MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **60** (2008) 301-312

ISSN 0340 - 6253

A numerical characterization of modified Hamori curve representation of DNA sequences

Igor Pesek [a], **Janez Žerovnik** [a,b]

[a] IMFM, Jadranska 19,
Ljubljana, Slovenia
[b]University of Maribor, Smetanova 17
Maribor, Slovenia
{igor.pesek,janez.zerovnik}@imfm.uni-lj.si

(Received March 21, 2008)

Abstract. We present new numerical characterization of DNA sequences that is based on the modified graphical representation proposed by Hamori. While Hamori embeds the sequence into Euclidean space, we use analogous embedding into the strong product of graphs, $K_4 \boxtimes P_n$, with weighted edges. Based on this representation, a novel numerical characterization is proposed which is based on the products of ten eigenvalues from the start and the end of the descending ordered list of the eigenvalues of the L/L matrices associated with DNA. The examination of similarities/dissimilarities among coding sequences of the first exon of the β-globin gene of different species illustrates the utility of the approach.

1 Introduction

Nowadays the automated DNA sequencing techniques have led to an explosive growth in the number and the length of DNAs sequences from different organisms. This has resulted in a large accumulation of data in the DNA databases, but has also called for the development of suitable techniques for rapid viewing and analysis of the data. Graphical representations of DNA sequences were initiated by Hamori [4] and later expanded by

many others, see the review [21] and a number of more recent papers, for example [8, 9, 10, 12, 13, 14, 15, 22, 23, 24], the list being by no means exhaustive.

The advantage of graphical representation of DNA sequences is that they allow visual inspection of data, helping in recognizing major differences among similar DNA sequences. These techniques provide useful insights into local and global characteristics and the occurrences, variations and repetition of the nucleotides along a sequence which are not as easily obtainable by other methods. Two-dimensional plots are obviously useful for visual communication of the results of an analysis, but can also be useful to help checking for the presence of an effect by human eye rather by a computer program, and finally, they are used for identifying unsuspected structures in the data. Recently, it has been shown that some of the graphical representations lead to numerical characterizations of DNA sequences and quantitative measures of the degree of similarity/dissimilarity between the sequences [13, 14, 15, 21, 23]. Similarly as topological indices used as molecular descriptors can dramatically improve the search for synthesis of compounds with a desired property [20], it is hoped that the numerical descriptors of DNA may be used to predict some properties of the DNA sequences. An important advantage of a characterization of structures by invariants, as opposed to use of codes, is the simplicity of the comparison of numerical sequences based on invariants. The price paid is a loss of information on some aspects of the structure that accompany any characterization based on invariants. The loss of the information, however, can in part be reduced by use of larger number of descriptors (invariants) [16, 17].

By a *graph* we mean a set $V(G)$ of vertices, together with a set $E(G)$ of edges. A graph is the *complete graph* K_n if any two of its distinct vertices are adjacent. A graph is the *path* P_n if it is isomorphic to a graph on n distinct vertices $v_1, v_2, \ldots, v_n$ and $n-1$ edges v_i, v_{i+1}, $1 \leq i < n$.

As the four bases A, G, C, and T are regarded independent, at least four dimensions are needed for an embedding that is free of using some arbitrary conventions. A number of graphical representations first embeds the DNA sequence into an Euclidean space of some dimension, using a projection to 2-D plot, where for the projection again some more or less arbitrary choice has to be made. In this paper, we essentially use a more dimensional presentation, but instead of working with Euclidean coordinates we

rather embed the sequence into a graph, more precisely into a strong product of K_4 times a path. A geometric representation would then be more than two dimensional as an isometric drawing of K_4 is only possible in three dimensions. In figures here we use a particular drawing of the graph, which in our opinion seems to give a good impression of the sequence to the observer. The one dimensional plot of K_4 is of course not isometric (i.e. the edges in the plot have different lengths) but we believe that the resulted drawing may be a reasonable compromise between the arbitrary projection(s) and a unique more dimensional embedding which can, of course, easily be found by an isometric embedding of the complete graph K_4 into Euclidean space, for example by mapping A, C, G, and T to the edges of a tetrahedron in 3D or to the four unit vectors in 4D. Furthermore, based on this graph representation we propose a novel numerical characterization of the DNA sequence. In contrast to some other numerical characterizations that are based on the graphical representations [9, 12, 15, 23], our representation is free of arbitrary choices because it is based on the graph and not on its drawing, i.e. embedding and projection. The numerical characterization uses eigenvalues of a matrix that is based on the graph distances. The numerical invariant is computed for the first exon of the β-globin gene for the 10 different species, a dataset shown in Table 1, that is used in many recent studies [8, 9, 10, 12, 13, 14, 15, 22, 23] and is taken from EMBL-EBI database [25]. This dataset is one of the primary tools for comparison of different graphical and numerical characterizations and was first used by Nandy [11] and later by other authors [8, 12, 13, 15, 22]. The reason why Nandy decided to use this gene lies in the fact that β globin sequences represent a conservative gene, that is, the gene that changes little from one species to another. The differences between the values of the invariant are used as a measure of similarity/dissimilarity among the species. We do not attempt to extensively comment the results because this is not an area of our expertise. However we wish to note that our results are not like those obtained by a similar computations which are based on eigenvalues of the graphical representations [12], but are based on graphs, therefore our approach is using less computational effort. For example in [12] one has to compute 12 different permutations of the graphical representation before the actual characterization, while our approach computes only one.

The rest of the paper is organized as follows. In the next section we recall the Hamori

curve representation and explain our modification. Our numerical characterization is explained in Section 3, and in Section 4 we give results of application to a data sample.

2 Modified Hamori curve representation

We based our research on DNA sequence representation introduced by Hamori [4]. In this method, the information content of a DNA sequence is mapped into a three-dimensional space function (H curve). The positive x-direction is used to count the number of bases in the sequence. At each point of x on the corresponding yz plane the four corners (NW, NE, SE and SW as four points on the compass) are taken to represent the four bases A, C, G and T. Basic rule for the construction of the sequence map is to move one unit in the corresponding direction depending on which nucleotide (base) is being plotted and to draw a connected line of all such points plotted, one for each unit in the x-direction. Thus a sequence like ATGGTGCACCTGACT... will generate a spiral along the x-axis.

H-curve representation is sensitive to the directions chosen for four bases. For example representation with bases $ACGT$ corresponding to four corners is different from $AGCT$, since the distance from base A to base G is different in this two cases.

We modified this approach by putting the corners of four bases on the K_4 and weighted all the edges in K_4 with 1. This way we avoided the drawback of the original representation. Edges in the x direction or along P_n are weighted with 1 if the base in the coding sequence is the same as the previous one and with $\sqrt{2}$ otherwise.

Formally, a sequence of the length n in this paper is a path in the strong product of the graphs K_4 and P_n. The *strong product* $G_1 \boxtimes G_2$ of graphs G_1 and G_2 has as vertices the pairs (g, h) where $g \in V(G_1)$ and $h \in V(G_2)$. Vertices (g_1, h_1) and (g_2, h_2) are adjacent if either $\{g_1, g_2\}$ is an edge of G_1 and $h_1 = h_2$ or if $g_1 = g_2$ and $\{h_1, h_2\}$ is an edge of G_2 or if $\{g_1, g_2\}$ is an edge of G_1 and $\{h_1, h_2\}$ is an edge of G_2. The strong product is one of the standard graph products [7].

Here K_4 is a complete graph on vertices A, C, G, T and P_n is a path on the vertices $1, 2, \ldots, n$. The edges of the product are weighted as follows:

$$W((i,j)(k,\ell)) = \begin{cases} 1 & i = k \text{ or } j = \ell \\ \sqrt{2} & i \neq k \text{ and } j \neq \ell \end{cases} \tag{1}$$

Figure 1 shows modified Hamori curve, where first few edges between the K_4's have weights indicated with the numbers on gray background. The factor K_4 is drawn on a

circle and projected to obtain a 2-D drawing. Any other possibly nicer drawing of the graph $K_4 \boxtimes P_n$ can be used [1, 2]. However, we find our way of drawing the graph and the path a reasonable compromise that can be used as a help for easier understanding of our concept. Note that all the edges within the vertical factor (K_4) and all the horizontal edges have weight 1 while all edges between K_4 factors that are not horizontal have weight $\sqrt{2}$. The motivation for choosing $\sqrt{2}$ is the intuitive assumption that the two factors in the product are orthogonal, hence the corresponding edge is the diagonal of a unit square.

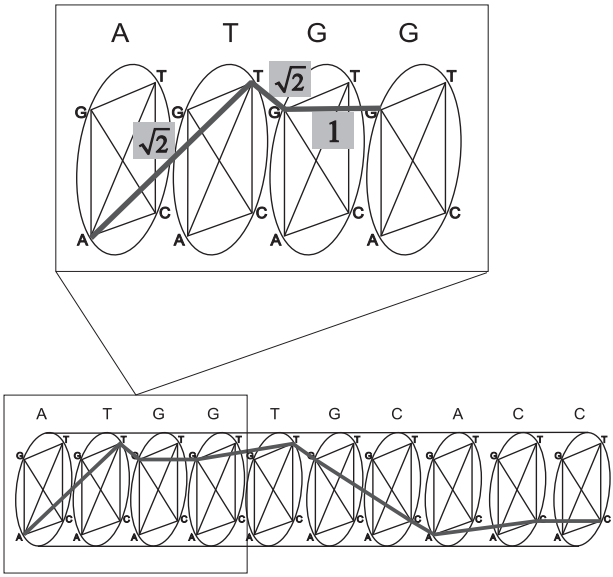

Figure 1: Modified Hamori curve

3 Numerical characterization of DNA sequences

In order to numerically characterize a DNA sequence given by the 2-D graphical representation based on our approach one can associate with a corresponding zigzag curve a matrix and consider matrix invariants that are sensitive to the form of the curve. This approach was first outlined and used by Randić et al. [13]. One of the possible matrices they use is the **L/L** matrix (the length/length matrix) whose elements are defined as the quotient of the distance between a pair of the vertices (dots) of the zigzag curve and the

Table 1: The coding sequences of the first exon of β-globin gene of 10 different species

Species	Coding sequence
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA-
(92 bases)	GGTGAACGTGGAGTAAGTTGGTGGTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAG-
(92 bases)	GTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAA-
(92 bases)	GGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAA-
(92 bases)	GGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Mouse	ATGGTGCACCTGACTGATGCTGAGAAGGCTGCTGTCTCTTGCCTGTGGGGAAA-
(92 bases)	GGTGAACTCCGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAA-
(90 bases)	GGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAA-
(92 bases)	GGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA-
(93 bases)	GGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAA-
(86 bases)	AGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA-
(105 bases)	GGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG

sum of distances between the same pair of vertices measured along the zigzag curve. Here we use analogous matrix based on the weighted graph representation of DNA, i.e. the entries of the **L/L** matrix are the quotients between the graph distance and the weighted graph distance. Using this weights we can construct **L/L** matrix as is shown in Table 2 where we used first 10 bases of the first exon of β-globin gene of human. For example, the first three entries of the first row are $\frac{1}{\sqrt{2}} \simeq 0.707$, $\frac{2}{\sqrt{2}+\sqrt{2}} \simeq 0.707$, and $\frac{3}{\sqrt{2}+\sqrt{2}+1} \simeq 0.783$.

Formally, we assign the matrix LL_x to the sequence x with

$LL_x(i,j) = \frac{j-i}{d((x_i,i),(x_j,j))}$

where $d((x_i,i),(x_j,j))$ is the distance in the weighted graph $K_4 \boxtimes P_n$. More precisely, $d((x_i,i),(x_j,j)) = \sum_{k=i}^{j-1} W((x_k,k)(x_{k+1},k+1))$ for $j > i$. (For $i = j$ we put $d((x_i,i),(x_j,j)) = 0$ and for $j < i$ we define $d((x_i,i),(x_j,j)) = d((x_j,j),(x_i,i))$.)

We will characterize the coding sequences of the first exon of β-globin gene of 10 species (including human), shown in the Table 1, by means of the leading eigenvalues, λ, of the **L/L** matrix. Eigenvalues of a matrix are one of the best known matrix invariants. If a matrix is symmetric, as is the case with all the matrices considered here, the eigenvalues are real. A set of eigenvalues can be viewed as a characterization of a structure, but as is well known such characterization is not unique. In other words, different graphs and different structures may have the same set of eigenvalues. Such graphs are known as isospectral and have received considerable attention in mathematics [5, 3] and chemistry [6], of which we only indicated some earlier contributions. While it was initially thought

that the complete coincidence of all eigenvalues may be an exception rather than a rule, the subsequent research revealed that isospectral graphs are more a rule than exception. That, however, does not diminish their utility, although they would fail to discriminate structures in testing for isomorphism [16]. On other hand, if two structures are similar they are likely to have similar eigenvalues and consequently similar product of leading eigenvalues. In a recent study in which the DNA sequence was characterized by average distances between various nucleic acid bases was shown that is very sensitive already when a single nucleic base has been changed [19].

Our characterization begins with computing the **L/L** matrix and then computing eigenvalues of this matrix. First few eigenvalues for each species are shown in descending order in the Table 3. In the next step we order the eigenvalues from largest to the smallest. Then we compute the product of first ten and last ten leading eigenvalues of such ordering. Species have different lengths of DNA sequence, shortest is DNA sequence of the bovine (86 bases) and longest of the Chimpanzee (105 bases), therefore we needed to find some common number of eigenvalues. We decided to take 10 eigenvalues from the start and the end of the descending ordered list of the eigenvalues. (Note that the choice of 10 was arbitrary. We have performed the same procedure using 9 and 11 eigenvalues and obtained very similar results which gives some evidence that the method is robust, i.e. not too sensitive to the choice of the number of eigenvalues.)

Formally, the numerical characterization of the sequence $x = (x_1, x_2, \ldots, x_n)$, $x_i \in \{A, C, G, T\}$ is a product of first ten and last ten eigenvalues of the descending ordered eigenvalues list of the matrix LL_x,

$$\Lambda(x) = \lambda_1(LL_x)\lambda_2(LL_x)\ldots\lambda_{10}(LL_x)\lambda_{n-9}(LL_x)\lambda_{n-8}(LL_x)\ldots\lambda_n(LL_x). \qquad (2)$$

4 Similarities/dissimilarities among the coding sequences of the first exon of β-globin gene of the different species

We will illustrate the use of novel quantitative characterization of the DNA sequences with the examination of the similarities/dissimilarities among the 10 coding sequences shown in Table 1. The analysis of similarity/dissimilarity is based on the assumption that two

Table 2: The upper triangle of the L/L matrix of the sequence ATGGTGCACC

Base	A	T	G	G	T	G	C	A	C	C
A	0	0.707	0.707	0.783	0.762	0.751	0.743	0.737	0.733	0.756
T		0	0.707	0.828	0.783	0.762	0.751	0.743	0.737	0.762
G			0	1.00	0.828	0.783	0.762	0.751	0.743	0.771
G				0	0.707	0.707	0.707	0.707	0.707	0.743
T					0	0.707	0.707	0.707	0.707	0.751
G						0	0.707	0.707	0.707	0.762
C							0	0.707	0.707	0.783
A								0	0.707	0.828
C									0	1.00
C										0

Table 3: The 10 leading eigenvalues, λ, of the $\mathbf{L/L}$ matrices for the coding sequences

Leading eigenvalues									
Human	Opussum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
70.224	68.652	70.267	69.071	69.792	68.464	70.390	71.033	66.916	80.584
0.065	0.023	0.092	0.227	0.327	0.044	0.307	0.069	0.312	0.087
-0.236	-0.156	-0.122	-0.017	-0.218	-0.237	-0.113	-0.236	-0.224	-0.213
-0.275	-0.363	-0.237	-0.205	-0.239	-0.248	-0.229	-0.275	-0.319	-0.273
-0.378	-0.397	-0.377	-0.244	-0.321	-0.388	-0.265	-0.396	-0.377	-0.349
-0.396	-0.440	-0.389	-0.353	-0.378	-0.397	-0.327	-0.378	-0.426	-0.381
-0.426	-0.444	-0.413	-0.467	-0.426	-0.425	-0.420	-0.419	-0.447	-0.396
-0.433	-0.460	-0.435	-0.488	-0.461	-0.432	-0.439	-0.432	-0.459	-0.415
-0.461	-0.471	-0.487	-0.501	-0.475	-0.447	-0.457	-0.461	-0.483	-0.433
-0.475	-0.485	-0.503	-0.513	-0.503	-0.484	-0.495	-0.475	-0.490	-0.461

DNA sequences are similar if the corresponding differences between the products of ten leading eigenvalues are small.

In Table 4 we give the similarity/dissimilarity matrix. The smallest entries in Table 4 are associated with the pairs (human, chimpanzee), (human, gorilla) and (gorilla, chimpanzee) which is in accordance with our intuitive expectations and, not surprisingly, also in accordance with other studies [8, 13]. On the other hand the largest entries in the similarity/dissimilarity matrix appear in rows belonging to bovine and opossum. On the basis of these findings we conclude that the presented numerical characterization via L/L matrices as well products of leading eigenvalues have captured some important features of the DNA sequences considered.

Formally we can define similarity relations as follows:

$$similarity(x, y) = |\Lambda(x) - \Lambda(y)|, \text{ where } x, y \text{ are sequences of the species}$$

Table 4: The similarity/dissimilarity matrix for the coding sequences of Table 1 based on product of first ten leading values of **L/L** matrix

Species	Human	Chimpanzee	Gorilla	Opussum	Gallus	Lemur	Mouse	Rabbit	Rat	Bovine
Human	0	0,002299	0,000652	0,00862	0,003545	0,009986	0,041185	0,005268	0,007074	0,100412
Chimpanzee	0,002299	0	0,001646	0,010919	0,005843	0,012284	0,038886	0,007566	0,004775	0,098113
Gorilla	0,000652	0,001646	0	0,009272	0,004197	0,010638	0,040532	0,00592	0,006421	0,099759
Opussum	0,00862	0,010919	0,009272	0	0,005075	0,001366	0,049805	0,003352	0,015694	0,109032
Gallus	0,003545	0,005843	0,004197	0,005075	0	0,006441	0,044729	0,001723	0,010618	0,103956
Lemur	0,009986	0,012284	0,010638	0,001366	0,006441	0	0,05117	0,004718	0,017059	0,110397
Mouse	0,041185	0,038886	0,040532	0,049805	0,044729	0,05117	0	0,046452	0,034111	0,059227
Rabbit	0,005268	0,007566	0,00592	0,003352	0,001723	0,004718	0,046452	0	0,012341	0,105679
Rat	0,007074	0,004775	0,006421	0,015694	0,010618	0,017059	0,034111	0,012341	0	0,093338
Bovine	0,100412	0,098113	0,099759	0,109032	0,103956	0,110397	0,059227	0,105679	0,093338	0

5 Conclusion

Our objective was to arrive at a numerical characterization of DNA sequences, which as can be seen from Table 4, may be accomplished in a relatively simple algebraic manner and make the proposed approach very attractive for the characterization of DNA sequences having 1,000 or more bases. Needles to say that the outlined approach is suitable for characterization of local fragments of DNA, which is precisely how one may look on the truncated DNA fragment considered in this work. Conceptually and computationally the approach is simple and therefore can be very useful in the field of the bioinformatics.

While the existence of isospectral graphs implies that there are structures that can not be distinguished by any spectral method, another important reason for losing the structural information may be due to the method of generating a graph (or matrix) from the sequence (e.g. choice of the representation and/or choice of the subset of eigenvalues) as pointed out by one of the referees. In our case, it is clear that the method does not distinguish between sequences in which the letters are permuted. Such sequences give rise to isomorphic graphs and hence the spectrums are identical. This can be a serious drawback if arbitrary sequences are compared, but it seems extremely unlikely that it would be of any importance when comparing DNA sequences. For example, when restricted to certain gene (or, DNA fragment) from different species it would mean that we have two working genes at two different species that differ by a permutation of letters which means that their distance measured in mutations is very large and hence extremely unlikely (or, impossible).

Acknowledgement. We authors wish to thank to the referees for constructive remarks. This work was supported in part by grants of the ARRS, Slovenian Research Agency.

References

[1] G. Di Battista, P. Eades, R. Tamassia, I. G. Tollis, Graph Drawing: Algorithms for the Visualization of Graphs, Prentice Hall PTR, New Jersey, 1998.

[2] G. Di Battista, P. Eades, R. Tamassia, I. G. Tollis, Algorithms for drawing graphs: an annotated bibliography, Comp. Geom-Theor. Appl. 4 (1994) 235–282.

[3] C. D. Godsil, D. A. Holton, B. D. McKay, The spectrum of a graph, Combinatorial Mathematics V, Lect. Notes. Math. 622 (1977) 91–117.

[4] E. Hamori, Graphical representation of long DNA sequences by methods of H curves, current results and future aspects. Biotechniques 7 (1989) 710–720.

[5] F. Harary, C. King, A. Mowshowitz, R.C. Read, Cospectral graphs and digraphs, Bull. London Math. Soc. 3 (1971) 321–328.

[6] W. C. Herndon, M. L. Ellzey, Isospectral graphs and molecules, Tetrahedron 31 (1975) 99–107.

[7] W. Imrich, S. Klavžar, Product Graphs: Structure and Recognition, John Wiley & Sons, New York, 2000.

[8] B. Liao, T. Wang, Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation, Chem. Phys. Lett. 388 (2004) 195–200.

[9] B. Liao, T. Wang, 3-D graphical representation of DNA sequences and their numerical characterization, J. Mol. Struc. - Theochem 681 (2004) 209–212.

[10] B. Liao, Y. Zhang, K. Ding, T. Wang, Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation, J. Mol. Struc. - Theochem 717 (2005) 199–203.

[11] A. Nandy, A new graphical representation and analysis of DNA sequence structure, Curr. Sci. India 66 (1994) 309–314.

[12] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Novel 2-D graphical representation of DNA sequences and their numerical characterization, Chem. Phys. Lett. 368 (2003) 1–6.

[13] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, Chem. Phys. Lett. 371 (2003) 202–207.

[14] M. Randić, M. Vračko, J. Zupan, M. Novič, Compact 2-D graphical representation of DNA, Chem. Phys. Lett. 373 (2003) 558–562.

[15] M. Randić, N. Lerš, D. Plavšić, S. Basak, A. Balaban, Four-color map representation of DNA or RNA sequences and their numerical characterization, Chem. Phys. Lett. 407 (2005) 205–208.

[16] M. Randić, M. Vračko, On the similarity of DNA primary sequences, J. Chem. Inf. Comput. Sci. 40 (2000) 599–606.

[17] M. Randić, M. Vračko, A. Nandy, S.C. Basak, On 3D graphical representation of DNA primary sequences and their numerical characterization, J. Chem. Inf. Comput. Sci. 40 (2000) 1235–1244.

[18] M. Randić, Condensed Representation of DNA Primary Sequences, J. Chem. Inf. Comput. Sci. 40 (2000) 50–66.

[19] M. Randić, S.C. Basak, Characterization of DNA Primary Sequences Based on the Average Distances between Bases, J. Chem. Inf. Model. 41 (2001) 561–568.

[20] D. Rouvray, Predicting Chemistry from Topology, Sci. Am. 254 (1986) 40–47.

[21] A. Roy, C. Raychaudhury, A. Nandy, Novel techniques of graphical representation and analysis of DNA sequences – A review, J. Bioscience 23 (1998) 55–71.

[22] J. Zupan, M. Randić, Algorithm for Coding DNA Sequences into "Spectrum-like" and "Zigzag" representations, J. Chem. Inf. Model. 45 (2005) 309–313.

[23] Y. Yao, X. Nan, T. Wang, Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation, Chem. Phys. Lett. 411 (2005) 248–255.

[24] D. Bielinska-Waz, T. Clark, P. Waz, W. Nowak, A. Nandy, 2D-dynamic representation of DNA sequences, Chem. Phys. Lett. 442 (2007) 140–144.

[25] http://www.ebi.ac.uk/