J Math Chem (2010) 47:41–51
DOI 10.1007/s10910-009-9535-3

ORIGINAL PAPER

Four-component spectral representation of DNA sequences

Dorota Bielińska-Wąż

Received: 14 February 2008 / Accepted: 22 January 2009 / Published online: 12 February 2009

Abstract A 2D-graphical representation of DNA sequences is presented. In this representation, the DNA sequence is represented by a four-component (A, C, T, G) spectrum taken as a superposition of Gaussian functions. The maxima of single Gaussians are determined by the positions of the bases in the sequence. Distribution moments of the four-component functions are proposed as descriptors of DNA sequences.

Keywords Descriptors · DNA sequences · Distribution moments

1 Introduction

Recently, a rapid growth of sequence data in DNA databases has been observed, resulting in over 100 billion bases in the DNA sequence databanks by 2005 [1]. Consequently, designing mathematical tools that aim at a quick identification or for similarity studies between sequences have become an urgent necessity. Very useful are the methods based on graphical representations [2]. One of the aims of graphical representations is to identify regions of interest or the distribution of bases along the sequence visually. The original 2D-graphical methods for representing a gene sequence are easy to visualize but, regretfully, they lead to significant degeneracy [3–5]. Recently, we proposed a 2D-dynamic representation of DNA sequences that removed many of these degeneracies [6]. Among other solutions one should mention multidimensional, even ranging up to 6D, graphical representations [7–9]. However, the multidimensional methods are difficult to visualize. In general, when the number of data is large, it can be easier to deal with mathematical descriptors that offer a numerical characterization of the graphs. Several approaches defining such descriptors are given in

D. Bielińska-Wąż (✉)
Instytut Fizyki, Uniwersytet Mikołaja Kopernika, Grudziądzka 5, 87-100 Toruń, Poland
e-mail: dsnake@phys.uni.torun.pl

[10–12]. In particular, the descriptors characterizing 2D-dynamic graphs are presented in [6,13,14].

In this paper, an easy to visualize 2D-graphical representation of DNA sequences is proposed. Graphically, it resembles a molecular spectrum and therefore this representation is called *spectral*. Another feature of this representation that refers to spectroscopy is the numerical characterization of the graphs: I propose the "intensity distribution moments" as descriptors of the DNA sequences. The method is very sensitive. As will be shown, even a difference of only one base in a pair between sequences can be visualized graphically in this approach. Another advantage of this method is its flexibility. In particular, the resolution of the graphs can be modified by a proper selection of the input data. A similar approach has been proposed recently in the theory of molecular similarity [15]. The two models, however deal with different physical objects, are methodologically similar since the spectra defined in the present work should be considered as mathematical tools that play the same role as intensity distributions in spectroscopy.

2 Theory

The DNA sequence is represented as an abstract spectrum given by a four-component function:

$$I^{\gamma}(x) = \sum_{i=1}^{N} n_i^{\gamma} \exp[-(x - \epsilon_i)^2], \tag{1}$$

where $\gamma = A, C, T, G$ stand for the bases of the sequence, adenine (A), cytosine (C), thymine (T) and guanine (G), x is the variable running along the sequence, and ϵ_i describes the position of the bases in the sequence and the occupation number of the base γ in the sequence

$$n_i^{\gamma} = \begin{cases} 1, & \gamma \text{occupies the } i\text{th position in the sequence,} \\ 0, & \text{otherwise.} \end{cases}$$

Accordingly, a single component of the spectral function is referred to as A, C, T, G spectrum. The ith base is represented by a single Gaussian function with the maximum located at

$$\epsilon_i = (i - 1)r, \quad r > 0 \tag{2}$$

where r is a properly chosen step that determines the differences between the maxima of the Gaussians, $\epsilon_1 = 0$ is the position of the maximum of the first Gaussian and $\epsilon_N = (N - 1)r$ is the position of the last Gaussian, with

$$N = N^A + N^C + N^T + N^G$$

standing for the total number of bases in the sequence. These maxima are localized in one of I^γ spectra. Let us take an example of a model sequence GAAACG. In this case $N^A = 3$, $N^C = 1$, $N^T = 0$, and $N^G = 2$. The spectra I^γ for the particular bases (γ) are as follows:

$$
\begin{aligned}
I^A(x) &= \exp[-(x-\epsilon_2)^2] + \exp[-(x-\epsilon_3)^2] + \exp[-(x-\epsilon_4)^2],\\
I^C(x) &= \exp[-(x-\epsilon_5)^2],\\
I^T(x) &= 0,\\
I^G(x) &= \exp[-(x-\epsilon_1)^2] + \exp[-(x-\epsilon_6)^2].
\end{aligned}
$$

The details of spectra are better visible when the step r is large, i.e. when the neighboring maxima are well separated. When the step is small then several Gaussian functions overlap and are represented by a single envelope with a larger height. In this sense, the step is related to the resolution of the spectrum. With an increasing r the resolution becomes larger.

A convenient characterization of spectra is given by their distribution moments. The qth moment ($q = 0, 1, 2, \ldots$) of the distribution $I^\gamma(x)$ is defined as

$$
M_q^\gamma = c^\gamma \int\limits_{R(x)} I^\gamma(x) x^q dx, \tag{3}
$$

where the normalization constant

$$
c^\gamma = \left(\int\limits_{R(x)} I^\gamma(x) dx \right)^{-1}
$$

and $R(x)$ is the range of x for which the integrand does not vanish. In some cases it may be convenient to split the sequences to several parts. This may be reflected by a proper choice of $R(x)$.

The centered moments for which $M_1^{\gamma\prime} = 0$ are defined as

$$
M_q^{\gamma\prime} = c^\gamma \int\limits_{R(x)} I^\gamma(x)(x - M_1^\gamma)^q dx. \tag{4}
$$

The scaled moments for which $M_1^{\gamma\prime\prime} = 0$ and $M_2^{\gamma\prime\prime} = 1$ read

$$
M_q^{\gamma\prime\prime} = c^\gamma \int\limits_{R(x)} I^\gamma(x) \left[\frac{x - M_1^\gamma}{\sqrt{M_2^\gamma - (M_1^\gamma)^2}} \right]^q dx. \tag{5}
$$

A very clear meaning have several lowest moments. The first moment M_1 describes the mean value of the distribution, the second centered moment $M_2^{\gamma\prime}$ is referred to

as the variance of the distribution, the third scaled moment, $M_3^{\gamma\prime\prime}$, is the skewness coefficient and the fourth one, $M_4^{\gamma\prime\prime}$, is the excess.

The spectra change with r and, as a consequence, also all their moments. Therefore, the descriptors practically independent of r (cf. the next chapter) are defined:

$$D_1^{\gamma} = \frac{M_1^{\gamma}}{r}, \tag{6}$$

$$D_2^{\gamma} = \frac{M_2^{\gamma\prime}}{r^2}, \tag{7}$$

$$D_3^{\gamma} = M_3^{\gamma\prime\prime}, \tag{8}$$

$$D_4^{\gamma} = M_4^{\gamma\prime\prime}. \tag{9}$$

I propose γ-descriptors D_1^{γ}, D_2^{γ}, D_3^{γ}, D_4^{γ} as new descriptors of DNA sequences.

3 Results and discussion

The new descriptors have been calculated for ten histone H4 DNA sequences [6,13, 14,16]:

1. Maize ZMH4C7
2. Maize ZMH4C14
3. Maize ZMH4A
4. Chicken GGHIST4A
5. Chicken GGHIST4B
6. Wheat TAH4091
7. Mouse MMHIST4
8. Rat RR4HIS
9. Human HSHIS4
10. Human HSHISAD

$N = 311$ for the sequence labeled by 9 and for all other cases $N = 312$.

Figures 1 and 2 show the four-component spectra for histone H4 gene of chicken GGHIST4A (label 4 in the list). The differences between the two figures are the steps ($r = 1.0$ in Fig. 1 and $r = 0.1$ in Fig. 2). The I^{γ} components show us the positions of the particular bases. The direct positions of the bases is particularly well observed when $r = 1.0$ ($N = 312$ in this case). The sequence starts with: $ATGTCTGGCAGAGGC$. The positions of the maxima of Gaussians are ϵ_1, ϵ_{10}, ϵ_{12} for A component. As a result (Fig. 1), we observe at the outset of A spectrum a single peak for ϵ_1 and one function with a double peak for closely located (separated by only one space) ϵ_{10} and ϵ_{12}. For component C, the maxima appear for x equal to ϵ_5, ϵ_9, ϵ_{15}. Therefore the outset of the C spectrum consists of three single and well separated Gaussians. Their heights are equal to 1. For the T component, the maxima appear for x equal to ϵ_2, ϵ_4, ϵ_6. They are quite closely located. Thus, the outset of the T spectrum is represented by one function with a triple peak. The number of maxima for the G spectrum is quite large in this short distance. There are six G

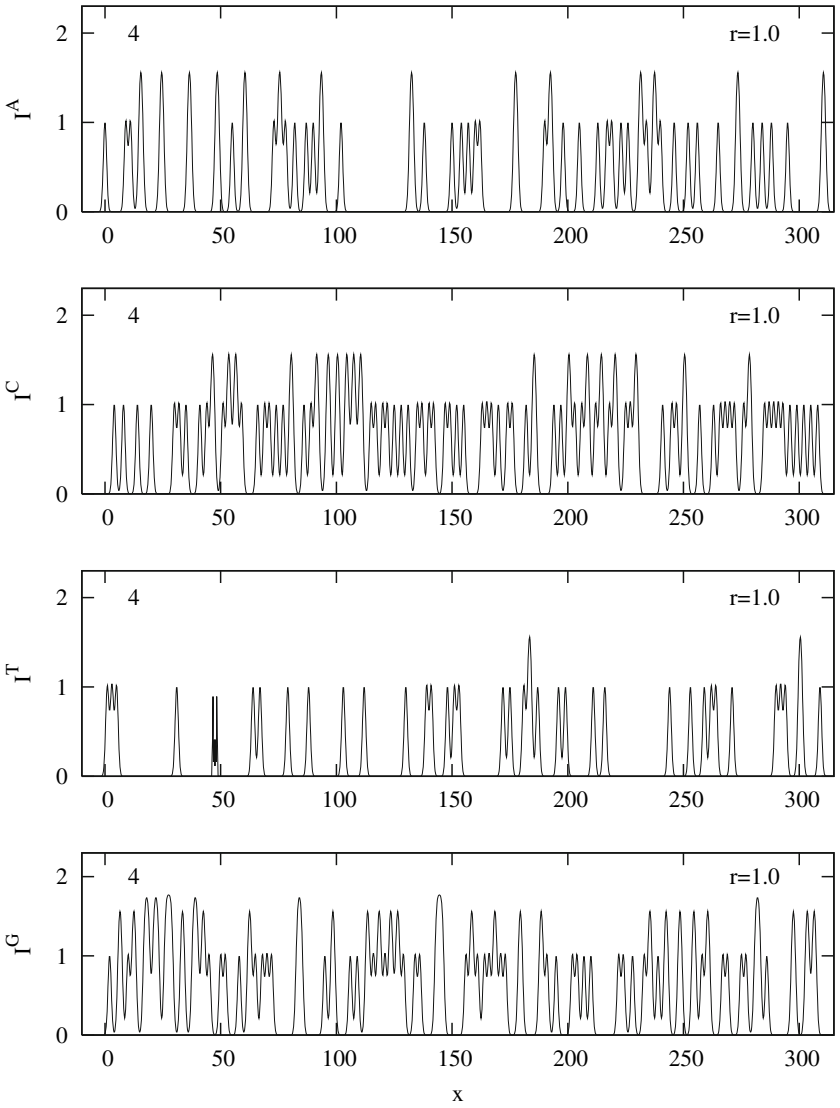

Fig. 1 Spectral representation for histone H4 gene of chicken GGHIST4A for $r = 1.0$

maxima ($\epsilon_3, \epsilon_7, \epsilon_8, \epsilon_{11}, \epsilon_{13}, \epsilon_{14}$). Therefore, single Gaussian functions overlap. Then, as a consequence of the summation of closely located Gaussians, maxima higher than one appear at the outset of spectrum G. Larger r separates single functions but for a good visualization an extension of the x axis is necessary in this case. In such cases it is better to split $R(x)$ to several parts. In the case when we are interested in a global comparison of sequences, taking a smaller r could be a better solution. In Fig. 2 the step has been decreased by a factor of 1/10. Such a visualization gives an approximate positions of the bases. For example, the T spectrum in Fig. 2 looks very much different

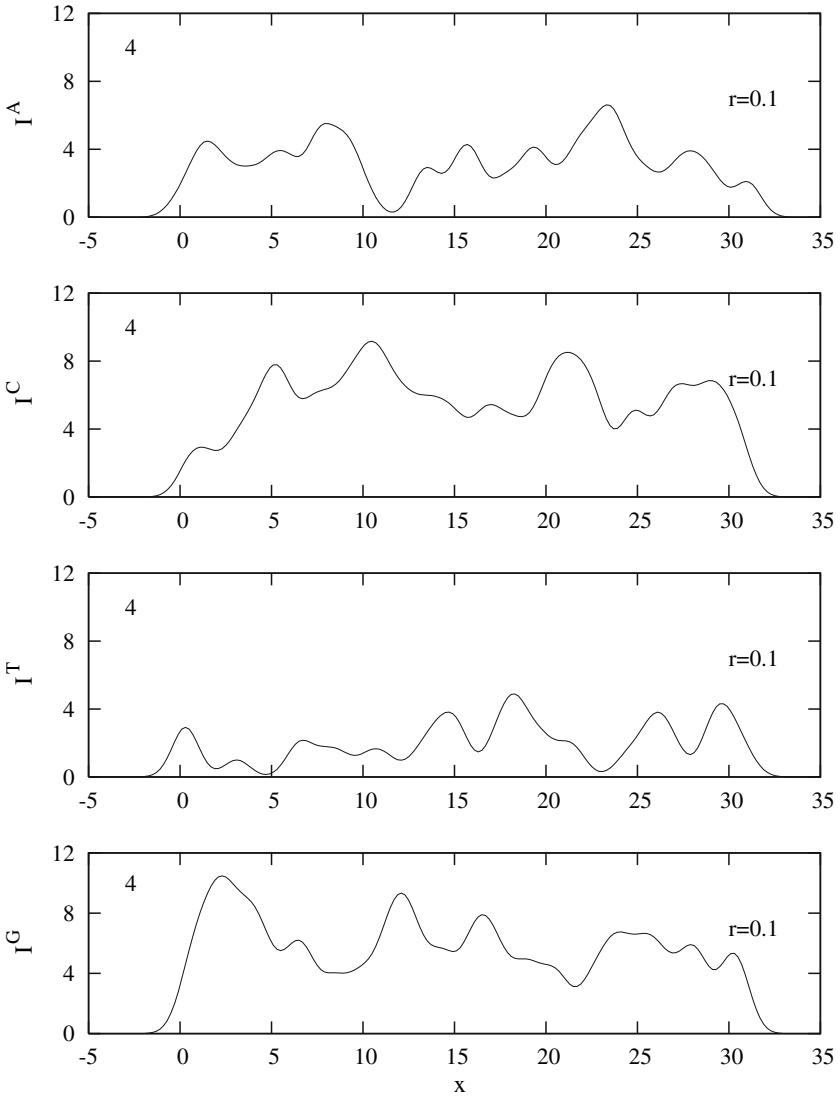

Fig. 2 Spectral representation for histone H4 gene of chicken GGHIST4A (the same as in Fig. 1) for $r = 0.1$

from the others. The maxima are small. This is a consequence of a large separations of T bases. Figure 1 confirms this observation where there is one-to-one correspondence between the positions of the bases in the sequence and the positions of the maxima in the spectrum.

Visually, a pair of spectra can be compared easily by plotting their differences. In Fig. 3, the differences between I^{γ} components for Chicken GGHIST4A (4) and Chicken GGHIST4B (5) are shown. The differences for the A and G components are

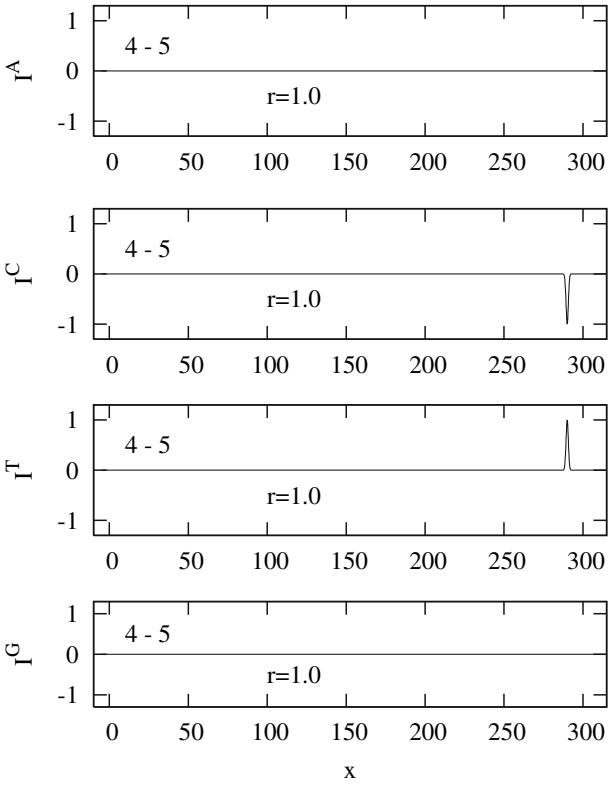

Fig. 3 Differences between the spectra for histone H4 gene of chicken GGHIST4A (4) and histone H4 gene of chicken GGHIST4B (5) for $r = 1.0$

exactly equal to zero. The two sequences differ by only one base. For $i = 291$ in the sequence labeled by 4, T appears and in this place in the sequence labeled by 5, C appears. This small difference between the two sequences is visible in Fig. 3, where single peaks for C and T are observed. The difference is also detected by the descriptors that are shown in the tables (Table 1 for A spectrum, Table 2 for C, and Table 3 for T one). The descriptors for the A and G components in the case of sequences labeled 4 and 5 are exactly the same (the G-descriptors are shown in Fig. 4). D_1^C for the sequence 5 is larger than D_1^C for the sequence 4. This index seems to be quite sensitive since only one C base in the sequence 5 located close to the end of the sequence causes an increase in the mean value of the distribution. Increasing the number of bases ($N^C = 105$ for sequence 5 and $N^C = 104$ for sequence 4) also causes an increase in the width of the distribution and such a relation between D_2^C is also observable (D_2^C for the sequence 5 is larger than D_2^C for the sequence 4). As a result of a larger concentration of the intensity for large x for sequence 5, D_3^C becomes smaller than for sequence 4. Also the excess for sequence 5 becomes smaller. Similar behavior of the T-descriptors can be observed for the two sequences ($N^T = 38$ for the sequence 4 and $N^T = 37$ for the sequence 5).

Table 1 *A*-descriptors representing ten histone H4 genes

No.	D_1^A	D_2^A	D_3^A	D_4^A
1	150.08	7570.5	0.0711	1.8300
2	152.95	7514.9	0.0488	1.8246
3	148.25	7657.2	0.0953	1.8133
4	155.79	8452.9	−0.0843	1.7009
5	155.79	8452.9	−0.0843	1.7009
6	152.08	7707.1	0.0251	1.7982
7	145.77	8582.5	0.0388	1.6546
8	149.74	8458.2	0.0618	1.6902
9	146.97	8769.1	0.0505	1.6290
10	144.74	8252.8	0.0416	1.6353

Numbers in the first column correspond to particular histone H4 genes defined in the text

Table 2 *C*-descriptors representing ten histone H4 genes

No.	D_1^C	D_2^C	D_3^C	D_4^C
1	168.43	7429.1	−0.1772	1.7860
2	166.90	7607.3	−0.2026	1.7768
3	170.08	7247.7	−0.2018	1.8272
4	159.58	7291.5	0.0582	1.8204
5	160.82	7382.6	0.0466	1.8055
6	167.45	7349.6	−0.1846	1.8377
7	167.99	6918.7	−0.0225	1.8096
8	166.43	6504.1	0.0114	1.8297
9	163.61	7657.4	−0.0519	1.7643
10	163.67	6795.5	0.0806	1.7789

Numbers in the first column correspond to particular histone H4 genes defined in the text

Table 3 *T*-descriptors representing ten histone H4 genes

No.	D_1^T	D_2^T	D_3^T	D_4^T
1	177.81	7092.6	−0.2965	2.2436
2	166.41	7666.4	−0.1874	2.1668
3	177.43	7255.0	−0.2800	2.1900
4	175.21	7978.1	−0.3142	2.1693
5	172.11	7828.0	−0.2858	2.1901
6	172.11	7828.0	−0.2858	2.1901
7	158.67	8075.1	−0.0540	1.9011
8	156.24	9904.4	−0.0391	1.6985
9	161.68	6467.8	0.0218	2.2451
10	164.68	8440.4	−0.1730	1.9681

Numbers in the first column correspond to particular histone H4 genes defined in the text

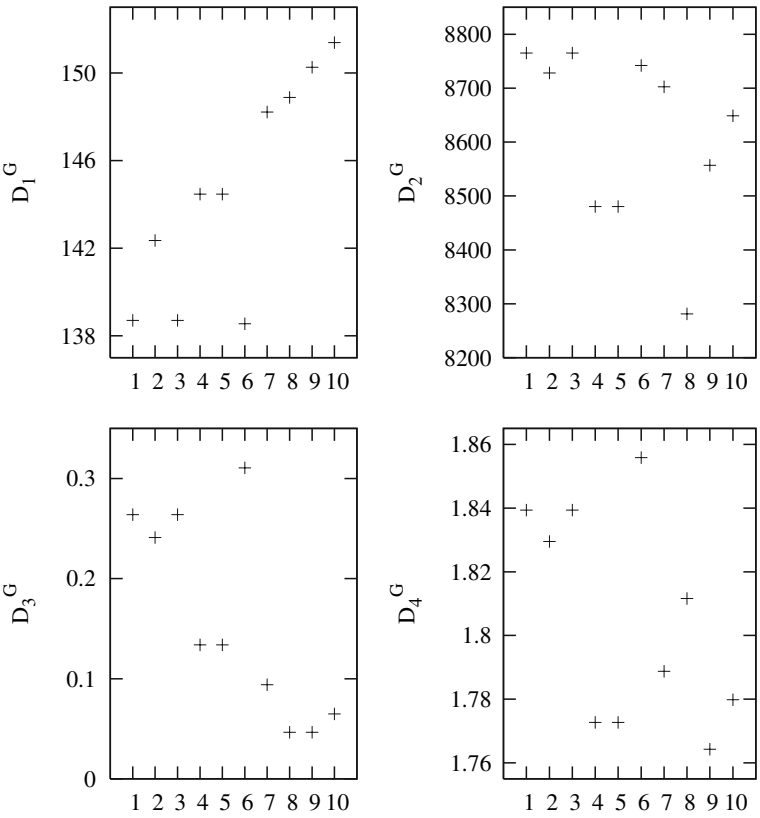

Fig. 4 G-descriptors for ten histone H4 DNA sequences. Numbers in the horizontal axes correspond to the particular labels of the sequences defined in the text

Figure 5 presents the dependencies of γ-descriptors on r. The first moment is proportional to r. Therefore D_1^γ is r-independent. The higher order descriptors depend on r, however this dependence is practically detectable for very small r only. For about $r > 0.2$ this dependence appears to be negligible. Intuitively (graphically) for small r the summation of single Gaussians results in an envelope that hides the information about the details of the distributions of particular bases. With an increasing resolution (increasing r), the information about the structure of the basis becomes larger until it reaches a maximum. Further separations of the Gaussians does not add any new information. If we reject the small r values we get γ-descriptors which are r-independent.

Moreover, in the case of G-descriptors (Fig. 4), the organisms which are evolutionary similar, as plants and vertebrates, cluster. Such a clustering is observed even for the first moments, which are smaller for the sequences labeled by 1,2,3,6 (plants) than for those for vertebrates. Similar conclusions have been drawn in our previous paper, where another kind of statistical distributions was considered. In that paper such clustering was observed for higher order moments [13].

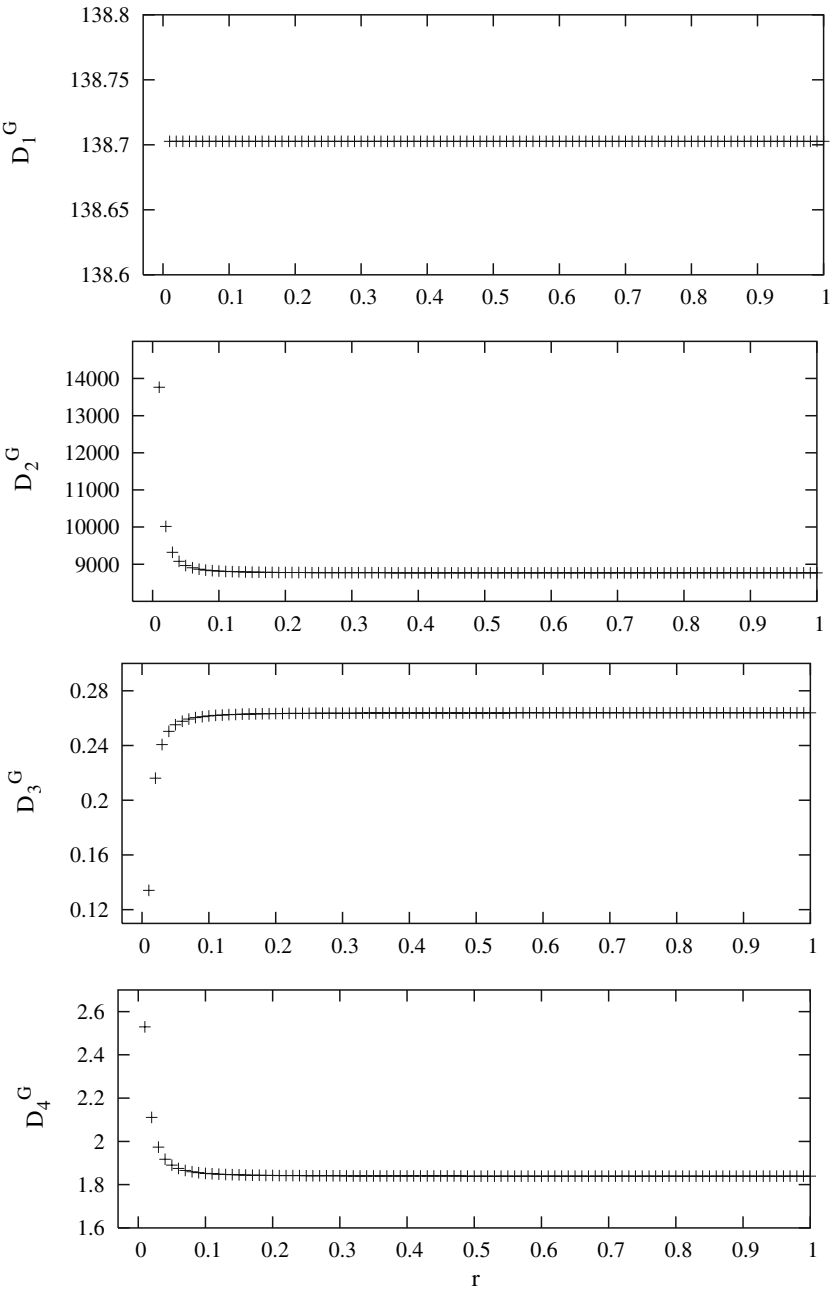

Fig. 5 The dependence of G-descriptors on r for Maize ZMH4C7

Summarizing, the presented method is a convenient graphical and numerical tool for determining distributions of bases along the sequences.

References

1. A. Nandy, M. Harle, S.C. Basak, ARKIVOC **ix**, 211 (2006)
2. M. Randić, M. Vracko, A. Nandy, S.C. Basak, J. Chem. Inf. Comput. Sci. **40**, 1235 (2000)
3. M.A.J. Gates, Theor. Biol. **119**, 319 (1986)
4. A. Nandy, Curr. Sci. **66**, 309 (1994)
5. P.M. Leong, S. Morgenthaler, Comput. Appl. Biosci. **11**, 503 (1995)
6. D. Bielińska-Wąż, T. Clark, P. Wąż, W. Nowak, A. Nandy, Chem. Phys. Lett. **442**, 140 (2007)
7. E. Hamori, J. Ruskin, J. Biol. Chem. **258**, 1318 (1983)
8. R. Chi, K. Ding, Chem. Phys. Lett. **407**, 63 (2005)
9. B. Liao, T. Wang, J. Chem. Inf. Comput. Sci. **44**, 1666 (2004)
10. A. Nandy, S.C. Basak, J. Chem. Inf. Comput. Sci. **40**, 915 (2000)
11. M. Randić, M. Vracko, N. Lers, D. Plavsić, Chem. Phys. Lett. **368**, 1 (2003)
12. G. Aguero-Chapin, H. González-Díaz, R. Molina, J. Varona-Santos, E. Uriarte, Y. González-Díaz, FEBS Lett. **580**, 723 (2006)
13. D. Bielińska-Wąż, W. Nowak, P. Wąż, A. Nandy, T. Clark, Chem. Phys. Lett. **443**, 408 (2007)
14. D. Bielińska-Wąż, P. Wąż, T. Clark, Chem. Phys. Lett. **445**, 68 (2007)
15. D. Bielińska-Wąż, P. Wąż, S.C. Basak, Eur. Phys. J. B **50**, 333 (2006)
16. C. Raychaudhury, A. Nandy, J. Chem. Inf. Comput. Sci. **39**, 243 (1999)