3D-Dynamic Representation of DNA/RNA Sequences: A Review

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

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

***Abstract*—The research aims to develop new bioinformatics techniques known in the literature as Graphical Representation Methods. This methodology allows for the calculation of numerical values describing deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) sequences, enabling both graphical and numerical analysis of their similarities and differences. This document provides an overview of a bioinformatics method we introduced, referred to as 3D-Dynamic Representation of DNA/RNA Sequences. In this method, the sequences are represented as sets of material points in 3D space, forming "3D-dynamic graph". Numerically, this three-dimensional dynamic graph is characterized by quantities analogous to those used in classical dynamics. The accuracy of this approach is high, allowing to distinguish sequences that differ by just one nucleobase. One application of this method is the characterization of viral genome sequences. Specifically, combining the 3D-Dynamic Representation of DNA/RNA Sequences with the random forest algorithm effectively classifies subtypes of influenza A virus strains.**

Keywords–supervised learning; bioinformatics; biostatistics; graphical methods; machine learning; random forest; Boruta algorithm

I. INTRODUCTION

This presentation describes a computational method we developed, known as the 3D-Dynamic Representation of DNA/RNA Sequences [1]–[4], which generalizes our previous 2D approach [5].

This approach is part of a broader category of techniques known as Graphical Representation Methods, which allow for both graphical and numerical comparisons of objects. Each method offers a unique perspective on similarity, and new techniques are continually being developed (for reviews, see [6]–[8]).

3D-Dynamic Representation of DNA/RNA Sequences aims to compare the sequences by representing them as sets of material points in 3D space, referred to as "3D-dynamic graphs." The distribution of these points and the calculation of their numerical characteristics ("descriptors") are described in Section II.

II. METHOD AND RESULTS

The method is based on shifts (or *walks*) in 3D space [1]. Nucleobases in a DNA/RNA sequence are represented by basis vectors: adenine A=(-1,0,1), cytosine C=(0,1,1), thymine/uracil T/U=(0,-1,1), and guanine G=(1,0,1). The walk begins at the origin point (0,0,0). This point is shifted by a basis vector corresponding to the first nucleobase in the sequence. At the end of this vector, a mass m=1 is placed, which serves as the starting point for the next shift representing the second nucleobase. This process is repeated for each nucleobase in the sequence. The resulting set of material points, which represents the entire sequence, is termed a 3D-dynamic graph (analogous to the 2D-dynamic graph used in the 2D method). Examples of 3D-dynamic graphs representing the complete genome sequences of embecovirus, the SARS-CoV-2 virus, and deltacoronavirus are shown in Figure 1. The differences between the sequences are clearly visible in the graphs. The calculations were conducted using nucleotide sequence data obtained from GenBank. FJ647223, MT192772, and KX443143.2 are the accession numbers corresponding to the sequences in this database.

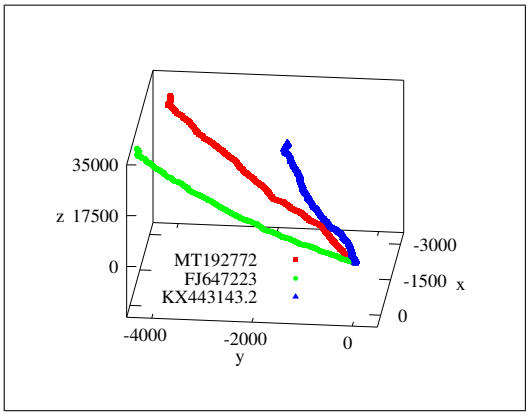

Figure 1. 3D-dynamic graphs.

We use the following descriptors (numerical characteristics of the 3D-dynamic graphs):

- Coordinates (μ_x, μ_y, μ_z) of the center of mass of the graph

$$\mu_\gamma = \frac{\sum_{i=1}^{N} m_i \gamma_i}{\sum_{i=1}^{N} m_i}, \quad \gamma = x, y, z, \tag{1}$$

where x_i, y_i, z_i are the Cartesian coordinates of mass m_i with point $(0, 0, 0)$ being the origin of the coordinate system and N is the length of the sequence. Since $m_i = 1$ for all material points, the total mass of the sequence is equal to the length of the sequence $N = \sum_{i=1}^{N} m_i$. The coordinates of the center of mass of the 3D-dynamic graph can then be expressed as:

$$\mu_\gamma = \frac{1}{N} \sum_{i=1}^{N} m_i \gamma_i, \quad \gamma = x, y, z. \tag{2}$$

- The principal moments of inertia (I_1, I_2, I_3) of the graph, where the moment of inertia tensor is defined by the matrix

$$\hat{I} = \begin{pmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{pmatrix} \quad (3)$$

with elements

$$I_{aa} = \sum_{i=1}^{N} m_i \left[(b_i^\mu)^2 + (c_i^\mu)^2 \right], \quad (4)$$

and

$$I_{ab} = I_{ba} = -\sum_{i=1}^{N} m_i a_i^\mu b_i^\mu, \quad (5)$$

where $\{a, b, c\} = \{x, y, z\}$, $a \neq b \neq c$ and the coordinates $(x_i^\mu, y_i^\mu, z_i^\mu)$ of m_i are determined in the center-of-mass of the graph coordinate system. The principal moments of inertia are equal to the solutions $I = I_1, I_2, I_3$ of the characteristic equation of $\hat{I}$:

$$\begin{vmatrix} I_{xx} - I & I_{xy} & I_{xz} \\ I_{xy} & I_{yy} - I & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} - I \end{vmatrix} = 0. \quad (6)$$

- Matrix elements of the moment of inertia tensor of the graph $(I_{xx}, I_{yy}, I_{zz}, I_{xy}, I_{xz}, I_{yz})$.
- Graph radius, defined as

$$g_R = \sqrt{\mu_x^2 + \mu_y^2 + \mu_z^2}. \quad (7)$$

- Descriptors D_k^γ,

$$D_k^\gamma = \frac{\mu_\gamma}{I_k}, \quad k = 1, 2, 3; \quad \gamma = x, y, z, \quad (8)$$

that depict a relation between the coordinates of the center of mass and the principal moments of inertia of the graph.
- Normalized principal moments of inertia of the graph (r_1, r_2, r_3):

$$r_k = \sqrt{\frac{I_k}{N}}. \quad (9)$$

- The values of C_{ik}.
The relative orientation of the new and old coordinate systems can be described by cosines of appropriately defined angles:

$$C_{ik} \equiv \cos(M_i, Q_k), \quad i, k = 1, 2, 3. \quad (10)$$

M_1, M_2 and M_3 mean the planes (X, Y), (X, Z) and (Y, Z), respectively. Similarly, Q_1, Q_2, Q_3 denote the planes (Ω_1, Ω_2), (Ω_1, Ω_3), (Ω_2, Ω_3).

The descriptors derived from the 3D-Dynamic Representation of DNA/RNA Sequences have proven effective for the similarity analysis of:

- histone H4 coding sequences of different species and α-globin coding sequences of different species [1];
- β-globin genes of different species [2];
- complete genome sequences of dengue virus [3];
- 20 most common subtypes of influenza A virus [4].

Notably, it has been demonstrated that combining the 3D-Dynamic Representation of DNA/RNA Sequences with the random forest algorithm effectively classifies subtypes of influenza A virus strains [4]. In these studies, the following 22 descriptors were considered: the 3 coordinates of the center of mass $\mu = \{\mu_\gamma : \gamma = x, y, z\}$; the 6 elements of the inertia tensor $J = \{I_{xx}, I_{yy}, I_{zz}, I_{xy}, I_{xz}, I_{yz}\}$; the 3 principal moments of inertia $I = \{I_k : k = 1, 2, 3\}$; the graph radius g_R; and the set of 9 parameters $D = \{D_k^\gamma : k = 1, 2, 3; \ \gamma = x, y, z\}$. The relevance of these descriptors was assessed using the Boruta algorithm, which employs Breiman's random forest concept to compute normalized importance.

Recently, we extended the 3D-Dynamic Representation of DNA/RNA Sequences to a four-dimensional method, applying it to the bioinformatics characterization of the SARS-CoV-2 virus [9] and to studies on the genetic diversity of Echinococcus multilocularis in red foxes in Poland [10]. In particular, the distribution of clusters in the classification maps generated using the 4D-Dynamic Representation of DNA/RNA sequences supports the hypothesis that SARS-CoV-2 may have originated in bats and pangolins [9].

III. Conclusion

In summary, 3D-Dynamic Representation of DNA/RNA Sequences allows for both graphical and numerical comparisons of the sequences, with enhanced classification effectiveness when combined with the random forest algorithm [4]. It is especially important to focus on developing tools that could be used to characterize unidentified viruses. In the future, we plan to evaluate the explainability of the systems under consideration using new descriptors of 3D-dynamic graphs, such as those that describe the direction of the sum of eigenvectors in the 3D space.

References

[1] P. Wąż and D. Bielińska-Wąż, "3D-dynamic representation of DNA sequences", J. Mol. Model. vol. 20, Art. No. 2141, 2014.

[2] P. Wąż and D. Bielińska-Wąż, "Non-standard similarity/dissimilarity analysis of DNA sequences", Genomics vol. 104, pp. 464–471, 2014.

[3] D. Bielińska-Wąż, D. Panas, and P. Wąż, "Dynamic representations of biological sequences", MATCH Commun. Math. Comput. Chem. vol. 82, pp. 205–218, 2019.

[4] D. Bielińska-Wąż, P. Wąż, and D. Panas, "Applications of 2D and 3D-Dynamic Representations of DNA/RNA Sequences for a Description of Genome Sequences of Viruses", Comb. Chem. High T. Scr. vol. 25, pp. 429–438, 2022.

[5] 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.

[6] A. Nandy, M. Harle, and S. C. Basak, "Mathematical descriptors of DNA sequences: development and applications", Arkivoc vol. ix, pp. 211–238, 2006.

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

[8] S. Mizuta, "Graphical Representation of Biological Sequences", In Bioinformatics in the Era of Post Genomics and Big Data; I.Y. Abdurakhmonov, Ed.; IntechOpen: London, UK, 2018.

[9] D. Bielińska-Wąż and P. Wąż, "Non-standard bioinformatics characterization of SARS-CoV-2", Comput. Biol. Med. vol. 131, Art. No. 104247, 2021.

[10] D. Bielińska-Wąż, P. Wąż, A. Lass, and J. Karamon, "4D-Dynamic Representation of DNA/RNA Sequences: Studies on Genetic Diversity of *Echinococcus Multilocularis* in Red Foxes in Poland", Life vol. 12, Art. No. 877, 2022.