J Math Chem (2014) 52:132–140
DOI 10.1007/s10910-013-0249-1

ORIGINAL PAPER

Descriptors of 2D-dynamic graphs as a classification tool of DNA sequences

Piotr Wąż · Dorota Bielińska-Wąż · Ashesh Nandy

Received: 1 May 2013 / Accepted: 16 August 2013 / Published online: 3 September 2013

Abstract A new tool of the classification of DNA sequences is introduced. The method is based on 2D-dynamic graphs and their descriptors. Using the descriptors created by centers of masses, moments of inertia, angles between the *x* axis and the principal axis of inertia of the 2D-dynamic graphs one can obtain classification diagrams in which similar sequences are clustered in separated areas.

Keywords Similarity/dissimilarity analysis of DNA sequences · Moments of inertia · Center of mass · Descriptors · Graphical representations of DNA sequences

1 Introduction

Similarity/dissimilarity analysis of DNA sequences is an important topic in many problems of biology and medicine. Very popular alignment methods often do not give the information detailed enough. For example, they do not distinguish which bases (A, C, T, or G) have been aligned. Recently, we have proposed some corrections to these methods which may enrich the information derived from these methods [1]. Alternatively, one may use methods called *Graphical Representations*. These approaches allow for both visual and numerical comparison of the objects. Recently, a variety of

P. Wąż (✉)
Department of Nuclear Medicine, Medical University of Gdańsk, Tuwima 15, 80-210 Gdańsk, Poland
e-mail: phwaz@gumed.edu.pl

D. Bielińska-Wąż
Department of Radiological Informatics and Statistics, Medical University of Gdańsk, Tuwima 15, 80-210 Gdańsk, Poland

A. Nandy
Centre for Interdisciplinary Research and Education, 404B Jodhpur Park,
Kolkata 700068, India

graphical representations have been created. In particular, easy for visualization 2D-methods are of interest, as for example [2–23]. A review of graphical representation methods may be found in [1,24,25].

The 2D graphical representations of DNA sequences have been used for many applications including phylogenetic relationships of coronavirus gene sequences [26], long range palindromes [27], characterization of avian flu neuraminidase genes [28], study of plant germplasm identificators [29], among others. The first order descriptors had also been used to propose a scale for grading toxicity of chemicals [30] and classification of SNP genes [31]. In this paper we propose to use the improved descriptors in our 2D-dynamic representation model to construct classification diagrams.

The studies on the classification of DNA sequences are the continuation of our earlier works [1,9,32] where we have constructed classification diagrams based on some other methods. In these diagrams different groups of the sequences are located in different parts of the plots. We have also shown that using these diagrams, sequences which differ by only one base can be distinguished [1].

2 Methods

In the present work we use a graphical representation of DNA sequences called by us 2D-dynamic representation. In this method a DNA sequence is represented by a 2D-graph described in [8]. The name of the representation comes from another area of science due to form of the descriptors (numerical characteristics) representing these graphs. The sequence is represented by material-like points which are treated as rigid bodies as in the Newtonian dynamics. We have proposed several descriptors related to the 2D-dynamic graphs: centers of mass [8], moments of inertia of the graphs [8], moments of the mass-density distribution [33,34], angles between the x axis and the principal axis of inertia of the graphs [35]. These descriptors were the basis for the creation of similarity measures between the sequences. We also performed a similarity/dissimilarity analysis using mass-overlaps of the 2D-dynamic graphs [35].

We have used a similar methodology, based on the distribution moments, which aims at the classification studies in other areas of science, as molecular physics [36–40], astrophysics [41,42], and dynamics [43,44].

In the present work we construct the classification diagrams using the descriptors built of the coordinates of the centers of the mass (μ_x, μ_y), the principal moments of inertia (I_{11}, I_{22}), and of the angles between the x axis and the principal axis of inertia of the 2D-dynamic graphs (α).

We define the coordinates of the center of the mass in the same way as it is in the dynamics:

$$\mu_x = \frac{\sum_i m_i x_i}{\sum_i m_i}, \tag{1}$$

$$\mu_y = \frac{\sum_i m_i y_i}{\sum_i m_i}, \tag{2}$$

where x_i, y_i are the coordinates of the mass m_i in the Cartesian coordinate system for which the point (0,0) is the origin, the same for all the sequences. The total mass of the graph is equal to the sum of the masses:

$$N = \sum_i m_i, \tag{3}$$

i.e. it is equal to the length of the sequence.

Also the moment of inertia tensor is defined as in the dynamics (see also [8]):

$$\hat{I} = \begin{pmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{pmatrix}, \tag{4}$$

where

$$I_{xy} = I_{yx} = -\sum_i m_i x_i^{\mu} y_i^{\mu}, \tag{5}$$

$$I_{xx} = \sum_i m_i \left(y_i^{\mu}\right)^2, \tag{6}$$

$$I_{yy} = \sum_i m_i \left(x_i^{\mu}\right)^2, \tag{7}$$

and x_i^{μ}, y_i^{μ} are the coordinates of the mass m_i in the Cartesian coordinate system for which the origin has been selected at the center of the mass.

Solutions $I = I_{11}, I_{22}$ of the second-order equation

$$\begin{vmatrix} I_{xx} - I & I_{xy} \\ I_{xy} & I_{yy} - I \end{vmatrix} = 0 \tag{8}$$

are referred to as the *principal moments of inertia*. They are equal to the moments of inertia associated with the rotations about the principal axes. The principal axes are defined by the Eigenvectors of the tensor of inertia.

Let us define the descriptors

$$D_k^{\gamma} = \frac{\mu_{\gamma}}{I_{kk}}, \tag{9}$$

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

The descriptors are related to some particular properties of the graphs and their interpretation is very intuitive and analogous as it is in dynamics. The moments of inertia are associated with the rotations about the principal axes. If the mass is concentrated close to the axis of rotation, the moment of inertia is small and it is easier to accelerate the spinning of the body. If the mass is dispersed, the moment of inertia is large and the acceleration of spinning is more difficult. Thus, these descriptors carry the information about the concentrations of masses around the axes.

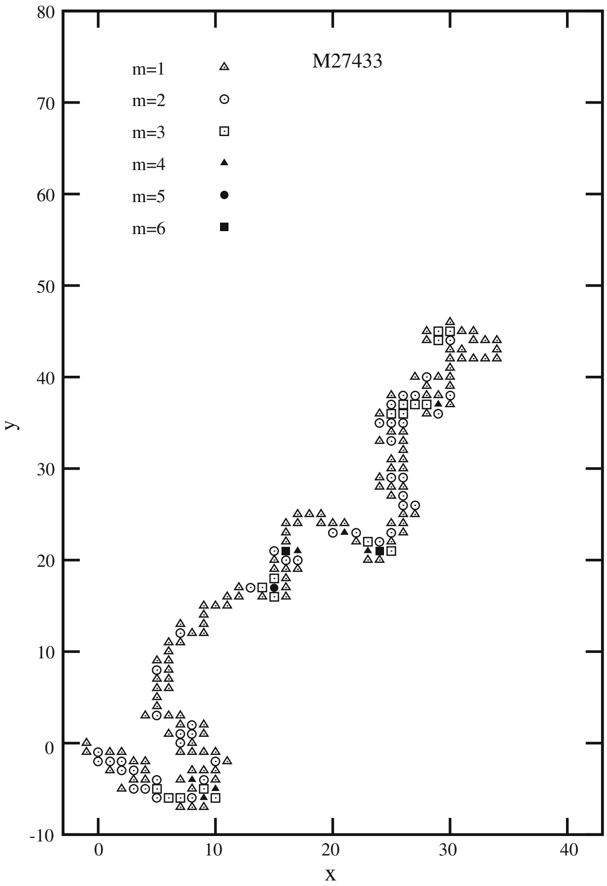

Fig. 1 2D-dynamic graph of histone H4 coding sequence of rat (M27433)

The location of the center of the mass of the 2D-dynamic graph depends on the number of particular bases in the sequence. Each base is represented by a 2D unit vector in the (x, y) plane: A = (−1,0), G = (1,0), C = (0,1), T = (0,−1). Since the graph is obtained using a method of walk in 2D space [8] the location of the graph depends on the relative number of particular bases. Thus, if for example, the number of A bases is larger than the number of G bases then the graph is shifted towards the negative x values by the appropriate amount.

In the present work we consider diagrams $D_k^{\gamma} - D_l^{\beta}, k \neq l$. We show that using these diagrams one can classify different groups of DNA sequences. We also use a diagram $\alpha - I_{22}$ for some detailed classification (see subsequent section).

3 Results and discussion

The descriptors for histone H4 coding sequences and alpha globin coding sequences of different species have been calculated using the values of centers of masses, moments

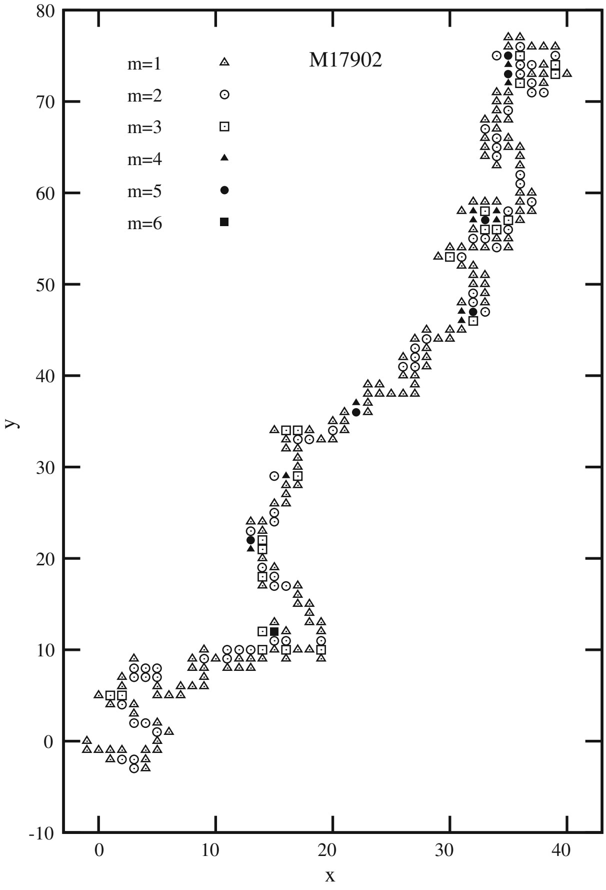

Fig. 2 2D-dynamic graph of alpha globin coding sequence of horse (M17902)

of inertia, and angles α between x axis and the principal axes of the 2D-dynamic graphs obtained in our earlier works [8,35]. The descriptors have been used to the construction of the classification diagrams.

Some examples of the 2D-dynamic graphs for the sequences under consideration are shown in Figs. 1, 2.

Figures 3, 4 show the classification diagrams based on the descriptors defined in Eq. 9 of the 2D-dynamic graphs. We observe that the descriptors corresponding to histone H4 coding sequences are located in different parts of the diagrams than the descriptors corresponding to alpha globin coding sequences. Each point corresponds to a different species (for details about histone H4 coding sequences see [1] and about alpha globin coding sequences see [8]).

Figure 5 shows a classification diagram $\alpha - I_{22}$ which allows to classify the sequences of evolutionary similar organisms: the descriptors corresponding to histone H4 coding sequences of plants are located in the upper part of the diagram and the ones of vertebrates in the lower part. A similar classification analysis of DNA

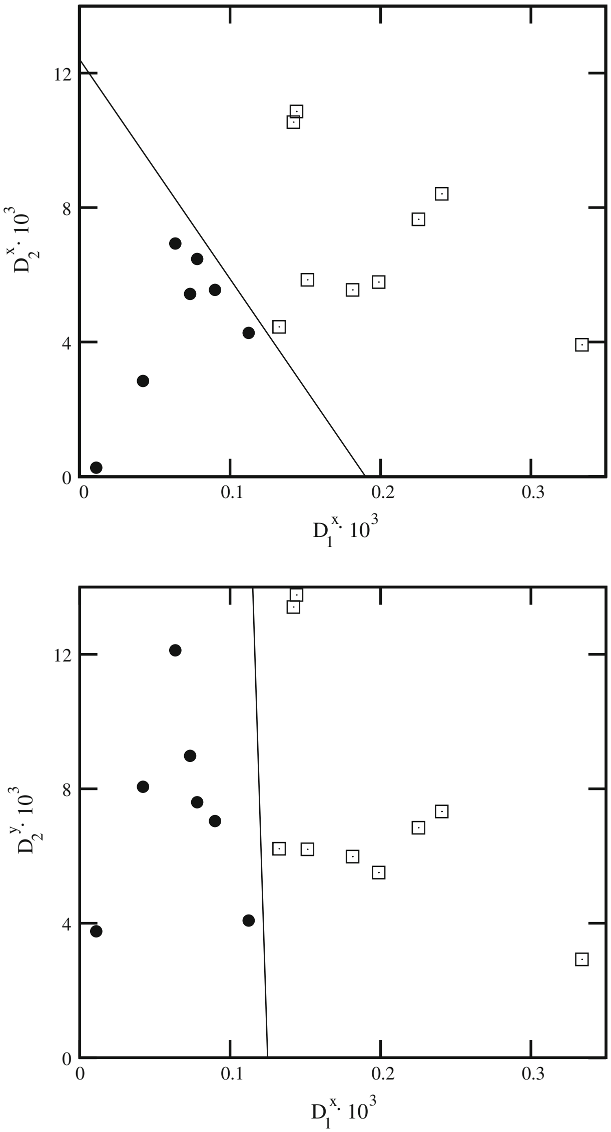

Fig. 3 Classification diagrams: $D_1^x - D_2^x$ (*top*) and $D_1^x - D_2^y$ (*bottom*). *Squares* correspond to histone H4 coding sequences and *circles* correspond to alpha globin coding sequences

sequences of plants and of vertebrates we have already performed using moments of the mass-density distributions of the 2D-dynamic graphs [1] and using the descriptors of Four-Component Spectral Representation of DNA sequences [1,32].

Summarizing, a variety of graphical representations of DNA sequences gives an opportunity of considering different properties of the sequences. Different aspects of similarity can be compared. It is interesting, that the ideas brought from different

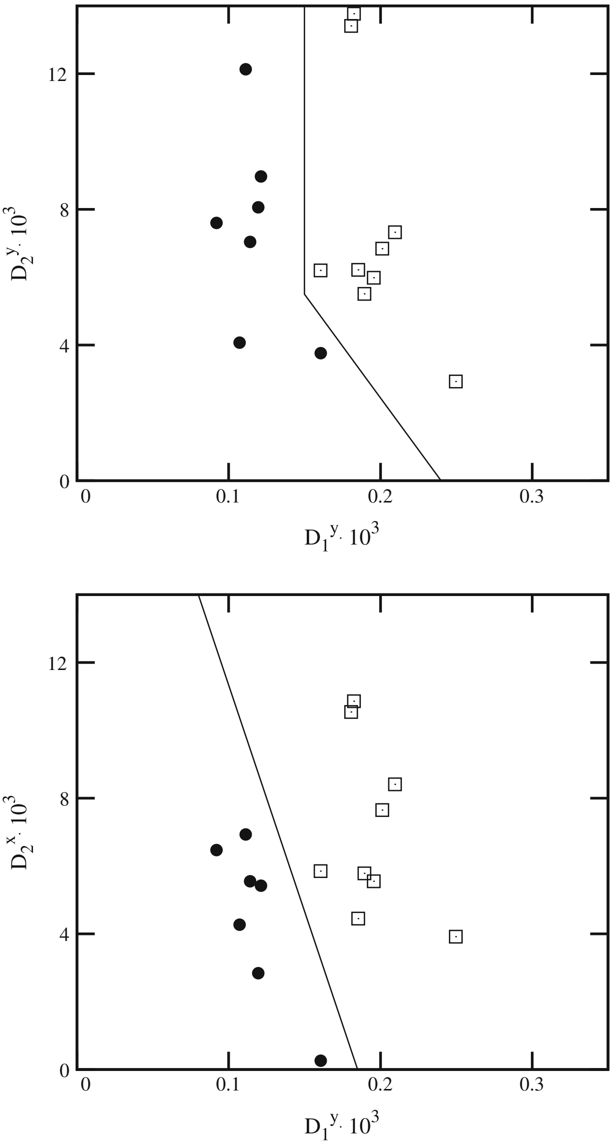

Fig. 4 Classification diagrams: $D_1^y - D_2^y$ (*top*) and $D_1^y - D_2^x$ (*bottom*). The *symbols* are the same as in Fig. 3

areas of science can be mixed and, in effect, we can reveal various aspects of similarity of the DNA sequences. In particular, Four-Component Spectral Representation [9] only visually resembles molecular spectrum. Also 2D-dynamic graphs are not real dynamical objects. However, using methods and terminology from other fields one can obtain a convenient and intuitive classification tool of the DNA sequences.

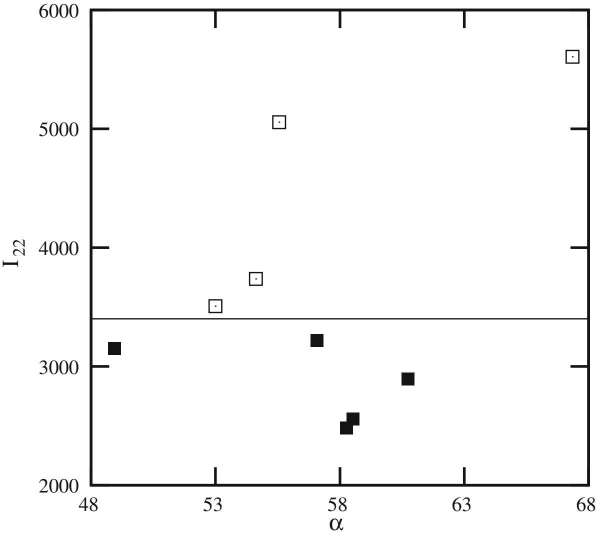

Fig. 5 $\alpha - I_{22}$ classification diagram. *Full squares* correspond to histone H4 coding sequences of vertebrates and *empty squares* correspond to histone H4 coding sequences of plants

References

1. D. Bielińska-Wąż, J. Math. Chem. **49**, 2345 (2011)
2. X. Guo, M. Randić, S.C. Basak, Chem. Phys. Lett. **350**, 106 (2001)
3. B. Liao, M. Tan, K. Ding, Chem. Phys. Lett. **414**, 296 (2005)
4. Y. Liu, X. Guo, L. Pan, S. Wang, J. Chem. Inf. Comput. Sci. **42**, 529 (2002)
5. G. Huang, B. Liao, Y. Li, Z. Liu, Chem. Phys. Lett. **462**, 129 (2008)
6. G. Huang, B. Liao, Y. Li, Y. Yu, Biophys. Chem. **143**, 55 (2009)
7. C. Li, J. Wang, Internet Electron. J. Mol. Des. **1**, 000 (2003)
8. D. Bielińska-Wąż, T. Clark, P. Wąż, W. Nowak, A. Nandy, Chem. Phys. Lett. **442**, 140 (2007)
9. D. Bielińska-Wąż, J. Math. Chem. **47**, 41 (2010)
10. Z.-J. Zhang, Bioinformatics **25**, 1112 (2009)
11. M. Randić, M. Vračko, N. Lerš, D. Plavsić, Chem. Phys. Lett. **368**, 1 (2003)
12. P.A. Scholes, *The Oxford Companion to Music*, 10th edn. (Oxford University Press, Oxford, UK, 1986)
13. C. Li, J. Wang, Comb. Chem. High Throughput Screen. **6**, 795 (2003)
14. J. Song, H. Tang, J. Biochem. Biophys. Methods **63**, 228 (2005)
15. B. Liao, T. Wang, J. Comput. Chem. **25**, 1364 (2004)
16. J. Wang, Y. Zhang, Chem. Phys. Lett. **423**, 50 (2006)
17. Y. Yao, T. Wang, Chem. Phys. Lett. **398**, 318 (2004)
18. M. Randić, Chem. Phys. Lett. **456**, 84 (2008)
19. H.I. Jefrey, Nucleic Acids Res. **18**, 2163 (1990)
20. H.I. Jefrey, J. Comput. Graph. **16**, 25 (1992)
21. M. Randić, M. Vračko, J. Zupan, M. Novič, Chem. Phys. Lett. **373**, 558 (2003)
22. M. Randić, Chem. Phys. Lett. **386**, 468 (2004)

23. M. Randić, N. Lerš, D. Plavsić, S.C. Basak, A.T. Balaban, Chem. Phys. Lett. **407**, 205 (2005)
24. A. Nandy, M. Harle, S.C. Basak, ARKIVOC **ix**, 211 (2006)
25. H. González-Diaz, L. Santana, E. Uriarte, Curr. Top. Med. Chem. **7**, 1025 (2007)
26. B. Liao, Y. Liu, R. Li, W. Zhu, Chem. Phys. Lett. **421**, 313 (2006)
27. S. Larionov, A. Loskutov, E. Ryadchenko, Chaos **18**, 013105 (2008)
28. A. Nandy, S.C. Basak, B.D. Gute, J. Chem. Inf. Model. **47**, 945 (2007)
29. I. Wiesner, D. Wiesnerová, Biologia Plantarum **54**, 353 (2010)
30. A. Nandy, S.C. Basak, J. Chem. Inf. Comput. Sci. **40**, 915 (2000)
31. A. Nandy, P. Nandy, S.C. Basak, Internet Electron. J. Mol. Des. **1**, 367 (2002)
32. D. Bielińska-Wąż, S. Subramaniam, J. Theor. Biol. **266**, 667 (2010)
33. D. Bielińska-Wąż, W. Nowak, P. Wąż, A. Nandy, T. Clark, Chem. Phys. Lett. **443**, 408 (2007)
34. D. Bielińska-Wąż, P. Wąż, W. Nowak, A. Nandy, S.C. Basak, in *AIP Conference Proceedings 963*, ed. by T.E. Simos, G. Maroulis (New York, 2007), pp. 28–30
35. D. Bielińska-Wąż, P. Wąż, T. Clark, Chem. Phys. Lett. **445**, 68 (2007)
36. D. Bielińska-Wąż, P. Wąż, S.C. Basak, Eur. Phys. J. B **50**, 333 (2006)
37. D. Bielińska-Wąż, P. Wąż, S.C. Basak, J. Math. Chem. **42**, 1003 (2007)
38. D. Bielińska-Wąż, P. Wąż, J. Math. Chem. **43**, 1287 (2008)
39. D. Bielińska-Wąż, W. Nowak, Ł. Pepłowski, P. Wąż, S.C. Basak, R. Natarajan, J. Math. Chem. **43**, 1560 (2008)
40. D. Bielińska-Wąż, P. Wąż, T. Clark, T. Puzyn, Ł. Pepłowski, W. Nowak, J. Math. Chem. **51**, 857 (2013)
41. P. Wąż, D. Bielińska-Wąż, A. Pleskacz, A. Strobel, Acta Phys. Pol. B. **39**, 1993 (2008)
42. P. Wąż, D. Bielińska-Wąż, A. Strobel, A. Pleskacz, Acta Astron **60**, 283 (2010)
43. P. Wąż, D. Bielińska-Wąż, Acta Phys. Pol. A **116**, 987 (2009)
44. P. Wąż, D. Bielińska-Wąż, Acta Phys. Pol. A **123**, 647 (2013)