A Two-Dimensional Computational Model for DNA/RNA Classification

Dorota Bielińska-Wąż
Department of Radiological Informatics and Statistics
Medical University of Gdańsk
80-210 Gdańsk, Poland
email: djwaz@gumed.edu.pl

Piotr Wąż
Department of Nuclear Medicine
Medical University of Gdańsk
80-210 Gdańsk, Poland
email: phwaz@gumed.edu.pl

***Abstract*—The 2D-Dynamic Representation of DNA/RNA Sequences, a two-dimensional computational model introduced by the authors, is reviewed for its application in the classification of deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) sequences. This method falls under the bioinformatics category known as Graphical Representations of Biological Sequences. The goal of these methods is to provide tools for the graphical and numerical classification of the sequences.**

Keywords–bioinformatics; machine learning; decision trees; descriptors

I. Introduction

Graphical Representations of Biological Sequences constitute a branch of alignment-free bioinformatics methods, focusing on the graphical and numerical classification of sequences [1] [2]. Each approach reveals different aspects of similarity, and a comprehensive review can be found in [3].

This document presents a method introduced by the authors and called 2D-Dynamic Representation of DNA/RNA Sequences [4]–[10]. Specifically, this method is combined with the C5.0 decision tree algorithm [10]. Details related to the graphical representation of the sequences within this method and the numerical characteristics of the graphs (”descriptors”) are described in Section II.

II. Methods and Results

Graphically, the sequences are represented by 2D-dynamic graphs (sets of material points in a 2D space). The graphs were obtained by following a ”walk” in the XY coordinate system, using the basis vectors representing the specific nucleobases: A = (-1,0), G = (1,0), C = (0,1), and T/U = (0,-1) [4] [9]. Examples of these graphs are shown in Figures 1-3.

The following descriptors of the 2D-Dynamic Representation of DNA/RNA Sequences, some of which are analogous to dynamics, are considered:

- Coordinates (μ_x, μ_y) of the centers of mass of the 2D-dynamic graphs [4]:

$$\mu_\gamma = \frac{1}{N}\sum_{i=1}^{p} m_i \gamma_i, \quad \gamma = x, y, \quad N = \sum_{i=1}^{p} m_i, \quad (1)$$

  where x_i, y_i are the coordinates of mass m_i in the Cartesian coordinate system for which $(0,0)$ is the origin of all the sequences and N is the length of the sequence (equal to the total mass of the graph) and p is the number of the material points in the graph.
- Principal moments of inertia (I_{11}, I_{22}) of the 2D-dynamic graphs [4].

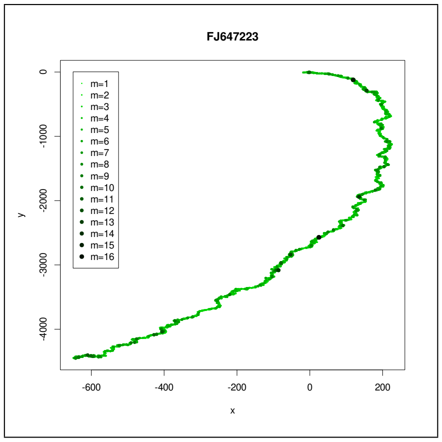

Figure 1. 2D-dynamic graph representing the complete genome sequence of embecovirus (GenBank accession number FJ647223).

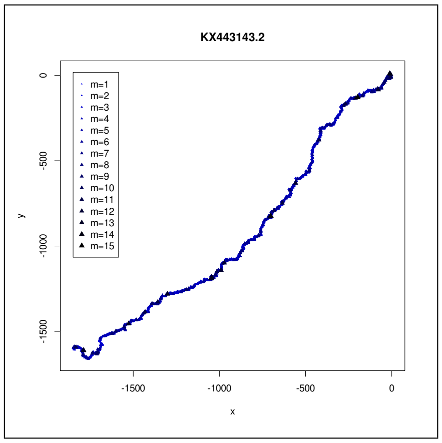

Figure 2. 2D-dynamic graph representing the complete genome sequence of deltacoronavirus (GenBank accession number KX443143.2).

The moment of inertia tensor is defined by the matrix

$$\hat{I} = \begin{pmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{pmatrix} \quad (2)$$

with elements

$$I_{xy} = I_{yx} = -\sum_{i=1}^{p} m_i x_i^\mu y_i^\mu, \quad (3)$$

$$I_{xx} = \sum_{i=1}^{p} m_i (y_i^\mu)^2, \quad (4)$$

Copyright (c) IARIA, 2024. ISBN: 978-1-68558-215-9

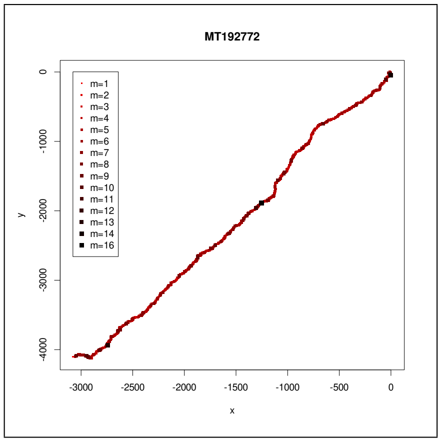

Figure 3. 2D-dynamic graph representing the complete genome sequence of the SARS-CoV-2 virus (GenBank accession number MT192772).

$$I_{yy} = \sum_{i=1}^{p} m_i (x_i^{\mu})^2, \qquad (5)$$

where x_i^{μ}, y_i^{μ} denote the coordinates of mass m_i in the Cartesian coordinate system with the origin at the center of mass of the graph. Principal moments of inertia are equal to the solutions $I = I_{11}, I_{22}$ of equation

$$\left| \begin{array}{cc} I_{xx} - I & I_{xy} \\ I_{xy} & I_{yy} - I \end{array} \right| = 0. \qquad (6)$$

- Moments of the mass-density distributions [5].

  The n-th moment of a discrete distribution ρ_E is defined as

  $$M_{E,n} = c_E \sum_i \rho_{E_i} E_i^n,$$

  $E = x, y$ and the normalization constant

  $$c_E = \left(\sum_i \rho_{E_i} \right)^{-1}.$$

  Moments normalized to a mean value equal to zero ($M'_{E,1} = 0$) are

  $$M'_{E,n} = c_E \sum_i \rho_{E_i} (E_i - M_{E,1})^n.$$

  The moments for which the variance is additionally equal to 1 ($M''_{E,2} = 1$) are also considered:

  $$M''_{E,n} = c_E \sum_i \rho_{E_i} \left[\frac{(E_i - M_{E,1})}{\sqrt{M_{E,2} - (M_{E,1})^2}} \right]^n.$$

- Angles between the x axis and the principal axes of inertia of the 2D-dynamic graphs [6].

  One of the angles smaller than $\frac{\pi}{2}$ is chosen.

- Mass overlaps of the 2D-dynamic graphs [6].
- Descriptors $(D_1^x, D_2^x, D_1^y, D_2^y)$ related to a relation between the coordinates of the center of mass and the principal moments of inertia of the 2D-dynamic graphs [7]:

  $$D_k^{\gamma} = \frac{\mu_{\gamma}}{I_{kk}}, \quad k = 1, 2; \quad \gamma = x, y. \qquad (7)$$

- Graph radius [8]:

  $$g_R = \sqrt{\mu_x^2 + \mu_y^2}. \qquad (8)$$

- Matrix elements of the moments of inertia tensor (I_{xx}, I_{yy}, I_{xy}) of the 2D-dynamic graphs [10].

2D-Dynamic Representation of DNA/RNA Sequences has been applied for the similarity analysis of:

- histone H4 coding sequences of different species [4]–[7];
- α-globin coding sequences of different species [4] [7];
- complete genome sequences of the Zika virus [8] [9];
- 20 most common subtypes of influenza A virus [10].

III. Conclusion

In summary, the 2D-Dynamic Representation of DNA/RNA Sequences is an effective tool for both graphical and numerical comparison of the sequences. Notably, combining this method with the C5.0 decision tree algorithm has yielded high mean accuracy in predicting the subtype of the influenza A virus, with over 90% correct predictions. This high number of correct predictions confirms the good explainability of the considered systems. Therefore, the method will be applied in the future to interpret our experimental data.

References

[1] K. E. Wade, L. Chen, C. Deng, G. Zhou, and P. Hu, ”Investigating alignment-free machine learning methods for HIV-1 subtype classification”, Bioinformatics Advances vol. 4, Art. No. vbae108, 2024.

[2] D. Bielińska-Wąż, P. Wąż, A. Błaczkowska, J. Mandrysz, A. Lass, and P. Gładysz, J. Karamon, ”Mathematical Modeling in Bioinformatics: Application of an Alignment-Free Method Combined with Principal Component Analysis”, Symmetry vol. 16, Art. No. 967, 2024.

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

[4] 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.

[5] 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.

[6] 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.

[7] P. Wąż, D. Bielińska-Wąż, and A. Nandy, Descriptors of 2D-dynamic graphs as a classification tool of DNA sequences, J. Math. Chem. vol. 52, pp. 132–140, 2014.

[8] A. Nandy, S. Dey, S.C. Basak, Bielińska-Wąż, and P. Wąż, Characterizing the Zika Virus Genome - A Bioinformatics Study, Curr. Comput. Aided Drug Des. vol. 12, pp. 87–97, 2016.

[9] D. Panas, P. Wąż, D. Bielińska-Wąż, A. Nandy, and S.C. Basak, 2D-Dynamic Representation of DNA/RNA Sequences as a Characterization Tool of the Zika Virus Genome, MATCH Commun. Math. Comput. Chem. vol. 77, pp. 321–332, 2017.

[10] D. Panas, P. Wąż, D. Bielińska–Wąż, A. Nandy, and S.C. Basak. An Application of the 2D-Dynamic Representation of DNA/RNA Sequences to the Prediction of Influenza A Virus Subtypes, MATCH Commun. Math. Comput. Chem. vol. 80, pp. 295–310, 2018.