# A Four-Dimensional Mathematical 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*—This document reviews the 4D-Dynamic Representation of DNA/RNA Sequences, a four-dimensional bioinformatics model developed and published by the authors for classifying deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) sequences.**

***Keywords–bioinformatics; alignment-free methods; descriptors***

## I. Introduction

Alignment-free bioinformatics methods represent a rapidly developing field and provide an efficient alternative to traditional alignment-based approaches [1]–[6]. Among alignment-free methods are Graphical Representations of Biological Sequences, which focus on the graphical and numerical analysis as well as the classification of biological sequences, with comprehensive reviews available in [7]–[10]. Each approach highlights different aspects of sequence similarity. This document describes the 4D-Dynamic Representation of DNA/RNA Sequences, an alignment-free method developed and published by the authors [11] [12]. While multidimensional in nature, the projections of its four-dimensional graphs into two- and three-dimensional spaces serve as graphical tools for analyzing sequence similarity. This method extends our earlier two-dimensional [13]–[19] and three-dimensional [20]–[23] approaches. We refer to this method as "dynamic" because it models sequences as clouds of material points, maintaining constant distances between one another, similar to the behavior of a rigid body in classical dynamics. The numerical characteristics of these clouds are designed to be analogous to those used in dynamics. Details of the method are provided in Section II, while the summary and our plans are presented in Section III.

## II. Method and Results

In this approach, a DNA/RNA sequence is represented as a *4D-dynamic graph* - a collection of material points with unit masses positioned in four-dimensional space. The methodology for constructing this graph is detailed in [11]. As descriptors of the 4D-Dynamic Representation of DNA/RNA Sequences, we proposed using the coordinates of the centers of mass of these 4D-dynamic graphs

$$\mu^k = \frac{1}{N}\sum_{i=1}^{N} x_i^k, \quad k = 1, 2, 3, 4 \tag{1}$$

and using the normalized principal moments of inertia of the 4D-dynamic graphs

$$r_k^{4D} = \sqrt{\frac{I_k}{N}}. \tag{2}$$

$N$ is the length of the sequence which is equal to the total mass of the 4D-dynamic graph ($m_i = 1$ of each material point):

$$N = \sum_{i=1}^{N} m_i. \tag{3}$$

The moment of inertia tensor in four-dimensional space is:

$$\hat{I} = \begin{pmatrix} I_{1\,1} & I_{1\,2} & I_{1\,3} & I_{1\,4} \\ I_{2\,1} & I_{2\,2} & I_{2\,3} & I_{2\,4} \\ I_{3\,1} & I_{3\,2} & I_{3\,3} & I_{3\,4} \\ I_{4\,1} & I_{4\,2} & I_{4\,3} & I_{4\,4} \end{pmatrix} \tag{4}$$

with the matrix elements

$$I_{j\,j} = \sum_{i=1}^{N} m_i \sum_{k=1}^{4} \left[\hat{x}_i^k (1 - \delta_{jk})\right]^2, \tag{5}$$

$$I_{j\,k} = I_{k\,j} = -\sum_{i=1}^{N} m_i \hat{x}_i^j \hat{x}_i^k. \tag{6}$$

The $m_i$ coordinates in the new coordinate system are $\hat{x}_i^k = x_i^k - \mu^k$ and $\delta_{jk}$ is the Kronecker-Delta. The new system is a Cartesian coordinate system in which the origin is chosen at the center of mass of the graph. The eigenvalues $I_k$ used in (2), called the principal moments of inertia, are obtained by solving the fourth-order secular equation:

$$\det(\hat{I} - I\hat{E}) = 0, \tag{7}$$

where $\hat{E}$ is the unit matrix.

The proposed descriptors have been applied to the construction of the similarity maps. These descriptors are represented on the axes of the maps. Examples of the similarity maps obtained using different descriptors are shown in Figures 1–3.

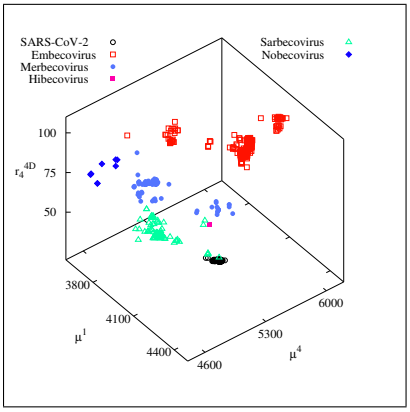


Figure 1. Classification map $\mu^1 - \mu^4 - r_4^{4D}$.

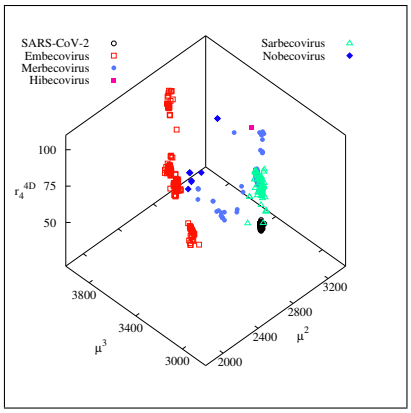


Figure 2. Classification map $\mu^2 - \mu^3 - r_4^{4D}$.

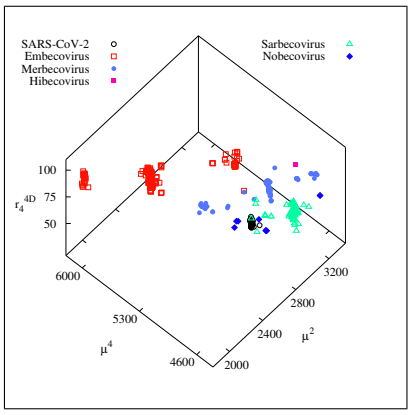


Figure 3. Classification map $\mu^2 - \mu^4 - r_4^{4D}$.

## III. Conclusion

We applied this approach to the bioinformatics characterization of the SARS-CoV-2 virus [11] and to studies on the genetic diversity of *Echinococcus multilocularis* in red foxes in Poland [12]. Specifically, 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 [11]. Our results align with those obtained using standard bioinformatics methods for the *Echinococcus multilocularis* [24] and SARS-CoV-2 [25] datasets. In our future work, we plan to introduce new graph descriptors and apply them to the characterization of other viruses.

## References

[1] M.K. Gupta, R. Niyogi, and M. Misra, "An alignment-free method to find similarity among protein sequences via the general form of Chou's pseudo amino acid composition", SAR QSAR Environ. Res. vol. 24, pp. 597–609, 2013.

[2] Y.S. Li, T. Song, J.S. Yang, Y. Zhang, and J.L. Yang, "An Alignment-Free Algorithm in Comparing the Similarity of Protein Sequences Based on Pseudo-Markov Transition Probabilities among Amino Acids", PLOS ONE vol. 11, Art. No. e0167430, 2016.

[3] A.K. Saw, B.C Tripathy, and S. Nandi, "Alignment-free similarity analysis for protein sequences based on fuzzy integral", Sci. Rep. vol. 9, Art. No. 2775, 2019.

[4] Y. Zhao, X. Xue, and X. Xie, "An alignment-free measure based on physicochemical properties of amino acids for protein sequence comparison", Comput. Biol. Chem. vol. 80, pp. 10–15, 2019.

[5] N. Ramanathan, J. Ramamurthy, and G. Natarajan, "Numerical Characterization of DNA Sequences for Alignment-free Sequence Comparison - A Review", Comb. Chem. High Throughput Screen. vol. 25, pp. 365–380, 2022.

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

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

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

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

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

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

[23] 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 Throughput Screen. vol. 25, pp. 429–438, 2022.

[24] J. Karamon, K. Stojecki, M. Samorek-Pieróg, E. Bilska-Zając, M. Różycki, E. Chmurzyńska, J. Sroka, J. Zdybel, and T. Cencek, "Genetic diversity of *Echinococcus multilocularis* in red foxes in Poland: The first report of a haplotype of probable Asian origin", Folia Parasitol. vol. 64, Art. No. 007, 2017.

[25] S.K. Gupta, R. Minocha, P.J. Thapa, M. Srivastava, and T. Dandekar, "Role of the Pangolin in Origin of SARS-CoV-2: An Evolutionary Perspective", Int. J. Mol. Sci. vol. 23, Art. No. 9115, 2022.