Graphical Bioinformatics – a Tool for the Characterization of Influenza Viruses

Dorota Bielińska-Wąż

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

Piotr Wąż

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

***Abstract*—One of the *Graphical bioinformatics* methods, called *2D-dynamic representation of DNA sequences*, is presented. A new application of this approach is discussed.**

Keywords–Bioinformatics; Alignment-free methods; Descriptors.

I. INTRODUCTION

Bioinformatics is a young, interdisciplinary area of research. It has been created in 1982 and it was related to projects aiming at creation and studying databases containing information about Deoxyribonucleic acid/Ribonucleic acid (DNA/RNA) and protein sequences. Nowadays, bioinformatical studies are defined as analysis and interpretation of these biological data using theoretical methods taken from mathematical, physical and biological sciences.

Deriving the information from the DNA or protein sequences became a modern tool for solving many problems in biology and in medicine. Bioinformatical studies not only constitute an important supplement of experimental works, but in many cases they can replace experiments. It is particularly important if animals have to be used in the experiments and if the experimental studies are expensive and time consuming. Therefore, developing high quality theoretical methods is of particular importance. Fast development in this area of research has been observed during the last two decades.

In particular, we have implemented nonstandard methods which have not been used in bioinformatics so far: ideas borrowed from the dynamics of the rigid body and from the statistical spectroscopy. We have constructed algorithms which can discover differences in a single base for a pair of DNA sequences. We can tell which base it is and indicate its approximate location in the sequence. Our methods can be applied to DNA sequences of an arbitrary length and the calculations can be performed in a short time.

The aim of our studies is both application of our methods to the studies which are directly related to biology and medicine and a construction of new methods which may be useful for studies, for example, the mutations of viruses or the creation of phylogenetic trees. We are going to predict directions of the mutations of the influenza viruses and the Zika virus by combining physico-chemical calculations with similarity analyses of DNA sequences (due to a large practical significance). Different modes of implementation of our methods and searching for methods useful for the description of the dynamics of the mutation processes is another aim of our studies. Such studies are particularly important in designing new kinds of vaccine.

Moreover, we plan to work on the classification of biological structures by finding new similarity properties, defining new descriptors which in a numerical way characterize these properties and on the construction of new methods of comparison of these structures. It is worth mentioning that finding new classification schemes often led to important developments in the understanding of the research area for which these schemes have been created (standard examples are the periodic table of elements in chemistry, the Herzsprug-Russel diagram in astronomy, the systematics of living organisms in biology).

In Section II we present the *Graphical bioinformatics* method called by us *2D-dynamic representation of DNA sequences.*

II. METHOD AND EXPECTED RESULTS

Methods known in the literature as Graphical Representations of DNA/RNA Sequences [1] combine ideas from different areas of science. They enable both graphical and numerical comparisons of DNA sequences. Contrary to the frequently used methods based on the alignment of the sequences [2], graphical representations allow us to consider each aspect of similarity separately. Another advantage of these methods is that they are not demanding computationally.

In this work, we present a graphical representation method introduced by us and called *2D-dynamic Representations of DNA Sequences* [3]. A specific feature of this approach is a simple adaptation of some ideas of the classical dynamics. Examples of 2D-dynamic graphs representing DNA sequence are shown in Figures 1 and 2. As we can see, the shapes of the graphs corresponding to different DNA sequences are different. The issues related to the choice and to the properties of the numerical characteristics of the graphs (descriptors), are also discussed. We have defined the following descriptors of 2D-dynamic graphs:

- Moments of the mass-density distributions,
- Angles between x axis and principal axis of inertia of the graphs,
- Coordinates of the centers of mass of the graphs,
- Principal moments of inertia of the graphs.

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

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

 ISBN: 978-1-61208-597-5

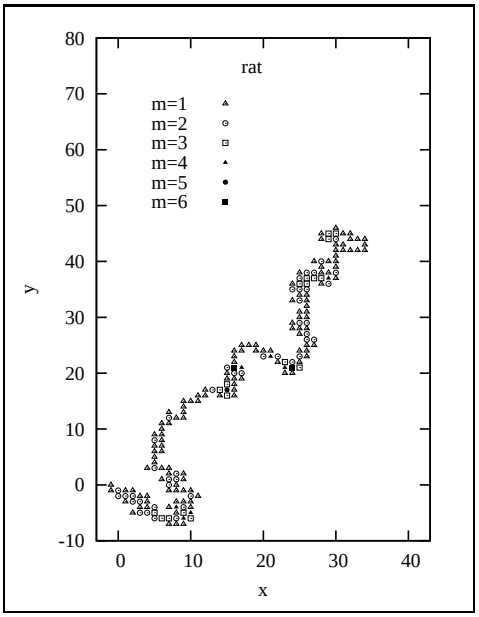

Figure 1. 2D-dynamic graph.

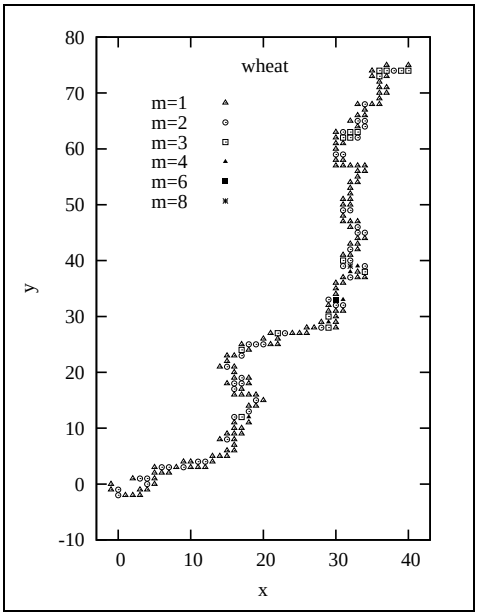

Figure 2. 2D-dynamic graph.

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

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

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

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

Coordinates (μ_x, μ_y) of the centers of mass of the graphs are

$$\mu_a = \frac{1}{N} \sum_i m_i a_i,$$

where $a = x, y$, and the normalization constant

$$N = \sum_i m_i.$$

x_i, y_i are the coordinates of the point mass with mass m_i of the 2D-dynamic graph.

The moment of inertia tensor is

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

where the matrix elements are

$$I_{xy} = I_{yx} = -\sum_i^N m_i x'_i y'_i,$$

$$I_{xx} = \sum_i^N m_i (y'_i)^2,$$

$$I_{yy} = \sum_i^N m_i (x'_i)^2.$$

Principal moments of inertia are the eigenvalues of $\hat{I}$: ω_{11}, ω_{22}. We have shown that the coordinates of the center of mass divided by the principal moments of inertia are good descriptors

$$D_k^\gamma = \frac{\mu_\gamma}{\omega_{kk}},$$

where $\gamma = x, y$ and $k = 1, 2$.

Summarizing,

- 2D-dynamic representation of DNA sequences allows for both graphical and numerical analysis of similarity/dissimilarity of DNA sequences.
- The accuracy is high.
- Negative:
  The history of emergence of a graph is lost since the graphs self-overlap. This point is corrected in 3D-dynamic representation of DNA sequences.

We have already applied this method for a characterization of the Zika virus genome [4]. The aim of the future work is an application of this approach to a characterization of influenza viruses. We expect that the nonstandard method can reveal some new features of the considered objects.

REFERENCES

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

[2] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic Local Alignment Search Tool", J. Mol. Biol. vol. 215, pp. 403–410, 1990.

[3] 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.

[4] 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: 978-1-61208-597-5