Analysis of DNA sequences similarity based on a new 3-D graphical representation method

Kshatrapal SINGH[1], Ashish KUMAR[1], Manoj Kumar GUPTA[2]

[1] Department of CSE, ITS Engineering College, Greater Noida, India

mekpsingh1@gmail.com, ashishcse29@gmail.com

[2] School of CSE, SMVD University, Katra (J&K), India

manoj.cst@gmail.com

Abstract: Finding the density of specific nucleotides is an important task in DNA sequence analysis. The technique, which is based on graphical representations of DNA sequences, allows for comparison, testing, and storage of different sequences. We create three typical curves in this research based on the categorization of four base DNA sequences. The DNA sequence fluctuation and geometrical centers of three curves organized into 12 component vectors were represented. The Euclidean distance between generated vectors is used to study and compare coding sequences for ten different species.

Keywords: Similarity analysis, graphical representation, geometrical centers, numerical characterization.

1. Introduction

DNA sequence basically represents strings of letters: A, C, T, G which specify the 4 nucleic acids: Adenine, Cytosine, Thymine as well as Guanine. In recent years various researchers pursued graphical representations of DNA sequence in their researches for identification of, similarities/dissimilarities and observation of such sequences. It has been well proved that this approach of analysis of a DNA sequences can lead to numerical characterization of the sequence (Jiyuan et al., 2015), (Kwon, 2015), (Nakamura et al., 2013).

There will be some gaps in both approaches; the 2-dimensional as well as the 3-dimensional representation because of the overlapping of DNA curves themselves. It is not feasible to reconstruct DNA sequences from such representations. This specific limitation has newly been corrected by process of adjustment of directions of various vectors assigned to 4 bases, in the case of 2-D graphical representation. However, there is no analogous limitation in the method of 4-D representations because there is no overlapping or intersecting of the curves. But the main feature of graphical visualization of DNA sequence lax here (Daa Young et al., 2016), (Young Lee et al., 2018). Randic (Randic, 2017) discovered the 2-dimensional graphical representation of DNA sequence, for which 4 strings of letters A, C, T and G are assign to 4 horizontal lines, which also keep off the drawback.

Recently, on the basis of chemical structure of bases 3-D graphical representation methods are introduced. In these methods some characteristic curves are drawn from DNA primary sequence. Coarse grained explanation of DNA main sequence is given by characteristic curves which indicate scattering of the base pair, and loss of information can be avoided by 3-D approach in which DNA curves intersect and overlap with themselves (Hwaangue Choo Dayoung & Kwon, 2016), (Kinetic, 2008), (Krzywinski et al., 2009).

2. Proposed methodology

While comparing DNA sequence we have to examine not only the structures of strings but also chemical structure also. IUB code table for the bases is as follows (Randic, 2017) (Table 1).

https://doi.org/10.33436/v31i3y202101

Table 1. IUB code for Bases A, C, G, T

IUB Code	N	V	B	H	D	K	S	W	M	Y	R
Bases	A,C,G,T	G,A,C	G,T,C	A,T,C	G,A,T	G,T	G,C	A,T	A,C	C,T	A,G

Now Table 2 classifies DNA sequence into following groups on the basis of chemical structure:

Table 2. DNA sequences classification

Sr. No.	Classification
1	Purine R = (A,G) or pyrimidine Y = (C,T)
2	Amino M = (A,C) or keto K = (G,T)
3	Weak H bond W = (A,T) or strong H bond S = (C,G)

We have given a new 3-dimensional graphical representation using the approach of DNA's main series on bases about Table 2 classification (Arora & Kansal, 2019), (Chen et al., 2008), (Tripathi & Kansal 2019). Assume S = $s_1s_2s_3\ldots s_i$ is a random primary sequence. For R/Y allocation, homomorphic mapping HM(S) = HM(s_1) HM(s_2)…, where

$$HM(s_i)=\begin{cases}(0,0,i) & if\ s_i \in R=(A,G)\\ (0,1,i) & if\ s_i = C\\ (1,0,i) & if\ s_i = T\end{cases}$$

This is used for mapping primary sequence in a plot set. For example, consider a primary sequence of Goat's DNA consisting of ACTGCTTAGT. Corresponding plot set is: [(0,0,1), (0,1,2), (1,0,3), (0,0,4), (0,1,5), (1,0,6), (1,0,7), (0,0,8), (0,0,9), (1,0,10)]. We give corresponding plot set curve name as RCT for this sequence. This zigzag curve which connects all point is known as RCT – characteristic curve is shown in Figure 1.

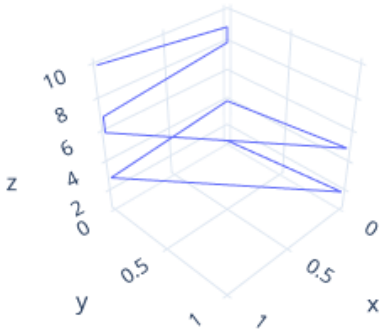

Figure 1. RCT characteristic curve for data ACTGCTTAGT

M/K classification given by:

$$HM(s_i)=\begin{cases}(0,0,i) & if\ s_i \in M=(A,C)\\ (0,1,i) & if\ s_i = G\\ (1,0,i) & if\ s_i = T\end{cases}$$

Corresponding plot set is: [(0,0,1), (0,0,2), (1,0,3), (0,1,4), (0,0,5), (1,0,6), (1,0,7), (0,0,8), (0,1,9), (1,0,10)]. Figure 2 shows MGT – characteristic curve.

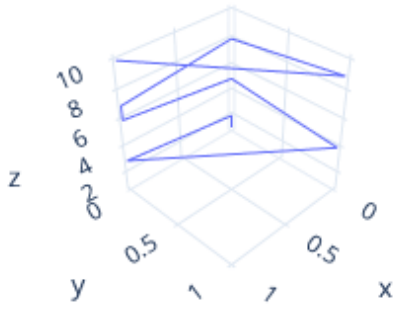

Figure 2. MGT characteristic curve for data ACTGCTTAGT

Now see the W/S classification:

$$HM(s_i)=\begin{cases}(0,0,i) & if\ s_i \in W=(A,T)\\ (0,1,i) & if\ s_i = C\\ (1,0,i) & if\ s_i = G\end{cases}$$

Corresponding plot set is: [(0,0,1), (0,1,2), (0,0,3), (1,0,4), (0,1,5), (0,0,6), (0,0,7), (0,0,8), (1,0,9), (0,0,10)]. Figure 3 shows WCG – characteristics curve.

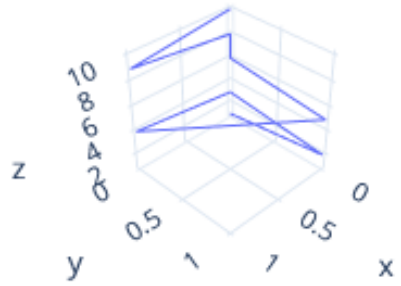

Figure 3. WCG characteristics curve of data ACTGCTTAGT

In a similar way, we draw another characteristic curve for opposite base pairs for each category of 4 bases. For R/Y we have YAG – characteristics curve, for M/K we get KAC – characteristic curve and for W/S allocation SAT – characteristics curve (Diaz et al., 2008).

Again we observe that for each category of 4 bases, on changing the location of 4 bases we can find other 6 characteristic curves, that are RTC, MTG, WGC, YGA, KCA and STA – characteristic curve symmetric to RCT, MGT, WCG, YAG, KAC and SAT – characteristic curve. So for each category of bases there are 4 characteristic curves and 12 different zigzag curves reflecting the same DNA sequence (Huangg et al., 2008). Therefore this approach considers both sequential adjacency as fine as chemical structure for primary sequences. Deprivation of information is avoided as curve overlap and intersects.

3. Numerical characterization

For DNA sequence, collection of spots are (x_i, y_i, z_i) for i = 1, 2, …, n. Here n represents limit of DNA sequences (Chaou, 2010), (Taang, et al. 2010). Use the following formula to find coordinates of geometrical centers:

$$x'=\frac{1}{n}\sum_{i=1}^{n}x_i, \qquad y'=\frac{1}{n}\sum_{i=1}^{n}yx_i, \qquad z'=\frac{1}{n}\sum_{i=1}^{n}z_i \tag{1}$$

We have 3 geometrical centers with respect t to patterns RCT, MGT and WCG respectively. A 9 component vector can be constructed which consists of 3 geometric center and Euclidean distance between end point of vectors (Zhaao et al., 2019). Fluctuation of vector for the coding sequence is given by:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}((x_i - x')^2 + (y_i - y')^2 + (z_i - z')^2) \quad (2)$$

We calculate 12 component vectors, 3 GC as well as fluctuations of primary DNA series. Distance ED among two vectors can be calculated by:

$$ED_{ij} = \sum_{k=1}^{3}\sqrt{[x'_k(i) - x'_k(j)]^2 + [y'_k(i) - y'_k(j)]^2 + [z'_k(i) - z'_k(j)]^2 + [s^2_k(i) - s^2_k(j)]^2} \quad (3)$$

The strengths of this new 3D graphical representation approach are the following:

- Graphical representations are respected which can change DNA groupings into visual bends as well as offer powerful mathematical descriptors.
- On account of its comfort and incredible mobility, strategies dependent on this 3 dimensional graphical representations have been widely applied in pertinent domains of bioinformatics.
- When we apply this technique to rationed gene families like the globin genes, we find that the guides of various genes appear to have unmistakable examples, suggesting that the examples could be used for global succession homology.
- It can describe the graphical representations for DNA arrangements precisely and get sensible outputs of similarities/dissimilarities within the examination of DNA sequences.
- It is preferable to the traditional invariants in predicting similarity and dissimilarity among different species.
- It can support appropriate sequence alignment tools for both computational scientists and molecular biologists.

4. Results and discussion

Code sequence of 10 particular species Human_, Goat_, Opossum_, Gallus_, Lemur_, Mouse_, Rabbit_, Rat_, Gorilla_ and Chimpanzee_ is addicted in Table 3.

Table 3. DNA sequence of 10 unlike breeds with accession number

Sr. No.	Species	Accession Number
1	Human_	U01317
2	Goat_	M15387
3	Opossum_	J03643
4	Gallus_	V00409
5	Lemur_	M15734
6	Mouse_	X06701
7	Rabbit_	V00882
8	Rat_	V00722
9	Gorilla_	U01519
10	Chimpanzee_	U02277

We have calculated 3 geometrical centers for RCT(x'_1,y'_1,z'_1), GC for MGT(x'_2,y'_2,z'_2) and WCG(x'_3,y'_3,z'_3) as well as fluctuation S_1^2, S_2^2, S_3^2 of 10 species of Table 3. This calculated information is listed in Table 4.

Table 4. Three GC and fluctuation of characteristic curves of the DNA sequence

Species	RCT(x'_1, y'_1, z'_1, S_1^2)	MGT(x'_2, y'_2, z'_2, S_2^2)	WCG(x'_3, y'_3, z'_3, S_3^2)
Human_	(0.2133, 0.1912, 0.9413, 0.5866)	(-0.1446, 0.1915, 0.9446, 0.6141)	(0.2151, -0.1441, 0.9442, 0.7026)
Goat_	(0.2012, 0.2008, 0.9411, 0.5561)	(-0.2011, 0.2036, 0.9412, 0.6254)	(0.2061, -0.2011, 0.9420, 0.6472)
Opossum_	(0.1054, 0.0598, 0.9443, 0.5845)	(-0.0792, 0.0610, 0.9447, 0.5769)	(0.1053, -0.0786, 0.9442, 0.7141)
Gallus_	(0.0614, 0.2514, 0.9501, 0.6130)	(-0.1455, 0.2558, 0.9446, 0.6030)	(0.0638, -0.1456, 0.9443, 0.6659)
Lemur_	(0.2488, 0.0814, 0.9414, 0.5589)	(-0.1665, 0.0816, 0.9446, 0.6560)	(0.2581, -0.1658, 0.9444, 0.6941)
Mouse_	(0.1984, 0.1524, 0.9449, 0.6047)	(-0.0877, 0.1535, 0.9448, 0.6131)	(0.1973, -0.0573, 0.9442, 0.7190)
Rabbit_	(0.2514, 0.1645, 0.9457, 0.5798)	(-0.2091, 0.1684, 0.9445, 0.6089)	(0.2581, -0.2094, 0.9441, 0.6829)
Rat_	(0.1784, 0.1056, 0.9458, 0.6587)	(-0.1442, 0.1041, 0.9446, 0.6356)	(0.1711, -0.1442, 0.9442, 0.6671)
Gorilla_	(0.2191, 0.1913, 0.94490.5867)	(-0.1524, 0.1945, 0.9446, 0.6149)	(0.2189, -0.1518, 0.9445, 0.7016)
Chimpanze e_	(0.2321, 0.1561, 0.9512, 0.6015)	(-0.1532, 0.1562, 0.9512, 0.61112)	(0.2326, -0.1532, 0.9506, 0.6939)

We gave the similarity/dissimilarity matrix as DNA sequence of Table 3 as per basis of ED among various end points of 12-component vectors. This matrix is characterized in Table 5.

Table 5. Similar/dissimilar matrix for DNA data on the basis of ED

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Chimpanzee
Human_	0	0.1048	0.2612	0.2371	0.1742	0.1031	0.1184	0.1574	0.0142	0.0572
Goat_		0	0.3132	0.2374	0.2012	0.1974	0.1001	0.1985	0.0944	0.1212
Opossum_			0	0.3026	0.2622	0.1899	0.3246	0.1765	0.2741	0.2532
Gallus_				0	0.3812	0.2564	0.3147	0.2702	0.2389	0.2801
Lemur_					0	0.1822	0.1451	0.1621	0.1768	0.1265
Mouse_						0	0.1947	0.1341	0.1152	0.1081
Rabbit_							0	0.1962	0.1074	0.918
Rat_								0	0.1654	0.1312
Gorilla_									0	0.0602
Chimpanzee_										0

On analysis of Table 5, we find that highest similarity is between (Human, Gorilla), (Human, Chimpanzee) and (Gorilla, Chimpanzee) pairs, highest dissimilarity seems like among (Gallus, lemur) and (Goat, opossum). This similarity/dissimilarity is the result of a relationship among these species in evolutionary sense.

5. Conclusion and future scope

Various DNA sequences analyses continue through comparative study focused on seeking similar sequences data. So, there is a requirement for a method to rapidly compare and seek for huge amount of DNA-related information. This alignment-based approach to comparing similarity is very authentic, but it takes more time and space difficulties are insufficient to handle enormous amounts of DNA data. To overcome this limitation, we presented a new comparison method for large-scale sequences. Our method converts genome data into a random 3-D sketch, and then replaces the sequence comparison problem with geometric object comparisons. This method examines not only the development of sequences, but also the chemical construction of main data. The amount of data flowed from sequence to graphical representation is likewise reduced.

There are various research concerns which flourish from the work done by us. We have worked on solving many complex problems in sequence alignment by alignment-free approach. In this paper, there are many other evolutionary and bioinformatics issues that could be associated to improvements. In the proposed 3 dimensional graphical representation of DNA, the z coordinate of nucleotide A, C, T and G could again be expanded to calculate universal z coordinate values.

REFERENCES

1. Arora, M. & Kansal, V. (2019). *Character Level Embedding with Convolution Neural Network for Text Normalization of Unstructured Data for Twitter Sentiment, Social Network Analysis and Mining*, 9(1), ISSN 1869-5450, pp 12:1-12:14, Springers, DOI 10.1007/s13278-019-0557-y, 2019.
2. Chaou, K. C. (2010). *Graphic rule for drug metabolism systems*. Curr. Drug Metabol, 2010;11:369–378.
3. Chen, W., Liao B., Liu Y., Zhu W., Su Z. (2008). *A numerical representation of DNA sequence and its applications*. MATCH Communications in Mathematical and in Computer Chemistry, 2008;60:291–301.
4. Daa Young, L., Kim Kyung Rim, Kim Taeyoong & Choo Hwan-Guee (2016). *Comparison-specialized visualization model for whole genome sequences*. Journal of WSCG, 24(2):43–52, 2016.
5. Diaz, G. H., Gonzaleez Diaz Y., Santana L., Ubeira F.M., Uriartee E. (2008). *Proteomics, networks, and connectivity indices*. Proteomics, 2008;8:750–778.
6. Huangg, G., Lioo B., Li Y., Liuu Z. H. (2008). *Curves: a novel 2D graphical representation for DNA sequences*. Chemical Physics Letters, 2008;462:129–132.
7. Hwaangue Choo Dayoung, Lee, D. Kwon (2016). *WebGL based Visualization System for Whole Genomes*. In Proceedings of KIISE, Korea Information Science Society, 2016, pp. 1414–1417.
8. Jiyuan, A., John Laii, Atul Sajjanhaar, J. Batra, Chenwei Wangg & Colleen C. Nelson (2015). *Jcircos: an interactive circos plotter*. Bioinformatics, 31(9):1463–1465, 2015.
9. Kinetic, A. J. (2008). *Plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs*. Canadian Journal of Chemistry, 2008;86:342–356.
10. Krzywinski, M., Jacqueline Scheein, InancBirol, J. Connors, Randy Gacoyne, Doug Horsmaan, S. J. Jones & Marco AMarra (2009). *Circos: an information aesthetic for comparative genomics*. Genome research, 19(9):1639–1645, 2009.
11. Kwon, D. (2015). *Whole genome data visualization and analysis using 3d random walk plot*. Master thesis, Pusan National University, 2015.
12. Nakamura, T., Keisi Taki, Hiroki Nomiyaa, Kazuhiro S. & Kuniaki Uehraa (2013). *A shape-based similarity measure for time series data with ensemble learning*. Pattern Analysis and Applications, 16(4):535–548, 2013.
13. Randic, M. (2017). *A group of 3D graphical representation of DNA sequences based on dual nucleotides*. International Journal of Quantum Chemistry, 2008;108:1485–1491.

14. Randic, M. (2017). *Graphical representations of DNA as 2-D map*. Chemical Physics Letters, 2017;386:468–471.

15. Taang, X. C., Zhaou P. P., Qiuu W.Y. (2010). *On the similarity/dissimilarity of DNA sequences based on 4D graphical representations*. Chin. Sci. Bull, 2010;55:701–704.

16. Tripathi, S. & Kansal, V. (2019). *Machine Translation Evaluation: Unveiling the role of Dense Sentence Vector Embedding for Morphologically Rich Language.* International Journal of Pattern Recognition and Artificial Intelligence, World Scientific Publishing, 34(1), March 2019, pp. 2059001-18, DOI 10.1142/S0218001420590016.

17. Young Lee, D., Hae S.Tak, Han-Ho Kim & Hwan-Gue Cho (2018). *Alignmentfree Sequence Searching over Whole Genomes Using 3D Random Plot of Query DNA Sequences*, Informatica 42 (2018), pp. 357–368.

18. Zhaao, L. P., Lv Y. H., Lii C., Yao M. H., Jin X. Z. (2019). *An S-curve-based approach of identifying biological sequences.* Acta Biotheor, 2019;58:1–13.

Kshatrapal SINGH, B. Tech. (CSE), M. Tech (CSE) & Pursuing Ph.D. from Dr. APJ AK Technical University, Lucknow (India). Currently he is working as an Assistant Professor at ITS Engineering College, Greater Noida. His specialization is Sequence Comparison, Compiler Design, Discrete Structure and Graph Theory.

Ashish KUMAR, B. Tech. (CSE), M. Tech (IIT-K) & Ph.D. from UPES, Dehradun (UK). Currently he is working as Professor and Head of the Department of Computer Science & Engineering at ITS Engineering College, Greater Noida. His specialization is Networking, Web Technology, Java and Software Engineering.

Manoj Kumar GUPTA, B. Tech. (CSE), M. Tech (CSE) & Ph.D. from IIT, Roorkee. Currently he is working as an Associate Professor and Head of the Department of Computer Science at Shri Mata Vaishno Devi University, Katra (J&K). His specialization is Sequence Comparison, Networking, Object Oriented Technologies, and Python etc.