Application of 3D-Dynamic Representation of DNA/RNA Sequences to a Characterization of the Zika Virus Genome

Piotr Wąż

Department of Nuclear Medicine
Medical University of Gdańsk, Poland
Email: phwaz@gumed.edu.pl

Dorota Bielińska-Wąż

Department of Radiological Informatics and Statistics
Medical University of Gdańsk, Poland
Email: djwaz@gumed.edu.pl

***Abstract*—A new method in bioinformatics, which is referred to as *3D-dynamic representation of DNA sequences* is briefly outlined. The aim of this work is an application of this method to the description of the Zika virus genome. We expect to reveal some new features of the sequences comparing to our previous results obtained by using the 2D-dynamic representation.**

Keywords–Bioinformatics; Alignment-free methods; Descriptors.

I. INTRODUCTION

The problem of classification is related to the problem of similarity of the objects. The objects arranged in simple, one-dimensional sets may be classified in a unique way according to a single aspect of similarity. The problem becomes more complicated if we consider multi-dimensional sets, i.e., objects characterized by several different aspects. The degree of similarity (classification to some particular groups) depends on the set of selected aspects, on the number of aspects considered and on the mathematical measure establishing the relations between different properties.

The aim of the studies is a creation of new bioinformatical models carrying information about similarity of the DNA sequences. This information is relevant for solving many biomedical problems. The inspiration for these studies has interdisciplinary character.

A sequence is defined as a sequence of symbols. In the case of the DNA this is a sequence composed of four letters corresponding to four nucleotides: A - adenine, C - cytosine, G - guanine, T - thymine.

The starting point in the methods mentioned above is a DNA sequence represented as a sequence of four symbols in the $5'$ to $3'$ direction. In the methods called in the literature *Graphical representations of DNA sequences*, the sequence of symbols is represented by graphs. The aim of these methods is the creation of both graphs and descriptors representing DNA sequences in a unique way. It may happen that a method can not distinguish between two or more different DNA sequences. In these cases, a nonuniqueness appears, which means that several different sequences are represented by the same graph or by the same descriptor. This kind of nonuniqueness in graphical bioinformatics is called in the English literature *degeneracy*.

Nonuniqueness of the description (degeneracy) is an undesired feature of the method. Removing this feature may be difficult in graphical representation methods. The graphs representing DNA sequences are plotted in two or three-dimensional space. The sequences are long and are composed of four different bases. A reduction of such complicated objects to simple, small, two or three-dimensional graphs corresponding to the perceptional abilities of humans, without a significant loss of information, is difficult and often leads to degeneracy. It is also not obvious how to assign nondegenerate descriptors to such graphs. One of the achievements of the present work is the construction of methods which are either free of this uncertainty or the remaining uncertainty is much lower than in the other formerly known methods.

Recently, we have proposed a new method of comparison of DNA/RNA sequences, called by us *3D-dynamic representation of DNA/RNA sequences* [1] [2]. This method belongs to a group of methods in bioinformatics called *graphical representation methods* (See for reviews [3]–[5]). These methods allow for both graphical and numerical comparison of the considered objects. The sequences are very long, and it is not obvious how to represent them graphically. Each method reveals different aspects of similarity, and therefore new approaches are created.

The aim of this work is an application of this method to the description of the Zika virus genome. We expect to reveal some new features of the sequences comparing to our previous results obtained by using the 2D-dynamic representation.

II. METHOD AND EXPECTED RESULTS

In *3D-dynamic representation of DNA sequences* method, the sequence is represented by a set of material points in a 3D space [1] [2]. The way of construction of the 3D-dynamic graph is described in [1]. The examples of 3D dynamic graphs representing histone H1 coding sequences of plants and of vertebrates are shown in Figures 1 and 2, respectively. The starting point of 3D-dynamic graphs is the origin of the coordinate system – the coordinates of this points are zeros. Each of the bases is represented by a basis vector. Therefore the dimensions of the graphs and their location contain information about the number of particular bases and about their distribution in DNA sequences.

The name of this method (*3D-dynamic representation of DNA sequences*) is related to the numerical characteristics of the graphs (called in the theory of molecular similarity *descriptors*), which are analogous to the ones used in the dynamics, i.e., coordinates of the centers of mass of the graphs, and moments of inertia of the graphs.

 ISBN: 978-1-61208-571-5

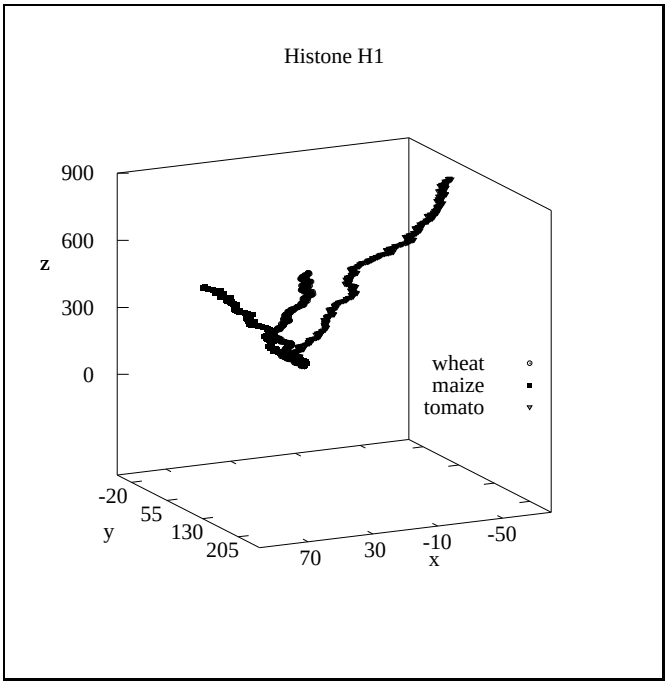

Figure 1. 3D-dynamic graphs.

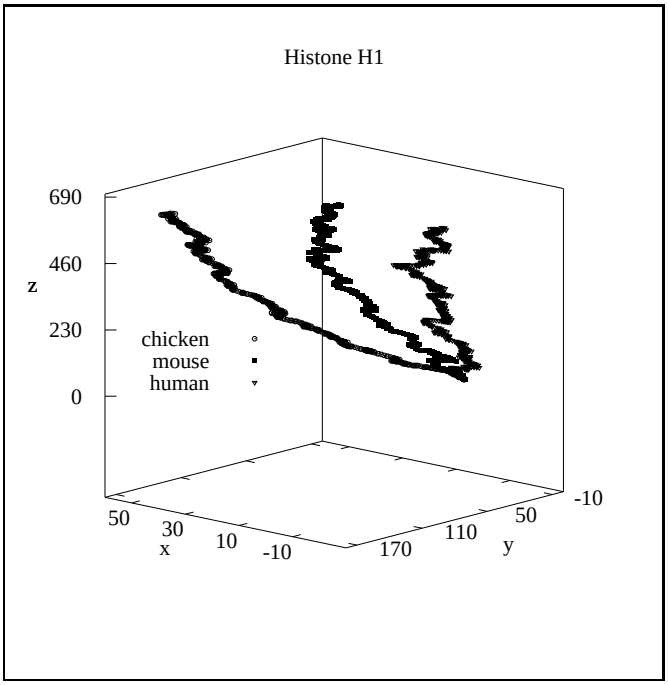

Figure 2. 3D-dynamic graphs.

The coordinates of the center of mass of the 3D-dynamic graph, in the $\{X, Y, Z\}$ coordinate system are defined as [1]

$$\mu_x = \frac{\sum_i m_i x_i}{\sum_i m_i}, \quad \mu_y = \frac{\sum_i m_i y_i}{\sum_i m_i}, \quad \mu_z = \frac{\sum_i m_i z_i}{\sum_i m_i}, \quad (1)$$

where x_i, y_i, z_i are the coordinates of the mass m_i. Since $m_i = 1$ for all the points, the total mass of the sequence is $N = \sum_i m_i$, where N is the length of the sequence. Then, the coordinates of the center of mass of the 3D-dynamic graph may be expressed as

$$\mu_x = \frac{1}{N}\sum_i x_i, \quad \mu_y = \frac{1}{N}\sum_i y_i, \quad \mu_z = \frac{1}{N}\sum_i z_i. \quad (2)$$

The tensor of the moment of inertia is given by the matrix

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

where the particular matrix elements are defined in [1]. The eigenvalue problem of the tensor of inertia is defined as

$$\hat{I}\omega_k = I_k\omega_k, \quad k = 1, 2, 3, \quad (4)$$

where I_k are the eigenvalues and ω_k are the eigenvectors. The eigenvalues I_1, I_2, I_3 are called the principal moments of inertia. As the descriptors we select the square roots of the normalized principal moments of inertia:

$$r_1 = \sqrt{\frac{I_1}{N}}, \quad r_2 = \sqrt{\frac{I_2}{N}}, \quad r_3 = \sqrt{\frac{I_3}{N}}. \quad (5)$$

III. Conclusion and Future Work

In the present work we describe the sequences of the Zika virus genome using 3D-Dynamic Representation of DNA/RNA Sequences. Recently, we have obtained some correlations of the descriptors with time using 2D-dynamic representation of DNA/RNA sequences [6]. Using the present method, some new features of the considered objects are revealed.

Summarizing, using graphical representation methods different aspects of similarity of the DNA sequences, can be considered separately. Only simple objects can be classified in a unique way in terms of their similarity. A pair of complex objects can be similar in one aspect and very different in another one. Using these methods one can indicate properties which are identical or very different for the same pair of the DNA sequences. Graphical representations of DNA sequences constitute both numerical and graphical tools for similarity/dissimilarity analysis of DNA sequences. They can be applied for solving a large class of problems in biology and medical sciences that require such an analysis.

References

[1] P. Wąż and D. Bielińska-Wąż, "3D-dynamic representation of DNA sequences", J. Mol. Model. vol. 20, 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] A. Nandy, M. Harle, and S. C. Basak, "Mathematical descriptors of DNA sequences: development and applications", Arkivoc ix, pp. 211–238, 2006.

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

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

[6] 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.

 ISBN: 978-1-61208-571-5