MOL2NET, International Conference Series on Multidisciplinary Sciences
http://sciforum.net/conference/mol2net-03

Some Comments on Mathematical Descriptors of Biomolecular Sequences and their Characteristics

Ashesh Nandy (E-mail: anandy43@yahoo.com) [a], Subhash C Basak[b]

[a]*Centre for Interdisciplinary Research and Education, 404B Jodhpur Park, Kolkata 700058, India*

[b]*University of Minnesota Duluth-Natural Resources Research Institute and Department of Chemistry and Biochemistry, University of Minnesota Duluth, 5013 Miller Trunk Highway, Duluth, MN 55811, USA*

Abstract

The advent of techniques of graphical representation and mathematical characterization of biomolecular sequences has seen the growth of a genre of non-alignment methods for analyses of their similarities/dissimilarities. The new descriptors are important and convenient to provide a quantitative measure of the composition and distribution of the basic units, allow discrimination amongst members of a family of similar sequences with a low computational overhead and hold promise for discovery of new systematics. These opportunities led to a plethora of models for graphical representation and numerical characterisations, but the question is how far the various sequence descriptors derived by these different mathematical approaches encode non-redundant information. We briefly consider the issues that when comparative studies of biomolecular sequences are undertaken, it is important to consider which properties are being considered and choose models that allow for computational closure and non-redundancy. We believe graphical representation and numerical characterization models have a significant role to play in non-alignment similarity/dissimilarity analysis of bio-molecular sequences, but the issues have to be approached with an eye to specific properties being investigated.

The advent of techniques of graphical representation and mathematical characterization of biomolecular (DNA/RNA/protein) sequences [1] have seen the growth of a genre of methods for their characterization that predicts properties without need for multiple alignments. In a recent paper analyzing the Zika virus nucleotide sequences, we had used such techniques to characterize the Zika virus genome.

The new descriptors of biomolecular sequences are important and convenient from several aspects:

1. They provide a formal quantitative measure of the biomolecular sequences that capture the essence of basic units' composition and their distribution;
2. Such quantitative information allow for the possibility of discrimination amongst members of a family of similar sequences;
3. Their computation overhead is limited since there is no recourse to computational models required in multiple alignments;

4. They provide a novel approach to routine tasks of bio-sequence analyses which may lead to discovery of new systematics.

The original impetus for graphical representation arose from the work of Hamori et al [2] who mapped a DNA sequence into a three-dimensional grid with, effectively, the four nucleotides mapped along the four cardinal directions in the xy plane and the nucleotide number measured along the z-axis. Plotting each base by taking a step in the designated direction and connecting to the next base point by a short line resulted in a three-dimensional curve, when all points were plotted, representing the distribution of bases along the sequence. Since such graphical representation was practically difficult when done on a paper or a computer screen, some authors proposed two-dimensional graphical representations that captured the essence of such representations by dispensing with the elevation along the z-axis that separated each nucleotide from its neighbors [1,3-5]. This resulted in degeneracies in the representation of the paths taken when the movements had to alternate between two bases placed on the same axis. Quantitative measures were devised to distinguish between two closely related sequences: Gates [3] proposed the Manhattan distance where distance between two nucleotides in a DNA sequence were measured as a sum of the distances measured in terms of steps taken along the axes to reach from one base to the target base., Raychaudhury and Nandy [7] proposed a geometrical measure of moments of the graphs and a graph radius as DNA descriptors.

Zeroth order: $X = N_G - N_A \quad Y = N_C - N_T$
First order: $\mu_x = \Sigma\, x_i / N \quad \mu_y = \Sigma\, y_i / N \quad g_{\underline{R}} = \sqrt{(\mu_{\underline{x}}^2 + \mu_{\underline{y}}^2)}$

where N_i (I = A, C, G, Y) represents the base composition numbers, X, Y are the end points of the 2D curve, N is the total number of bases, μ_x, μ_y are weighted average co-ordinates and $g_{\underline{R}}$ is the graph radius which measures the base distribution in terms of the graph spread and thus represents a sequence descriptor.

Although such 2D graphical representations provided a reasonable view of the base distribution along a sequence, the degeneracy feature was a stumbling block that inspired many authors to rushin to fill the void [1]. Randic et al [7] resurrected the 3D model in a new form and also instituted a D/D matrix whose eigenvalues λ_i were supposed to represent the sequence descriptor. The issue of (a) designing a descriptor for a DNA sequence, (b) devising a means to discriminate between sequences and (c) avoiding degeneracy became a Holy Grail of the DNA descriptor world. By last count there may be over a hundred models devised to accomplish this task and new models are being proposed yet (see e.g., 8,9].

Given this plethora of models, and that all of them proposed descriptors computed either through geometrical methods or eigenvalues of underlying matrices, it is pertinent to ask how far the various sequence descriptors derived by different mathematical approaches encode redundant information [10]. The 2D-dynamic model of Bielinska-Waz et al [11], where the redundancies of the original Nandy 2D model [4] was to be removed by apportioning masses at each vertex, reproduced the exact same descriptor values at the 1st order moments level as the naïve 2D model [4] applied to the Zika virus genomes [12]. In a separate exercise where seven different representations of DNA sequences were analyzed for descriptor computations as applied to various globin and other genes [10], some were

found to be strongly correlated; those that were not must be representing some specific characteristics of the base composition and distribution, but it has not been clear what these are.

Thus, when comparative studies of biomolecular sequences are undertaken, it is important to consider which properties one is looking at and choose models that allow for fair computation, non-redundancy with other models and improvement over multiple alignment systems. In another paper elsewhere in this conference [13] we show that a commonly held belief in the similarity of Dengue type 2 virus and Zika virus genomes, supported by a BLAST analysis, the nucleotide sequences are indeed quite different as demonstrated in a 2D graphical representation. Non-alignment, graphical representation and numerical characterization models thus have a significant role to play in similarity/dissimilarity analysis of bio-molecular sequence analysis, but the issue has to be approached with an eye to specific properties being investigated.

References

1. A Nandy, M Harle, S C Basak, Mathematical descriptors of DNA sequences: development and applications, *ARKIVOC* Vol. 9, 211-238, 2006.
2. Hamori E and Ruskin J 1983 H curves, a novel method of representationo f nucleotide series especially suited for long range DNA sequences. J; . BioI. Chem. 258 1318-1327
3. Gates M A 1986 A simple way .to look at DNA; J. Theor. BioI. 119 319-328
4. A Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and Application to Globin Genes, *Current Sc* 66(4), 309-314, 1994.
5. Leong P M and Morgenthaler S 1995 Random walk and gap plots of DNA sequences; Comput Applic. Bio.Sci. 11, 503-507
6. C Raychaudhury and A Nandy, "Indexing Scheme and Similarity Measures for Macromolecular Sequences", *Journal of Chemical Information and Computer Science*, 39, 243-247, 1999.
7. M Randic, M Vracko, A Nandy and S C Basak, On 3-D representation of DNA primary sequences, *J Chem Infor and Comput Sc* 40, 1235-1244, 2000.
8. Rajendra Kumar Bharti, Archana Verma and K.S. Vaisla. A New 2.D RAK Method of Representation and Analysis of a DNA Sequence Advances in Computer Science and Engineering, Ch.7, pp 65-72. Ed P. Grag and J. Ranjan. Macmilan Publishers Inds. Ltd., New Delhi, 2009.
9. Manoj Kumar Gupta, Rajdeep Niyogi, Manoj Misra. A 2D Graphical Representation of Protein Sequence and Their Similarity Analysis with Probabilistic Method MATCH Commun. Math. Comput. Chem. 72 (2014) 519-532.
10. Dwaipayan Sen, Subhadeep Dasgupta, Indrajit Pal, Smarajit Manna, Subhash C Basak6, Ashesh Nandy and Gregory D. Grunwald. Intercorrelation of major DNA/RNA Sequence Descriptors - A Preliminary Study. *Curr Comp-Aided Drug Des*, 2016, *12(3)*, 216-228.
11. D Bielinska-Waz, T. Clark, P. Waz, W. Nowak, A. Nandy2D-dynamic representation of DNA sequences, *Chem. Phys. Lett.*442/1-3 pp. 140-144, 2007.
12. *Nandy A, Dey S, Basak SC, Bielinska-Waz D, Waz P. Characterizing the Zika Virus Genome – A Bioinformatics Study. Curr. Comp.-Aided Drug Des* 2016*, 12, 87-97.*
13. Dey S, Roy P, Nandy A, Basak SC, Das S. Comparison of Base Distributions in Dengue, Zika and Other Flavivirus Envelope and NS5 Genes. *MOL2NET* 2017, 3,