A New Computational Method of Comparison of DNA Sequences

Piotr Wąż

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

Dorota Bielińska-Wąż

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

***Abstract*—A new method of comparison of deoxyribonucleic acid (DNA) sequences, 3D-dynamic representation of DNA sequences, is presented. This method allows for both graphical and numerical similarity/dissimilarity analysis of the sequences. This method is a generalization of our previous method called by us 2D-dynamic representation of DNA sequences. The methodology is taken from physics: the DNA sequence is represented by a set of "material points" in a 3D space. Using this nonstandard approach we have obtained high accuracy: a difference in a single base can be recognized. We can indicate which base it is (cytosine, guanine, adenine, or thymine) and its approximate location in the DNA sequence.**

Keywords–Bioinformatics; Alignment-free methods; Descriptors.

I. INTRODUCTION

The aim of the presented studies is the 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 main idea in our work is an application of methodological concepts derived from the classical mechanics to bioinformatics. The application of concepts of classical mechanics to the classification of biochemical objects, in particular to the formulation of new criteria determining the degree of similarity of DNA sequences led to the creation of new methods.

II. THEORY AND RESULTS

In this section, we briefly review a new method in bioinformatics which is referred to as *3D-dynamic representation of DNA sequences* [1][2]. The name of this method is related to the descriptors which are analogous as the ones used in the dynamics. The method used to create a 3D-dynamic graph was described in [1]. Two examples of such graphs are shown in Fig. 1.

This method belongs to the group of methods in bioinformatics called *graphical representation methods* (See for reviews [3][4]).

The correctness of these kind of methods is usually shown using standard sets of data: β-globin and histone H4 coding sequences of different species.

As the descriptors (numerical characteristics) of 3D-dynamic graphs, we have proposed [1] the followings:

- Coordinates of the centers of mass of the graphs (μ_x, μ_y, μ_z),
- Normalized principal moments of inertia of the graphs (r_1, r_2, r_3),
- The values of the cosines of properly defined angles.

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

$$I_{xx} = \sum_i^N m_i \left[(y_i')^2 + (z_i')^2\right], \quad I_{yy} = \sum_i^N m_i \left[(x_i')^2 + (z_i')^2\right],$$

$$I_{zz} = \sum_i^N m_i \left[(x_i')^2 + (y_i')^2\right], \quad I_{xy} = I_{yx} = -\sum_i m_i x_i' y_i', \quad (4)$$

$$I_{xz} = I_{zx} = -\sum_i m_i x_i' z_i', \quad I_{yz} = I_{zy} = -\sum_i m_i y_i' z_i'.$$

x_i', y_i', z_i' are the coordinates of m_i in the Cartesian coordinate system for which the origin has been selected at the center of mass.

The eigenvalue problem of the tensor of inertia is defined as

$$\hat{I}\omega_k = I_k \omega_k, \quad k = 1, 2, 3, \quad (5)$$

 ISBN: 978-1-61208-513-5

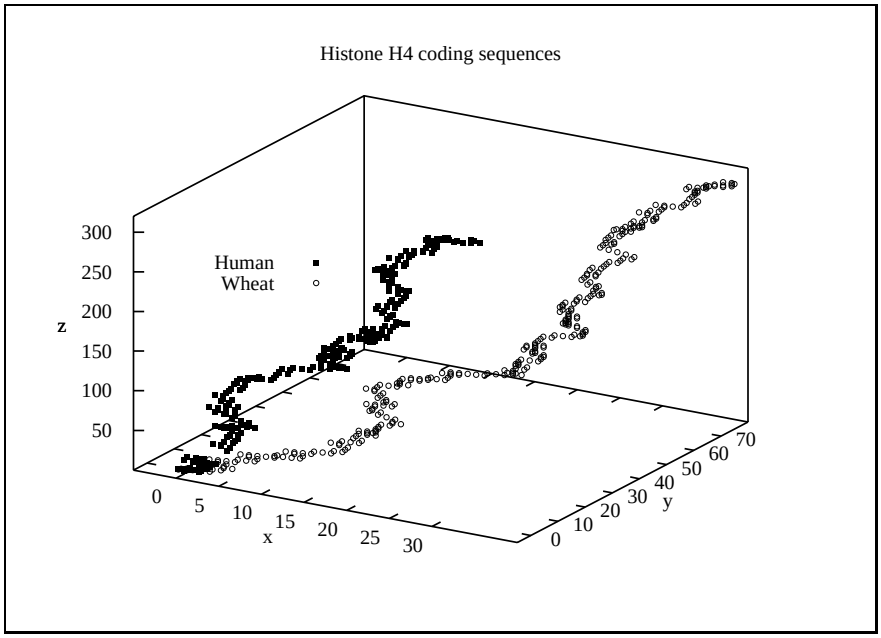

Figure 1. 3D-dynamic graphs.

TABLE I. SIMILARITY/DISSIMILARITY MATRIX BASED ON $D = \frac{\mu_z}{r_3}$ FOR THE SECOND EXON OF β-GLOBIN GENE OF DIFFERENT SPECIES.

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0.0000	0.9419	0.9881	0.9952	0.9990	0.8130	0.9992	0.9886	0.1638	0.9971	0.0050
Goat		0.0000	0.7957	0.9997	0.9826	0.6893	0.9867	0.8044	0.9514	0.9493	0.9416
Opossum			0.0000	0.9999	0.9147	0.9365	0.9349	0.0422	0.9901	0.7519	0.9881
Gallus				0.0000	1.0000	0.9991	1.0000	0.9999	0.9942	1.0000	0.9952
Lemur					0.0000	0.9946	0.2360	0.9110	0.9992	0.6564	0.9990
Mouse						0.0000	0.9959	0.9392	0.8436	0.9843	0.8120
Rabbit							0.0000	0.9320	0.9994	0.7375	0.9992
Rat								0.0000	0.9905	0.7409	0.9886
Gorilla									0.0000	0.9975	0.1680
Bovine										0.0000	0.9970
Chimpanzee											0.0000

where I_k are the eigenvalues and ω_k – the eigenvectors. The eigenvalues are obtained by solving the third-order secular equation

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

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:

$$r_1 = \sqrt{\frac{I_1}{N}}, \quad r_2 = \sqrt{\frac{I_2}{N}}, \quad r_3 = \sqrt{\frac{I_3}{N}}. \qquad (7)$$

Using this approach one can calculate similarity values between the DNA sequences. For this purpose, we have introduced the similarity measure [2]

$$S^{ij} = 1 - exp(-|D_i - D_j|), \qquad (8)$$

where i and j denote two sequences. The measure is normalized: $0 \leq S \leq 1$. For the descriptors, which are identical in both sequences ($D_i = D_j$) the similarity value $S = 0$.

An example of the calculations using this approach is shown in Table 1. This is the similarity/dissimilarity matrix for the second exon of β-globin gene of different species [2]. Small values of S correspond to large degree of similarity related to the considered descriptor D. As we can see the largest similarity human-other species is for chimpanzee. Such result is obtained for many descriptors for these data.

III. CONCLUSIONS

- 3D-dynamic representation of DNA sequences facilitates both graphical and numerical comparison of DNA sequences.
- The method is sensitive:
  It can recognize a difference in only one base.
- The new normalized similarity measure is a good tool for similarity analysis of DNA sequences.

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

 ISBN: 978-1-61208-513-5