, we select monocot species as the specimen for genome sequence analysis. 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 some monocot species. By our research, we identified the codon pattern from *Turfgrass* species among the other monocot species.

Key words: Markov chain · *Turfgrass* · Genome sequence · matrix frequency

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 neighbour bases could help in the description of nucleotide sequences [1]. However, [1] has followed the scale independence of genetic sequence method to investigate local and global homology [2]. The two patterns identified 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 neighbours over equivalent dinucleotide [3, 4]. In another work it was analyzed the parity ratio of protein sequences based on Chargaff's rule [5].

Mathematical characterization of DNA sequences could help in the understanding of structural relationships among different whole genomes along the chromosomes. The degenerated 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 [6]. The genomic signature as expressed in terms of short nucleotide usage extends and generalizes the genomic signature and it takes advantages of whole genome data reveals genome wide trends [7]. In two other works [8 (a), (b)] we analyzed the matrix frequency analysis of *Oryza sativa* (japonica cultivar - group) and *Homo sapiens* complete genomes. The 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 genome sequences of some monocot species such as pineapple, banana, barley, duckweed, *gladiolus*, onion and *turfgrass*. We have generated a matrix frequency of the above genomes.

Corresponding Author: K. Manikandakumar, Department of Physics, Bharathidasan University College (W), Orathanadu, 614 625, Tanjavore, Tamil Nadu, India. E-mail: bioinfokm@gmail.com

MATERIAL AND METHODS

A simple model, which permits the simulation of these features of nucleotide residues, is discrete time Markov chain [9]. In this model, a 4 × 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^{th} element of the i^{th} row of *P* which defines the probability that base *j* follows base *i*. The row sum of *P* must equal 1. Using this matrix a simulated nucleotide residue are 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 can be used to select the next and the coming bases of the consecutive simulated sequences to the same length of the original nucleotide residues.

In first-order Markov chain model, the successive bases of a residue depend only on the preceding base. The probabilities in the matrix *P*, is estimated by direct calculation of the residues of 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}) of the some monocot species such as pineapple, banana, barley, duckweed, *gladiolus*, onion and *turfgrass* genomes, are given in Table 1.

The first-order Markov chain model successfully recreates other genomes. The lack of banding suggests approximate equality of the frequencies of the bases A, T, G, and C, confirmed by direct calculation from the residues. The first-order Markov chain model will not give

Table 1: Matrix frequency of Doublet Codon

Name of the	Pineapple	Banana	Barley	Duckweed		Onion	
Doublet Codon	(*Ananas comosus*)	(*Musa sapientum*)	*(Hordeum vulgare)*	(*Lemnaceae*)	*Gladiolus*	*(Allium cepa)*	*Turfgrass*
AA	0.349	0.317	0.360	0.347	0.313	0.339	0.276
AC	0.182	0.173	0.164	0.155	0.188	0.186	0.273
AG	0.193	0.213	0.188	0.191	0.228	0.197	0.239
AT	0.276	0.297	0.288	0.307	0.271	0.278	0.212
CA	0.325	0.323	0.350	0.291	0.336	0.318	0.254
CC	0.223	0.224	0.190	0.232	0.221	0.227	0.252
CG	0.196	0.173	0.131	0.168	0.179	0.200	0.271
CT	0.255	0.280	0.329	0.309	0.264	0.255	0.222
GA	0.264	0.334	0.354	0.345	0.319	0.262	0.244
GC	0.270	0.195	0.168	0.176	0.201	0.270	0.298
GG	0.225	0.230	0.190	0.237	0.242	0.228	0.276
GT	0.242	0.242	0.288	0.241	0.238	0.241	0.182
TA	0.212	0.248	0.239	0.253	0.217	0.214	0.187
TC	0.197	0.220	0.201	0.213	0.208	0.202	0.306
TG	0.244	0.214	0.201	0.192	0.253	0.244	0.297
TT	0.348	0.318	0.359	0.342	0.322	0.339	0.210

Table 2: Matrix frequency of Triplet Codon

Name of the	Pineapple	Banana	Barley	Duckweed		Onion	
Triplet Codon	(*Ananas comosus*)	(*Musa sapientum*)	*(Hordeum vulgare)*	(*Lemnaceae*)	*Gladiolus*	*(Allium cepa)*	*Turfgrass*
AAA	0.377	0.317	0.379	0.385	0.318	0.368	0.256
AAC	0.349	0.340	0.370	0.279	0.366	0.343	0.252
AAG	0.260	0.350	0.386	0.363	0.346	0.262	0.293
AAT	0.249	0.279	0.275	0.277	0.231	0.253	0.214
CAA	0.319	0.343	0.372	0.350	0.338	0.312	0.291
CAC	0.359	0.340	0.382	0.311	0.304	0.347	0.263
CAG	0.301	0.306	0.350	0.333	0.282	0.293	0.218
CAT	0.173	0.221	0.220	0.254	0.212	0.175	0.162
GAA	0.342	0.353	0.349	0.357	0.331	0.330	0.287

Table 2: Continued

Name of the	Pineapple	Banana	Barley	Duckweed		Onion	
Triplet Codon	(*Ananas comosus*)	(*Musa sapientum*)	*(Hordeum vulgare)*	(*Lemnaceae*)	*Gladiolus*	*(Allium cepa)*	*Turfgrass*
GAC	0.318	0.321	0.341	0.304	0.344	0.307	0.230
GAG	0.285	0.338	0.355	0.356	0.341	0.279	0.222
GAT	0.203	0.247	0.239	0.288	0.223	0.208	0.209
TAA	0.341	0.260	0.328	0.282	0.257	0.332	0.267
TAC	0.282	0.297	0.322	0.280	0.327	0.279	0.275
TAG	0.231	0.328	0.326	0.325	0.296	0.230	0.258
TAT	0.208	0.234	0.219	0.215	0.203	0.209	0.164
ACA	0.172	0.171	0.171	0.142	0.184	0.174	0.267
ACC	0.196	0.237	0.206	0.241	0.233	0.204	0.298
ACG	0.275	0.197	0.181	0.183	0.203	0.275	0.286
ACT	0.184	0.198	0.200	0.205	0.188	0.189	0.339
CCA	0.207	0.173	0.168	0.155	0.187	0.206	0.232
CCC	0.214	0.204	0.168	0.220	0.238	0.218	0.230
CCG	0.257	0.214	0.157	0.178	0.218	0.257	0.288
CCT	0.226	0.245	0.234	0.219	0.244	0.230	0.279
GCA	0.179	0.147	0.146	0.146	0.168	0.188	0.241
GCC	0.236	0.215	0.185	0.220	0.211	0.241	0.255
GCG	0.285	0.200	0.164	0.162	0.188	0.286	0.359
GCT	0.192	0.209	0.178	0.188	0.184	0.204	0.283
TCA	0.175	0.199	0.164	0.180	0.220	0.181	0.417
TCC	0.241	0.233	0.189	0.240	0.204	0.242	0.227
TCG	0.262	0.179	0.161	0.179	0.201	0.262	0.248
TCT	0.193	0.232	0.195	0.227	0.220	0.197	0.335
AGA	0.177	0.238	0.194	0.189	0.236	0.181	0.327
AGC	0.175	0.153	0.127	0.168	0.133	0.180	0.246
AGG	0.201	0.221	0.175	0.222	0.234	0.205	0.279
AGT	0.220	0.219	0.204	0.190	0.265	0.222	0.294
CGA	0.224	0.204	0.165	0.188	0.224	0.227	0.225
CGC	0.199	0.180	0.145	0.185	0.203	0.206	0.256
CGG	0.227	0.241	0.212	0.238	0.234	0.234	0.293
CGT	0.284	0.199	0.176	0.195	0.230	0.282	0.330
GGA	0.222	0.212	0.219	0.208	0.242	0.224	0.218
GGC	0.187	0.193	0.121	0.169	0.194	0.191	0.288
GGG	0.215	0.215	0.168	0.231	0.203	0.217	0.222
GGT	0.277	0.230	0.208	0.211	0.269	0.271	0.288
TGA	0.157	0.187	0.174	0.181	0.202	0.160	0.147
TGC	0.225	0.171	0.131	0.155	0.192	0.223	0.293
TGG	0.248	0.244	0.209	0.258	0.283	0.249	0.317
TGT	0.223	0.210	0.208	0.184	0.244	0.226	0.265
ATA	0.274	0.274	0.256	0.283	0.262	0.276	0.150
ATC	0.280	0.270	0.296	0.312	0.269	0.273	0.204
ATG	0.264	0.231	0.259	0.232	0.218	0.258	0.143
ATT	0.347	0.304	0.320	0.328	0.315	0.336	0.154
CTA	0.250	0.279	0.294	0.307	0.251	0.254	0.253
CTC	0.228	0.275	0.305	0.283	0.254	0.230	0.251
CTG	0.215	0.239	0.281	0.250	0.265	0.216	0.202
CTT	0.316	0.335	0.370	0.332	0.314	0.313	0.230
GTA	0.258	0.287	0.287	0.288	0.260	0.258	0.254
GTC	0.259	0.271	0.353	0.308	0.251	0.260	0.227
GTG	0.214	0.247	0.313	0.252	0.268	0.217	0.197
GTT	0.327	0.313	0.375	0.313	0.324	0.317	0.221
TTA	0.327	0.354	0.334	0.356	0.321	0.327	0.169
TTC	0.252	0.298	0.358	0.325	0.277	0.255	0.205
TTG	0.260	0.250	0.305	0.237	0.220	0.259	0.177
TTT	0.376	0.323	0.378	0.374	0.333	0.368	0.235

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 digroup XY, is estimated directly from the nucleotide residues trigroup frequencies n_{XYZ} using the formula $P_{XYZ} = n_{XYZ} / (n_{XYA} + n_{XYT} + n_{XYG} + n_{XYC})$. Dinucleotide frequencies (n_{XY}) and Markov Chain probabilities (P_{XY}) for the monocot species genomes, are given in Table 2.

RESULTS AND DISCUSSIONS

The structure of DNA is specific to each species and undergoes only slight variations along the whole genome. Diversity among species is considerable and is primarily a consequence of base concentration, stretches of bases with unusual frequencies. The frequencies of occurrence, allows one to point out the basis of the genome. In our analysis, the matrix frequency calculation of some sequences of monocot plant genome is presented. We analysed each species genome sequences. Table 1 has shown by the first order Markov chain matrix frequency of genome sequences and it is representing in dinucleotide codons. The Table 2 is shown by the second order Markov chain matrix frequency of genome sequences and it is representing in trinucleotide codons.

Matrix Frequency of the First Order Markov Chain Model: We have generated and analysed the first order Markov chain matrix frequency for 16 (4 × 4) nucleotide doublet codons for monocot plant species such as pineapple, banana, barley, duckweed, *gladiolus*, onion and *turfgrass* genome sequences. The matrix frequency for each doublet codon is given in Table 1. The duplets of monocot genome sequence frequency is shown in Table 1, among the above species analysis, the *turfgrass* species doublet codons are identified, it has having some pattern from among other species. i.e., the doublet codon frequencies are either high or low for comparing other species genome frequencies. For example, the AC doublet codon frequency of *turfgrass* species is 0.273. But, the other species doublet codon are having the frequency is < 0.200. In the same way, the GT doublet codon frequency of *turfgrass* species is 0.182. The other species doublet frequency is having > 0.235. The TT doublet codon frequency of *turfgrass* is 0.210. The other species doublet codon frequency is > 0.310. Therefore the *turfgrass* species is having the pattern comparing the other monocot species.

Matrix Frequency of the Second Order Markov Chain Model: We have generated and analysed the second order Markov chain matrix frequency for 64 (4 × 4 × 4) nucleotide triplet codons for monocot species complete genome sequences and the matrix frequency for each tripet codon is given in Table 2. From Table 2, it results that the AAA triplet codon of *turfgrass* species have the frequency of 0.256, but for the other monocot species the frequency is > 0.310. The TAT triplet codon of *turfgrass* species is 0.164. The monocot speices triplet frequency is > 0.200. The ACA triplet codon of *turfgrass* speices is having the frequency is 0.267. The other monocot species is having the triplet frequency is <0.190. The ACT triplet codon frequency of *turfgrass* species is 0.339. The other monocot species have the triplet frequency is < 0.210. In the above analysis, we have identified the *turfgrass* species as having the pattern comparing between the other monocot species.

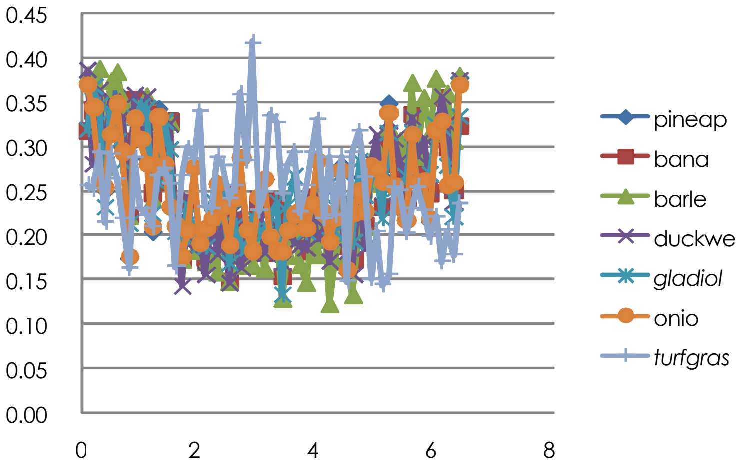

Fig. 1: Graphical Representation of *Turfgrass*

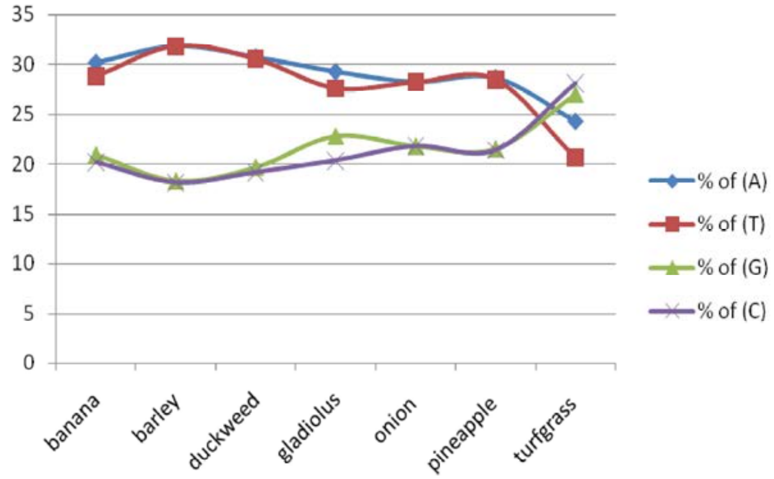

Fig. 2: Geometrical Representation of *Turfgrass*

Table 3: Nucleotide content analysis of different monocot species

Sl. No.	Name of the species	Total No. of Residues	% of (A)	% of (T)	% of (G)	% of (C)	% of (A + G)	% of (T + C)	% of (A+G) / (T+C)
1.	Banana	485854	30.18	28.81	20.87	20.15	51.05	48.96	1.042687908
2.	Barley	14619781	31.83	31.83	18.24	18.1	50.07	49.93	1.002803925
3.	Duckweed	616440	30.72	30.52	19.61	19.15	50.33	49.67	1.013287699
4.	*Gladiolus*	93219	29.27	27.61	22.79	20.33	52.06	47.94	1.085940759
5.	Onion	14598671	28.24	28.21	21.76	21.79	50	50	1
6.	Pineapple	9950361	28.65	28.47	21.49	21.38	50.14	49.85	1.005817452
7.	*Turfgrass*	234227	24.28	20.63	27	28.09	51.28	48.72	1.052545156

Pattern Identification from Graphical Representation of *Turfgrass*: We identified the pattern from graphical representation of monocot species (Fig. 1 and Fig. 2). This figure shows, the *turfgrass* species having the different pattern from other monocot species. The other monocot species have the same line, but the *turfgrass* species only got the different line either high or low from the other species (Fig-1). From Fig-2, the *turfgrass* species has the different position from other species. i.e., the *turfgrass* species nucleotides of A, T, G and C are goes to opposite direction of other species nucleotides.

From Table 3, we identified the different pattern of *turfgrass* species from other species like the nucleotides contents. The other monocot species are having the A and T nucleotide contents are high compared from *turfgrass* species, G and C nucleotide contents are low compared from *turfgrass* species. The A and T nucleotide content of *turfgrass* species having the ratio is low from other monocot species. The G and C nucleotide contents of *turfgrass* species is having the ratio is high from the monocot species. i.e., the other monocot species are having the A+T and G+C nucleotide contents are, the banana having A+T is 58.99% and G+C having 41.01%. The barley having A+T is 63.66% and G+C having 36.34%, the duckweed having A+T is 61.23% and G+C having 38.77%. A+T is 58.99% and G+C having 41.01%, the *gladiolus* having A+T is 56.88% and G+C having 43.12%, the onion having A+T is 56.45% and G+C having 43.55% and the pineapple having A+T is 57.12% and G+C having 42.88%. But, the *turfgrass* species having the A+T is 44.91% and G+C having 55.09%.

CONCLUSION

Using these new techniques, the first order Markov chain, second order Markov chain of the genome sequence analysis, we can easily identify the pattern and nature of the *turfgrass* genome sequences. The probabilities defining these models can be calculated directly and easily from the raw DNA sequences, implying that 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 some monocot species genome sequences such as pineapple, banana, barley, onion, duckweed, *gladiolus* and *turfgrass*. However, our analysis is identified, the *turfgrass* species as having the some pattern as compared to the above mentioned monocot species. This result is confirmed by two analyses such as matrix frequency of first order Markov chain and second order Markov chain methods. The future analysis can integrate these procedures into one logical, individual gene and consequently into protein sequences.

REFERENCES

1. Mohanty, A.K. and A.V.S.S. Narayana Rao, 2000. Factorial Moments Analyses Show a Characteristic Length Scale in DNA Sequences. Phys. Rev. Lett., 84: 1832-1835.
2. Almeida, J.S., J.A. Carrico, A. Maretzek, P.A. Noble and M. Fletcher, 2001. Analysis of genomic sequences by Chaos Game Representation. Bioinformatics, 17: 429-437.
3. Nussinov, R., 1980. Some rules in the ordering of nucleotides in the DNA. Nucleic Acids Res., 8: 4545-4562.
4. Nussinov, R., 1981. Nearest neighbor nucleotide patterns: Structural and biological implications. J. Biol. Chemistry, 256: 8458-8462.
5. Manikandakumar, K., S. Muthu Kumaran and R. Srikumar, 2010. Analysis of parity ratio of protein sequences: A new approach based on Chargaff's rule, Romanian J. Biophysics, 20: 183-191.
6. Deschavanne, P.J., A. Giron, J. Villain G. Fagot and B. Fertil, 1999. Genomic Signature: Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences. Mol. Biol. Evol., 16: 1391-1399.
7. Karlin, S. and C. Burge, 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet, 11: 283-290.
8. (a) Manikandakumar, K., S. Muthukumaran and R. Srikumar, 2009. Matrix Frequency Analysis of *Oryza sativa* (japonica cultivar-group) Complete Genomes. Journal of Computer Science and Systems Biol., 2: 159-166.

   (b) Manikandakumar, K., S. Muthukumaran, R. Srikumar, K. Gokulraj and S. Santhosh Baboo, 2009. Analysis of *Homo sapiens* (Human) Chromosomes Complete Genome Using Matrix Frequency. *nst* Life Sci. and Bioinformatics, 1: 57-66.
9. Goldman, N., 1993. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representation of DNA sequences. Nucleic Acids Res., 21: 2487-2491.