Application of Chaos Game in Tri-Nucleotide Representation for the Comparison of Coding Sequences of β-Globin Gene

Subhram Das, Nobhonil Roy Choudhury, D.N. Tibarewala and D.K. Bhattacharya

Abstract In this paper, we use 2D tri-nucleotide representation based on chaos game theory. We extend the representation from 2D to 3D by taking the third coordinate as the multiple of the first two ones. Complete coding sequences of β globin genes of 10 species are now compared using four types of descriptors—1. Mean of the components of the represented sequences, 2. Standard deviation of the components of the represented sequence, 3. Highest eigen value of M/M matrix and 4. Highest eigen value of J/J matrix. The results in the four cases are critically examined. It is found that the use of J/J matrix with highest eigen value as the descriptor is the best one among the others.

Keywords Chaos game · 2D tri-nucleotide representation · J/J matrix · Highest eigen value

1 Introduction

In bioinformatics, the basic studying strategy for both DNA and protein sequences is to make proper comparisons of both. There are mainly two types of comparison methods—one is based on alignment technique and the other one is based on alignment-free technique. The later one is preferred, as it is less time consuming. Anyway it is mostly based on mononucleotide representation. One such graphical

S. Das (✉) · N.R. Choudhury
Computer Science & Engineering, Narula Institute of Technology, Kolkata, India
e-mail: subhram@gmail.com

N.R. Choudhury
e-mail: nobhonil30390@gmail.com

D.N. Tibarewala · D.K. Bhattacharya
Bio-Science & Engineering, Jadavpur University, Kolkata, India
e-mail: biomed.ju@gmail.com

D.K. Bhattacharya
e-mail: dkb_math@yahoo.com

S. Bhattacharyya et al. (eds.), *Industry Interactive Innovations in Science, Engineering and Technology*, Lecture Notes in Networks and Systems 11,
DOI 10.1007/978-981-10-3953-9_54

representation is first given by Hamori and Ruskin in 1983 [1]. Graphical representations are found to vary from 2D to 6D. However, directly working with mononucleotides (A, G, C & T) leads to a lot of information loss. So Di- and Tri-nucleotide representations were thought of. The mononucleotide models cannot represent the Di- and Tri- nucleotides without complex calculations [2]. So such representations were found out independently [3–15]. However, the following limitations still remain—1. For 3D representation, the represented values are only 64 in number. Naturally the mean value of such represented coordinates is not of much interest. Even if the cumulative values give much variation, still the use of mean value is not a very satisfactory descriptor. So the final comparison based on means of two types of represented points (normal and cumulative) may not be applicable for comparison of a larger variety of samples [16]. 2. Standard deviation shows the spread of the data rather than determining a theoretical centre, and the cumulative components reduce redundancy. But even the calculations of comparisons based on standard deviation as the descriptor on the cumulative data set is also found to be non-satisfactory [17]. 3. There is always a risk in taking cumulative values, as the resulting time series becomes stochastic. 4. Numerical values and the signs used for tri-nucleotide representation [16, 17] appear to be very much artificial.

In order to avoid these difficulties we have made a very simple approach. We take only the 64 different values obtained by Chaos game representation [18] as the 2D representation of tri-nucleotides and make the representation 3D by taking the third coordinate as the multiple of the first two. As the represented values are now different from those obtained by earlier methods, so to check the improvement in the results, we choose sequentially the descriptors as mean, standard deviation, highest eigen values of M/M matrix and J/J matrix. Compared to the traditional matrix, the J/J matrix can investigate the composition, distribution and chemical properties of bases; it can also picture the biological significance of the sequence [19]. So we try for both M/M and J/J matrix with the expectation that J/J might give better results.

What makes our representation better than the previous ones [1–17] based on tri-nucleotides is that it is a much easier method and hence more efficient. The essential difference lies in getting the 2D tri-nucleotide representations with the help of chaos game theory. The 2D coordinates are obtained in a very natural way using chaos game with only the initial values of the nucleotides. It is known that the exon of β globin genes of different species is essential for pharmaceutical purposes. So we have preferred choosing coding sequences of β globin genes for the purpose of sequence comparison.

2 Methodology

We use the chaos game values shown in the graphical representation of [18] on the non-overlapping triplets of the given DNA sequence for the calculation of different statistical parameters to be used in the analysis of the paper.

Each nucleic acid triplet consists of three coordinates (x, y, z), the first two are obtained from the above chaos game representation, the third one being generated by the multiplication of the first two coordinates. Let N = M/3 be the number of codons in the sequence, where M is the length of the DNA sequence.

Let $x = (x_1, x_2, x_3, \ldots, x_N)$, $y = (y_1, y_2, y_3, \ldots, y_N)$, $z = (z_1, z_2, z_3, \ldots, z_N)$ be the 3D represented points.

Then mean of x, y, z is given by μ_x, μ_y and μ_z respectively, where $\mu_x = \sum_{i=1}^{N} x_i/N$, $\mu_y = \sum_{i=1}^{N} y_i/N$, $\mu_z = \sum_{i=1}^{N} z_i/N$

Then standard deviation of x, y, z is given by V_x, V_y and V_z respectively, where

$$V_x = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu_x)^2}{N}},\ V_y = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \mu_y)^2}{N}},\ V_z = \sqrt{\frac{\sum_{i=1}^{N} (z_i - \mu_z)^2}{N}}$$

The M/M matrix is calculated as

$$M_{i,j} = \frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}}{|x_i - x_j| + |y_i - y_j| + |z_i - z_j|}$$

The J/J matrix is calculated as

$$J_{i,j} = \frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 +}}{|x_i - x_j| + |y_i - y_j|} + \frac{\sqrt{(z_i - z_j)^2}}{|z_i + z_j|}$$

Table 1 The complete coding sequences of β globin genes of 10 species

Species	NCBI ID	Location of each exon	Length of CDS
Human	U01317	19541..19632, 19755..19977, 20833..20961	444
Duck	X15739	291..382, 495..717, 1742..1870	444
Opossum	J03643	467..558, 672..894, 2360..2488	444
Gallus	V00409	465..556, 649..871, 1682..1810	444
Mouse	V00722	275..367, 484..705, 1334..1462	444
Rabbit	V00882	277..368, 495..717, 1291..1419	444
Rat	X06701	310..401, 517..739, 1377..1505	444
Tufted monkey	AY279115	946..1037, 1168..1390, 2218..2346	444
Woolly monkey	AY279114	952..1043, 1174..1396, 2227..2355	444
Hare	Y00347	1485..1576, 1703..1925, 2492..2620	444

Table 2 Distance matrix for the coding sequences of β globin genes of 10 species based on highest eigen value of J/J Matrix

Species	Duck	Gallus	Hare	Human	Mouse	Rabbit	Rat	Tufted monkey	Woolly monkey	Opossum
Duck	0	0.5421	4.532	1.481	2.034	4.781	2.115	3.345	2.768	0.587
Gallus		0	3.990	0.939	1.492	4.239	1.573	2.803	2.226	1.129
Hare			0	3.050	2.498	0.248	2.416	1.187	1.764	5.119
Human				0	0.552	3.299	0.633	1.863	1.286	2.068
Mouse					0	2.746	0.081	1.310	0.733	2.621
Rabbit						0	2.665	1.435	2.012	5.368
Rat							0	1.229	0.652	2.702
Tufted monkey								0	0.577	3.932
Woolly monkey									0	3.355
Opposum										0

We calculate the similarity/dissimilarity between the coding sequences based on the distance matrix measured by

(1) Euclidean distances between three component vectors (μ_x, μ_y, μ_z) of pair of sequences. (2) Euclidean distances between the three component vectors (V_x, V_y, V_z) of pair of sequences. (3) Distance measured by modulus of the difference of highest eigen values of the M/M matrix. (4) Distance measured by modulus of the difference of highest eigen values of the J/J matrix.

The smaller the entry in the distance matrix is, more similar the DNA sequences are. Therefore, we can say that the distances between evolutionary closely related species are smaller, while those between evolutionary distant species are larger. We draw the phylogenetic tree based on similarity/dissimilarity matrix using UPGMA in MEGA4 software [20].

3 Result and Discussion

Table 1 shows the information regarding corresponding sequences of 10 different species and Table 2 shows the distance matrix of the complete coding sequences of β globin genes of 10 different species based on highest eigen value of J/J matrix. Distance matrix using Euclidian distance and corresponding Phylogenic trees are also obtained similarly in other three cases. We observe that the phylogenetic tree Fig. 1 using their highest eigen value of J/J matrix generates the best result among others. From Fig. 1 we also observe that the more similar species pairs are like Mouse—Rat, Tufted Monkey—Woolly Monkey, Hare—Rabbit, Gallus—Duck are come closer to each others. Our phylogenetic tree agrees with that found in [16] for the species taken in common.

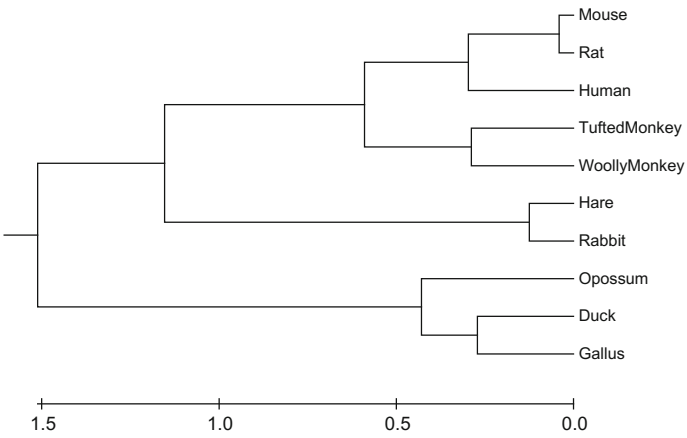

Fig. 1 Phylogenetic tree of 10 different species based on their complete coding sequence of β globin genes using their highest eigen value of J/J matrix

4 Conclusion

In this paper, we propose a new method of nucleotide representation using chaos game theory. By comparing four descriptor, we conclude that highest eigen value for J/J matrix is the best among the four descriptors: mean, standard deviation, highest eigen value for M/M matrix and highest eigen value for J/J matrix for comparison of complete coding sequences of β globin genes for the above 10 species. We therefore conclude that our method is effective for evaluating sequence similarities on an intuitive basis. However, our method is experimented on only 10 different sequences; in near future we like to apply our method on large numbers of species.

References

1. Hamori, E., Ruskin, J.: H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. J. Biol. Chem. **258**, 1318–1327 (1983)
2. Guo, F.B., Ou, H.Y., Zhang, C.T.: ZCURVE: a new system for recognizing protein–coding genes in bacterial and archaeal genomes. Nucl. Acids. Res. **31**, 1780–1789 (2003)
3. Zhang, C.T., Zhang, R.: Analysis of distribution of bases in the coding sequences by a diagrammatic technique. Nucl. Acids Res. **19**, 6313–6317 (1991)
4. Zhang, R., Zhang, C.T.: Z curves, an intuitive tool for visualizing and analyzing the DNA sequences. J. Biomol. Struct. Dyn. **11**, 767–782 (1994)
5. Nandy, A.: A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. Curr. Sci. **66**, 309–314 (1994)
6. Randic, M., Vracko, M., Lers, N., Plavsic, D.: Novel 2–D graphical representation of DNA sequences and their numerical characterization. Chem. Phys. Lett. **368**, 1–6 (2003)
7. Randic, M., Vracko, M., Zupan, J., Novic, M.: Compact 2–D graphical representation of DNA. Chem. Phys. Lett. **373**, 558–562 (2003)
8. Liao, B., Wang, T.M.: Analysis of similarity/dissimilarity of DNA sequences based on 3–D graphical representation. Chem. Phys. Lett. **388**, 195–200 (2004)
9. Randic, M.: Graphical representations of DNA as 2–D map. Chem. Phys. Lett. **386**, 468–471 (2004)
10. Liao, B., Wang, T.M.: 3–D graphical representation of DNA sequences and their numerical characterization. J. Mol. Struct. (Theochem) **681**, 209–212 (2004)
11. Chi, R., Ding, K.Q.: Novel 4D numerical representation of DNA sequences. Chem. Phys. Lett. **407**, 63–67 (2005)
12. Yao, Y.H., Nan, X.Y., Wang, T.M.: A new 2D graphical representation—Classification curve and the analysis of similarity/dissimilarity of DNA sequences. J. Mol. Struct. (Theochem) **764**, 101–108 (2006)
13. Liao, B., Ding, K.Q.: A 3D graphical representation of DNA sequences and its application. Theor. Comput. Sci. **358**, 56–64 (2006)
14. Song, J., Tang, H.W.: A new 2–D graphical representation of DNA sequences and their numerical characterization. J. Biochem. Biophys. Methods **63**, 228–239 (2005)
15. Zhang, Z.J.: DV–Curve: a novel intuitive tool for visualizing and analyzing DNA sequences. Bioinformatics, vol. 25, pp. 1112–1117 (2009)
16. Yu, J., Wang, J., Sun, X.: Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation. MATCH Commun. Math. Comput. Chem. **63**, 493-512 (2010)

17. Das, S., Palit, S., Mahalanabish, A.R., Choudhury, N.R.: A new way to find similarity/dissimilarity of DNA sequences on the basis of dinucleotides representation. In: Computational Advancement in Communication Circuits and System, pp. 151–160. Springer (2015)
18. Randic, M., Zupan, J., Balaban, A.T.: Unique graphical representation of protein sequences based on nucleotide triplet codons. Chem. Phys. Lett. **397**, 247–252 (2004)
19. Luo, J., Guo, J., Li, Y.: A new graphical representation and its application in similarity/dissimilarity analysis of DNA sequences. In: 4th International Conference on Bioinformatics and Biomedical Engineering (2010). doi:10.1109/ICBBE.2010.5515203
20. Kumar, S., Nei, M., Dudley, J., Tamura, K.: MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief. Bioinform. **9**, 299–306 (2008)