2D-dynamic Representation of DNA Sequences - Computational and Graphical Tool for Similarity Analysis in Bioinformatics

Dorota Bielińska-Wąż

Department of Radiological Informatics and Statistics
Medical University of Gdańsk
Tuwima 15, 80-210 Gdańsk, Poland
Email: djwaz@gumed.edu.pl

Piotr Wąż

Department of Nuclear Medicine
Medical University of Gdańsk
Tuwima 15, 80-210 Gdańsk, Poland
Email: phwaz@gumed.edu.pl

***Abstract*—A new nonstandard method of comparison of deoxyribonucleic acid (DNA) sequences called by us 2D-dynamic Representation of DNA Sequences is presented. This approach is based on a method known in the literature as Nandy plots but in the present method the degeneracy (non-uniqueness) of the Nandy plots has been removed. 2D-dynamic Representation is computationally not demanding and there are no limitations on the lengths of the DNA sequences. Using this method, one can compare DNA sequences both graphically and numerically.**

Keywords–Bioinformatics; Alignment-free methods; Descriptors.

I. INTRODUCTION

A variety of problems in bioinformatics is large and new approaches are still constructed. Molecular biology is a young area. Its beginning may be dated to 1953 when Watson and Crick discovered the structure of DNA [1]. In 1995 the genome of bacteria *Haemophilus influenzae* has been sequenced for the first time [2]. The project on human genome *Human Genome Project* has been finished in 2003. According to the data in 2013, the database GenBank contains the nucleotide sequences coming from 260 000 described species [3]. The increase of the amount of information available in databases stimulated the development of bioinformatical methods.

Graphical representations of DNA sequences constitute both numerical and graphical tools for similarity/dissimilarity analysis of DNA sequences. They belong to the class of approaches known in the literature as alignment-free methods. Examples of these kind of methods may be found in [4]–[24] (for reviews see [25] [26]). These methods can be applied for solving a large class of problems in biology and medical sciences that require such an analysis. One of such approaches has been introduced by us and we call it *2D-dynamic representation of DNA sequences* [27]–[31].

II. METHOD AND RESULTS

2D-dynamic representation of DNA sequences is based on shifts in a two-dimensional space [27]. The DNA sequence is represented by material points with different masses in a two-dimensional space. This method is an improvement of *Nandy plots* [6], in which particular bases are represented by two orthogonal pairs of colinear basis vectors. Such a choice of the vectors leads to the possibility of shifts back and forth along the same trace. The so called repetitive walks lead to degeneracy: different sequences may be represented by the same graphs. In order to remove the degeneracy, points with masses which are a multiplicity of the unit mass have been introduced. After a unit shift, a point with unit mass is localized. If the ends of the vectors during the shifts coincide, then the mass of this point increases accordingly. In order to compare the DNA sequences numerically, we have proposed several numerical characteristics (called descriptors in the theory of molecular similarity) of the 2D-dynamic graphs [28]–[30]. We have shown that our numerical approach allows for the classification of the DNA sequences [31]. 2D-dynamic representation of DNA sequences is also a good graphical tool for sequence comparison. Examples of 2D-dynamic graphs of the whole genomes of the Zika virus are shown in Fig. 1 (HQ234500 Nigeria 1968) and in Fig. 2 (KU312312 Suriname 2015). The shapes and the details of the 2D-dynamic graphs give the information about the DNA sequences.

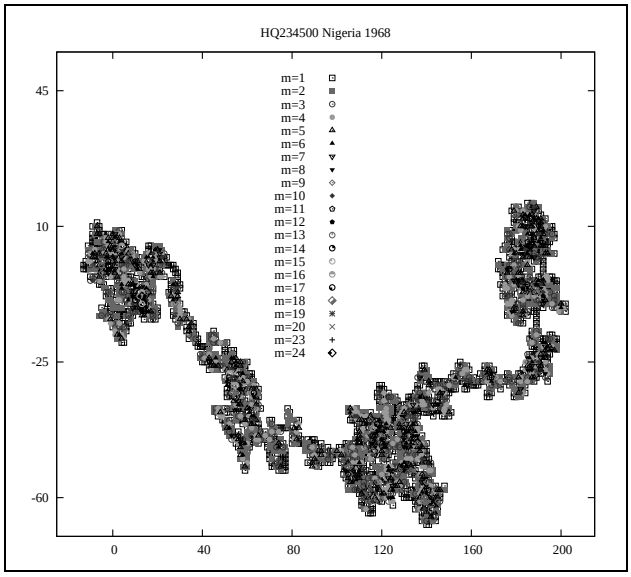

Figure 1. 2D-dynamic graph.

III. CONCLUSION

2D-dynamic representation of DNA sequences is both graphical and numerical tool for similarity/dissimilarity analy-

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-513-5

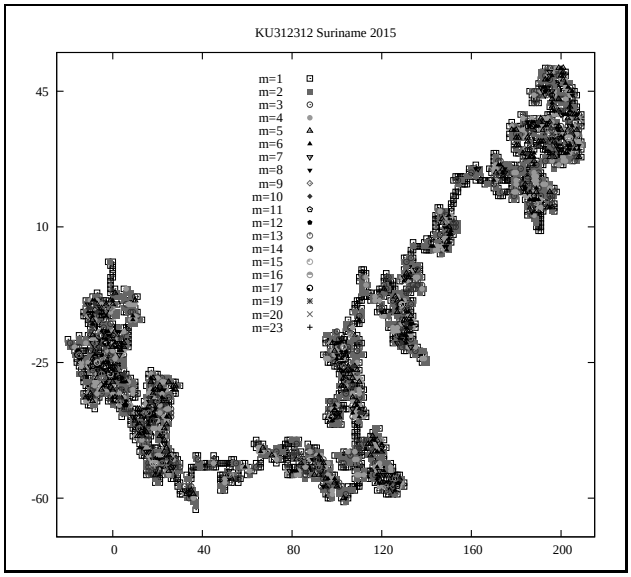

Figure 2. 2D-dynamic graph.

sis of DNA sequences. It can be applied to all problems in biology and medicine, which require such an analysis. An example of an application of 2D-dynamic representation of DNA sequences may be found in our recent work [32]. We have shown that a mutation of the Zika virus genome can be described both graphically, and numerically using for example the so called centers of mass of the 2D-dynamic graphs:

$$\mu_x = \frac{\sum_i m_i x_i}{\sum_i m_i}, \quad \mu_y = \frac{\sum_i m_i y_i}{\sum_i m_i}, \tag{1}$$

where x_i, y_i are the coordinates of the mass m_i in the 2D-dynamic graph. Some other descriptors of the 2D-dynamic graphs will be also applied for characterizing the Zika virus genome in a subsequent article.

REFERENCES

[1] J. D. Watson and F. H. C.Crick, "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid", Nature vol. 171, pp. 737–738, 1953.

[2] R. D. Fleischmann et al., "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd", Science vol. 269, pp. 496–512, 1995.

[3] D. A. Benson et al.,"GenBank", Nucleic Acids Res.vol. 41 (Database issue), pp. D36–D42, 2013.

[4] E. Hamori and J. Ruskin, "H Curves, a novel method of representation of nucleotide series especially suited for long DNA sequences", J. Biol. Chem. vol. 258, pp. 1318–1327, 1983.

[5] M. A. Gates, "Simpler DNA sequence representations", Nature vol. 316, p. 219, 1985.

[6] A. Nandy, "A new graphical representation and analysis of DNA sequence structure. I: Methodology and application to globin genes", Current Science vol. 66, pp. 309–314, 1994.

[7] P. M. Leong and S. Morgenthaler, "Random walk and gap plots of DNA sequences", Comput. Appl. Biosci. vol. 11, pp. 503–507, 1995.

[8] R. Chi and K. Ding, "Novel 4D numerical representation of DNA sequences", Chem. Phys. Lett. vol. 407, pp. 63–67, 2005.

[9] Q. Dai, X. Liu, and T. Wang, "A novel graphical representation of DNA sequences and its application", J. Mol. Graph. Model. vol. 25, pp. 340–344, 2006.

[10] H. González-Díaz et al., "Generalized lattice graphs for 2D-visualization of biological information", J. Theor. Biol. vol. 261, pp. 136–147, 2009.

[11] P. He and J. Wang, "Numerical characterization of DNA primary sequence", Internet Electron. J. Mol. Des. vol. 1, pp. 668–674, 2002.

[12] N. Jafarzadeh and A. Iranmanesh, "C-curve: a novel 3D graphical representation of DNA sequence based on codons", Math Biosci. vol. 241, pp. 217–224, 2013.

[13] B. Liao, Q. Xiang, L. Cai, and Z. Cao, "A new graphical coding of DNA sequence and its similarity calculation", Physica A vol. 392, pp. 4663–4667, 2013.

[14] Y.-Z. Liu and T. Wang, "Related matrices of DNA primary sequences based on triplets of nucleic acid bases", Chem. Phys. Lett. vol. 417, pp. 173–178, 2006.

[15] M. Randić and M. J. Vračko, "On the similarity of DNA primary sequences", J. Chem. Inf. Comput.Sci. vol. 40, pp. 599–606, 2000.

[16] M. Randić, X. Guo, and S. C. Basak, "On the characterization of DNA primary sequences by triplet of nucleic acid bases", J. Chem. Inf. Comput. Sci. vol. 41, pp. 619–626, 2001.

[17] M. Randić and A. T. Balaban, "On a four-dimensional representation of DNA primary sequences", J. Chem. Inf. Comput. Sci. vol. 43, pp. 532–539, 2003.

[18] M. Randić, M. Vračko, N. Lerš, and D. Plavšić, "Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation", Chem. Phys. Lett. vol. 371, pp. 202–207, 2003.

[19] X. Yang and T. Wang, "Linear regression model of short k-word: A similarity distance suitable for biological sequences with various lengths", J. Theor. Biol. vol. 337, pp. 61–70, 2013.

[20] Y. Yao and T. Wang, "A class of new 2-D graphical representation of DNA sequences and their application", Chem. Phys. Lett. vol. 398, pp. 318–323, 2004.

[21] Y. Yao, X. Nan, and T. Wang, "Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation", Chem. Phys. Lett. vol. 411, pp. 248–255, 2005.

[22] J.-F. Yu, J.-H. Wang, and X. Sun, "Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation", MATCH Commun. Math. Comput. Chem. vol. 63, pp. 493–512, 2010.

[23] Z. Zhang, T. Song, X. Zeng, Y. Niu, Y. Jiang, L. Pan, and Y. Ye, "ColorSquare: A colorful square visualization of DNA sequences", MATCH Commun. Math. Comput. Chem. vol. 68, pp. 621–637, 2012.

[24] S. Zhang, Y. Zhang, and I. Gutman, "Analysis of DNA sequences based on the fuzzy integral", MATCH Commun. Math. Comput. Chem. vol. 70, pp. 417–430, 2013.

[25] D. Bielińska-Wąż, "Graphical and numerical representations of DNA sequences: Statistical aspects of similarity", J. Math. Chem. vol. 49, pp. 2345–2407, 2011.

[26] M. Randić, M. Novič, and D. Plavšić, "Milestones in Graphical Bioinformatics", Int.J.Quant.Chem.vol. 113, pp. 2413–2446, 2013.

[27] D. Bielińska-Wąż, T. Clark, P. Wąż, W. Nowak, and A. Nandy, "2D-dynamic representation of DNA sequences", Chem. Phys. Lett. vol. 442, pp. 140–144, 2007.

[28] D. Bielińska-Wąż, W. Nowak, P. Wąż, A. Nandy, and T. Clark, "Distribution moments of 2D-graphs as descriptors of DNA sequences", Chem. Phys. Lett. vol. 443, pp. 408–413, 2007.

[29] D. Bielińska-Wąż, P. Wąż, and T. Clark, "Similarity studies of DNA sequences using genetic methods", Chem. Phys. Lett. vol. 445, pp. 68–73, 2007.

[30] D. Bielińska-Wąż, P. Wąż, W. Nowak, A. Nandy, and S. C. Basak, "Similarity and dissimilarity of DNA/RNA sequences", Computation in Modern Science and Engineering vol. 2, Proceedings of the International Conference on Computational Methods in Science and Engineering Corfu, Greece, 25-30 September, 2007, Eds T.E. Simos and G. Maroulis, American Institute of Physics, pp. 28–30, 2007.

[31] P. Wąż and D. Bielińska-Wąż, A. Nandy, "Descriptors of 2D-dynamic graphs as a classification tool of DNA sequences", J. Math. Chem. vol. 52, pp. 132–140, 2013.

[32] A. Nandy, S. Dey, S. C. Basak, D. Bielińska-Wąż, and P. Wąż, "Characterizing the Zika virus genome - A bioinformatics study", Curr. Comput. Aided Drug Des. vol. 12, pp. 87–97, 2016.

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-513-5