Journal of Biomedical Informatics 72 (2017) 1–7

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Spectral-dynamic representation of DNA sequences

Dorota Bielińska-Wąż [a,*], Piotr Wąż [b]

[a] Department of Radiological Informatics and Statistics, Medical University of Gdańsk, Tuwima 15, 80-210 Gdańsk, Poland
[b] Department of Nuclear Medicine, Medical University of Gdańsk, Tuwima 15, 80-210 Gdańsk, Poland

ARTICLE INFO

Article history:
Received 13 February 2017
Revised 3 May 2017
Accepted 1 June 2017
Available online 3 June 2017

Keywords:
Alignment-free methods
Moments of inertia
Similarity/dissimilarity analysis of DNA sequences
Descriptors

ABSTRACT

A graphical representation of DNA sequences in which the distribution of a particular base $B = A, C, G, T$ is represented by a set of discrete lines has been formulated. The methodology of this approach has been borrowed from two areas of physics: spectroscopy and dynamics. Consequently, the set of discrete lines is referred to as the B-spectrum. Next, the B-spectrum is transformed to a rigid body composed of material points. In this way a *dynamic representation* of the DNA sequence has been obtained. The centers of mass of these rigid bodies, divided by their moments of inertia, have been taken as the descriptors of the spectra and, thus, of the DNA sequences. The performance of this method on a standard set of data commonly applied by authors introducing new approaches to bioinformatics (the first exons of β-globin genes of different species) proved to be very good.

1. Introduction

The fast increase of data in DNA databases stimulated the development of computational methods aiming at an analysis of this information.

The most commonly used program (with about 5000 citations each year) for the comparison of primary biological sequence information is BLAST (Basic Local Alignment Search Tool) [1].

A decade ago, Randić and coauthors introduced *graphical alignment* of biosequences methods [2,3]. It is worthwhile to point out that compared to computer based programs on protein alignment, these algorithms do not involve any empirical parameters or approximations (in contrast, the BLAST uses empirical parameters).

Protein Alignment Problem has been very recently solved exactly [4–6]. As we read in Ref. [6] a comparison of the exact solution (based on the use of matrices) and BLAST of two proteins having 170 amino acids, BLAST aligned 89 amino acids, while exact solution aligned 95 (i.e. six more).

Alignment-free methods constitute a fast developing branch of bioinformatics. These methods are of interdisciplinary character and researchers representing different areas of natural sciences bring to them new ideas. Therefore, reports on this subject appear in journals which traditionally have been assigned to physics, chemistry, biology, computer science, and many other fields of science, as for example [7–43]. Reviews may be found in Refs. [44–46].

The basic quantities used in these methods are called by the authors *descriptors*, i.e. some numerical values characterizing the biological (DNA, RNA, protein) sequences [47]. The descriptors can be related to some graphical structures (plots) which give graphical representations of the sequences. As a consequence, the same sequences can be compared both graphically and numerically. This kind of approach may be exemplified by developed by us *2D-Dynamic Representation of DNA Sequences* [48–55] and its 3-dimensional generalization [56,57]. We call these methods "dynamic" because the numerical description of the graphs is based on concepts taken from the classical dynamics. In particular, as one of the descriptors of the 2D-dynamic and 3D-dynamic graphs we introduced properly defined, respectively, 2D and 3D moments of inertia [48,56]. In the 3D case the degeneracy, resulting from the overlapping of the 2D-dynamic graphs, has been removed. Our method has also been generalized to three dimensions by Aram and Iranmanesh [58]. As a consequence, two different methods derived from *2D-Dynamic Representations of DNA Sequences* and having the same name (*3D-Dynamic Representation of DNA Sequences*) [56,58] are present in the literature.

The idea of characterizing biological sequences by moments of inertia, introduced by us for the 2D-dynamic graphs [48], has been adopted in several other methods. In particular, Yao et al. applied this idea representing protein sequences by 2D moments of inertia [59] and by 3D moments of inertia [60]. The 3D moments of inertia have also been applied to characterize graphs representing protein sequences by Hou et al. [61]. Recently, we have also introduced 20D moments of inertia as new characteristics of protein sequences [62].

* Corresponding author.
E-mail addresses: djwaz@gumed.edu.pl (D. Bielińska-Wąż), phwaz@gumed.edu.pl (P. Wąż).

http://dx.doi.org/10.1016/j.jbi.2017.06.001
1532-0464/

In the present work 1D moments of inertia are introduced as new characteristics of DNA sequences. These descriptors are constructed from some specific *spectra* representing the sequences. In a way, the present approach is related to the one used in our recent work on molecular spectra in which 1D moments of inertia have been proposed as new descriptors of infrared molecular spectra [63,64]. The aim of the present article is to demonstrate that a similar methodology can be applied to an arbitrary system of discrete objects, *a spectrum*, in particular to the spectrum representing the DNA sequence.

The new method is a branch of the graphical methods called by the authors *spectral representations* of biological sequences, as for example introduced by us *Four-Component Spectral Representation of DNA Sequences* [65,66]. These methods differ from each other by the mathematical definitions of the spectra and by the descriptors [67–72,65,66,73].

A set of descriptors defined for a single sequence constitute a data basis for this object. In the case of molecules a similar approach results in computational techniques known as Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR) commonly used in computational pharmacology, toxicology, eco-toxicology [74,75]. We have recently demonstrated that 1D moments of inertia of molecular spectra can be applied in protein QSAR studies to predict environmentally relevant properties of chloronaphthalenes [64]. Analogously to the molecular descriptors, the descriptors of biological sequences have found their applications in QSAR studies, for representing other sources of information like mass spectra of blood serum in clinical proteomics and molecular dynamics trajectories [76–84].

The new descriptors of the DNA sequences proposed in this work, composed using the dynamic description of some properly defined abstract spectra, seem to be of importance in the context of their potential use in biomedical sciences. In particular, as it is shown in the present work, they can be applied to hierarchical cluster analysis.

2. Theory

Let us represent the distribution of a particular base in the DNA sequence by a series of lines. The length of each line is equal to 1 and its position in the series corresponds to the position of the base in the sequence. The graphical appearance of such a representation resembles an atomic, molecular, or stellar spectrum composed of a series of sharp spectral lines. Therefore it is referred to as a *spectral representation* of the sequence. In order to make this resemblance closer, the terminology used in spectroscopy has also been introduced.

Thus, the sequence corresponding to the base B, with $\mathrm{B} = \mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}$ is called the *spectrum* of B or the B-spectrum, the position of the ith line in the B-spectrum is denoted v_i^{B}, $i = 1, 2, \ldots N_{\mathrm{B}}$, and called *frequency*.

The length of the line (in our case equal to 1 for all lines) is denoted $I_{\mathrm{B}}(v_i^{\mathrm{B}})$, $i = 1, 2, \ldots N_{\mathrm{B}}$, and called the *intensity* of the line.

Since N_{B} is equal to the number of lines in the B-spectrum, i.e. to the number of B bases in the sequence, we have

$$N_{\mathrm{A}} + N_{\mathrm{C}} + N_{\mathrm{G}} + N_{\mathrm{T}} = N, \tag{1}$$

where N is the length of the DNA sequence. For example, a model sequence ATGGTT is represented by the B-spectra composed of the following lines:

$$I_{\mathrm{A}}(v_1^{\mathrm{A}}) = 1, \qquad v_1^{\mathrm{A}} = 1,$$
$$I_{\mathrm{G}}(v_1^{\mathrm{G}}) = I_{\mathrm{G}}(v_2^{\mathrm{G}}) = 1, \qquad v_1^{\mathrm{G}} = 3, \quad v_2^{\mathrm{G}} = 4,$$
$$I_{\mathrm{T}}(v_1^{\mathrm{T}}) = I_{\mathrm{T}}(v_2^{\mathrm{T}}) = I_{\mathrm{T}}(v_3^{\mathrm{T}}) = 1, \qquad v_1^{\mathrm{T}} = 2, \quad v_2^{\mathrm{T}} = 5, \quad v_3^{\mathrm{T}} = 6,$$

and there are no lines in the C-spectrum.

The B-spectra corresponding to a model sequence ATGACTTTGCTGAGT are shown in Fig. 1.

Let us assume that the spectral lines of the B-spectrum are set vertical. Their projections to the horizontal axis give us four sets of points with the following coordinates:

1. $(v_1^{\mathrm{A}}, 0), (v_2^{\mathrm{A}}, 0), \ldots \left(v_{N_{\mathrm{A}}}^{\mathrm{A}}, 0\right)$,
2. $(v_1^{\mathrm{C}}, 0), (v_2^{\mathrm{C}}, 0), \ldots \left(v_{N_{\mathrm{C}}}^{\mathrm{C}}, 0\right)$,
3. $(v_1^{\mathrm{G}}, 0), (v_2^{\mathrm{G}}, 0), \ldots \left(v_{N_{\mathrm{G}}}^{\mathrm{G}}, 0\right)$,
4. $(v_1^{\mathrm{T}}, 0), (v_2^{\mathrm{T}}, 0), \ldots \left(v_{N_{\mathrm{T}}}^{\mathrm{T}}, 0\right)$.

Let us assign to each point $(v_i^{\mathrm{B}}, 0)$ a mass m_i^{B}. In this way we obtain four massive bodies composed of point masses distributed along the horizontal axis. The moments of inertia of these bodies are equal to

$$M_{\mathrm{B}} = \sum_{i=1}^{N_{\mathrm{B}}} m_i^{\mathrm{B}} \left(\tilde{v}_i^{\mathrm{B}}\right)^2, \tag{2}$$

where

$$\tilde{v}_i^{\mathrm{B}} = v_i^{\mathrm{B}} - v_a^{\mathrm{B}}, \tag{3}$$

and $(v_a^{\mathrm{B}}, 0)$ are the coordinates of the centers of mass of the bodies with

$$v_a^{\mathrm{B}} = \frac{1}{N_{\mathrm{B}}} \sum_{i=1}^{N_{\mathrm{B}}} v_i^{\mathrm{B}}. \tag{4}$$

Hereafter, for simplicity, $m_i^{\mathrm{B}} = 1$ has been set for each point. Consequently, the total mass of a spectrum is equal to the total number of points

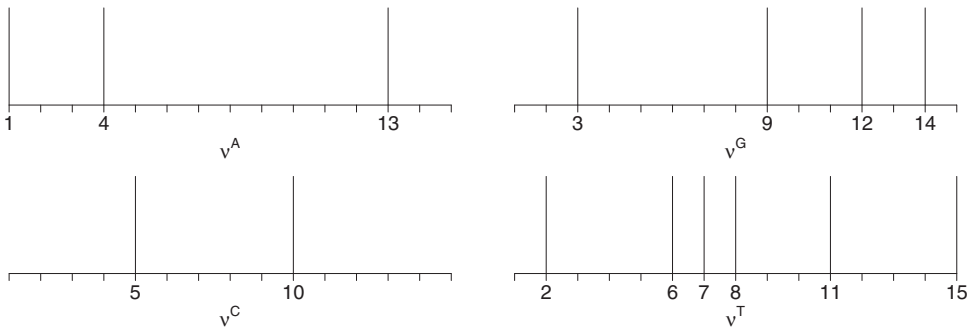

Fig. 1. B-spectra corresponding to a model sequence ATGACTTTGCTGAGT.