

Algorithm for Coding DNA Sequences into “Spectrum-like” and “Zigzag” Representations

Jure Zupan* and Milan Randić

National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia 61115

Received November 9, 2004

An algorithm for encoding long strings of building blocks, like 4 DNA bases (adenine - A, cytosine - C, thymine - T, and guanine - G), 20 natural amino acids (from Alanine Ala to Valine - Val, plus the stop triplet), or all 64 possible base triplets (from AAA to TTT), into “zigzag” or “spectrum-like” representations is suggested. The new encoding scheme can be derived in the 3-, 2-, or 1-dimensional form depending on the user's wishes. The only information, besides the string for which the “spectrum-like” representation is sought, is the initial positioning of the complete set of units from which the string is composed, i.e., four positions for A, C, G, and T, or 20 positions for natural amino acids plus stop, etc. This initial positioning can be initialized in either the 3-, 2-, or 1-D form. As an illustration of the suggested encoding scheme of the visual and chemometric comparison of the first 10 exon strings of the beta globin gene of 10 different species, each string consisting of about 100 basic amino acids long is shown.

INTRODUCTION

Because of the ease of handling, the visual attractiveness, and the numerous supporting methods already available for handling, the low dimensional graphical representation of complex multivariate data is very convenient for extracting the hidden information. This became apparent with a gradual expansion of purely graphical representations of DNA into accompanying numerical analyses based on the association of various distance-related matrices with the pictorial representations of DNA.^{1–3} Among alternative graphical representations of DNA sequences^{4–10} of special interest are graphical representations confined within a well defined planar,^{11–13} or 3-dimensional box¹⁴ of known dimension, the form of 2-D or 3-D “zigzag” curves,^{3,14–16} or in a limited interval on one axis in the form of a “spectrum-like” curve.³ It is worthwhile to point out that any connected curve can be represented as a graph, and within the graph theory many valuable numerical invariances can be calculated which in turn can be correlated to various physicochemical or biological properties. We should also add that one can arrive at a numerical characterization of DNA sequences also without a graphical representation of DNA as discussed in the literature by considering the frequency occurrence of pairs of bases,¹⁷ by recording the sequential occurrence of individual nucleic bases,¹⁸ or by considering the representation of DNA in a 4-D space.^{19,20} In this article we are focusing the attention on the algorithmic procedure for encoding the DNA or primary structures of protein sequences into visually more suitable representations.

In the present work an intention to give an algorithm that will produce a simple code or representation that can further enable the exploring of important features as the autocorrelation, Fast Fourier Transform functions, or any mathematical invariance imbedded in the graphs of these low

dimensional representations will be outlined. The obtained simpler for manual and computer handling and visually more attractive representation must retain as much of the original information as possible, possibly all. One immediate use of such a representation is that it can be employed as the input for mathematical models, either in the simple polynomial or more complex artificial neural network models.

THE ALGORITHM

The algorithm we are suggesting incorporates simultaneously the calculation of the 1-D, 2-D and 3-D representations all of which are unique and uniform for equally long strings. Additionally, 2-D and 3-D representations are reversible, i.e., the entire initial string of units can be reproduced only from the two or three coordinates of the last point, i.e., from only two (2) or three (3) numbers, respectively. Even the 1-D representation fulfills this condition, if besides its end point value an additional number is supplied. Our algorithm exploits the approach of Jeffrey,¹¹ who presented graphically DNA having 10,000–100,000 nucleic acids in a single planar square in which each of the four corners is assigned to one of the four nucleic acids A, T, G and C with the coordinates (–1,–1), (–1,1), (1,1), and (1,–1), respectively. To arrive at the graphical representation he started at the center of the square at the point (0,0) and moved halfway toward the corner assigned to the first nucleic acid in the sequence to be coded, and then he continued from this point halfway toward the corner of the second nucleic acid in the sequence, etc. In Figure 1 we illustrate the three initial steps obtained by following the DNA sequence ATG which can be seen in the first exon of β -globin of all 10 species listed in Table 1.

The assignment of four bases (A, C, T, and G) to one of the corners of the square in Figure 1 is of course arbitrary, leading to 24 (=4!) possible different coding schemes, four of which can be mapped onto other ones by rotation

* Corresponding author phone: 386-1-4760-279; fax: 386-1-4760-300; e-mail: jure.zupan@ki.si.

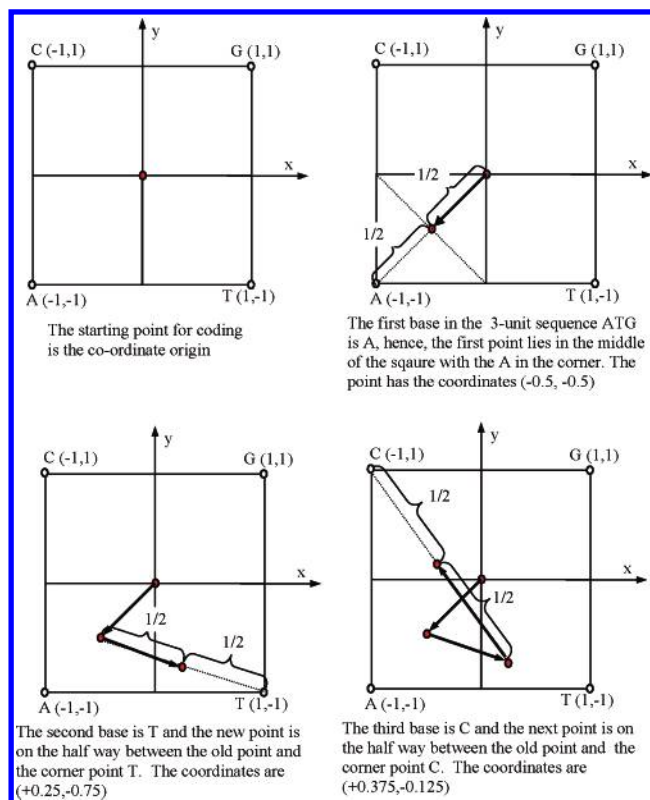


Figure 1. Coding the sequence ATC starting at the upper left and ending at the lower right corner.

producing only six different maps of which three can be mapped to itself by reflection (mirroring) leaving finally only three distinctly different zigzag curves. The three being associated with diagonally assigning A–G, A–T and A–C. By extending this coding scheme to 3-D using the assignment of four basic amino acids A, C, T, and G placed on the corners of the tetrahedron (Figure 2c), only two different (left and right oriented) variations of the coding can be obtained. Of course, one can consider representation of the DNA sequence in the 4-D space,^{19,20} that does not suffer from similar limitations, which may have an advantage for computer processing of the DNA data, but at the same time the visual interpretation and fast inspection of DNA sequences are lost.

The initial parameters of the algorithm from which all 2-D zigzag sequences in this work have been generated are the above-mentioned coordinate pairs of the four basic units (A, C, T, and G) given in Figure 1. For the examples given further on we have used the corners of a tetrahedron within the cube having the edges two units long and with the center at the coordinate origin. For 1-D representation only the x-coordinate values are used. The exact values of input parameters are given in Table 2.

It is interesting to note that the pure 1-D representation does not distinguish between pairwise bases adenine (A) and cytosine (C) on one side and thymine (T) and guanine (G) on the other side giving both bases in the pair the same starting point -1 or $+1$, respectively. By not differentiating between both constituent elements of the base pairs there is an inherent loss of information which in a symbolic way marks the binary nature of the information coded in the DNA or in the RNA carrier. This loss of information will be elaborated in more details later on.

Table 1. First Exon of β -Globin Gene for Ten Different Species

bovine	86 bases ATGCTGACTGCTGAGGAGAAGGCTGCCTGCACCG CCTTTTGGGGCAAGGTGAAAGTGATGAAG TTGGTGGTGAAGCCCTGGGCAG
gallus	92 bases ATGGTGCACCTGACTGCTGAGGAGAAGCAGCTCA TCACCGGCCTCTGGGGCAAGGTCAATGTGGC CGAATGTGGGGCCGAAGCCCTGGCCAG
goat	81 bases ATGGTGAAGTCTGCTGAGGAGAAGGCTGCCTGTCAC CGGCTTCTGGGGCAAGGTGAAATGGATGTTG TCTGAGGCCCTGGGCAG
gorilla	94 bases ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCC GTTACTGCCCTGTGGGGCAAGGTGAACGTGG ACGAAGTCGGTGGTGAAGCCCTGGGCAGGA
human	92 bases ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCG TTACTGCCCTGTGGGGCAAGGTGAACGTGGA TTAAGTTGGTGGTGAAGCCCTGGGCAG
lemur	90 bases ATGACTTTGCTGGTGCTGAGGAGAATGCTCATGT CACCCTCTGTGGGGCAAGGTGGATGTAGAG AAAGTTGGTGGCGAGGCCTGGGCAG
mouse	93 bases ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCT GTCTCTGCCTGTGGGGCAAGGTGAACCCGAT GAAGTTGGTGGTGAAGCCCTGGGCAG
opossum	92 bases ATGGTGCACCTGACTTCTGAGGAGAAGAAGTGCAC TCACTACCCTGTGGTCTAAGGTGCAGTTGAC CAGACTGGTGGTGAAGCCCTGGGCAG
rabbit	90 bases ATGGTGCATCTGTCCAGTGAAGGAGAAGTCTGCGG TCATGTCCCTGTGGGGCAAGGTGAATGTGGA AGAAGTTGGTGGTGAAGCCCTGGGC
rat	92 bases ATGGTGCACCTAAGTATGCTGAGAAGTCTACTG TTAGTGGCCTGTGGGGAAAGGTGAACCCCTGAT AATGTTGGCGCTGAAGCCCTGGGCAG

If the j -th unit of the N units long DNA sequence is written as $seq(j)$, then the recursive equation for obtaining the j -th element of the 3-D, 2-D, and 1-D “spectrum-like” representations $S(x_j, y_j, z_j)$ from its predecessor is the following:

$$S(x_{j+1}, y_{j+1}, z_{j+1}) = \frac{S(x_j, y_j, z_j) + S(x_{seq(j+1)}, y_{seq(j+1)}, z_{seq(j+1)})}{2} \quad (1a)$$

$$S(x_{j+1}, y_{j+1}) = \frac{S(x_j, y_j) + S(x_{seq(j+1)}, y_{seq(j+1)})}{2} \quad (1b)$$

$$S(x_{j+1}) = \frac{S(x_j) + S(x_{seq(j+1)})}{2} \quad (1c)$$

providing that the $S(x_0, y_0, z_0) = (0, 0, 0)$.

The recursion formulas (1a–1c) in the vector form are identical for all three dimensions. Table 3 and Figure 2 show the first 10 coordinate points in the 1-D, 2-D, and 3-D representation for the first 10 bases of the first exon of β -globin of human as calculated by the above equations. In general, the recursion formulas do not limit the number of basic units to four (A, C, T, and G for example), but any finite set of building blocks can be used. This is of interest when one considers graphical representations of proteins. The only prerequisite is that the positions in 1-D, 2-D, or 3-D space of all building blocks are known and fixed. For example

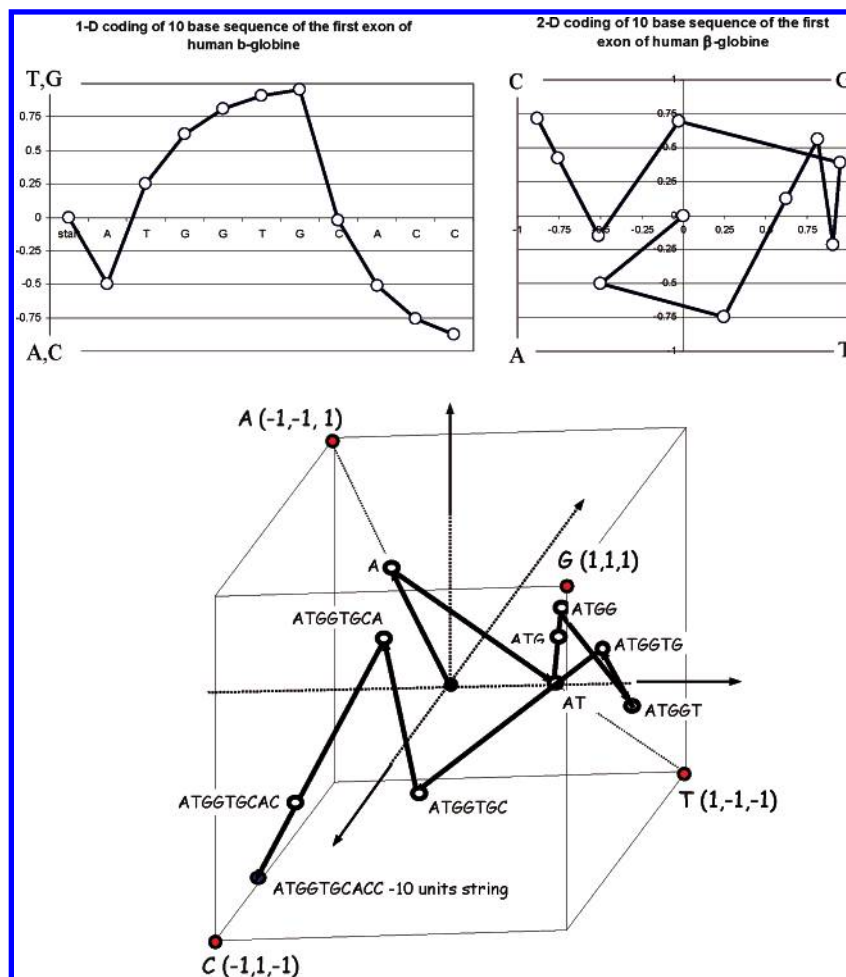


Figure 2. Three representations (1-D, 2-D, and 3-D) of the DNA sequence ATGGTGCACC.

Table 2. Starting Coordinates for Coding DNA Sequences in 3-D, 2-D, and 1-D, Representation, Respectively

	3-D	2-D	1-D
starting point	$(x_0, y_0, z_0) = (0, 0, 0)$	$(0, 0)$	(0)
adenine (A)	$(x_a, y_a, z_a) = (-1, -1, 1)$	$(x_a, y_a) = (-1, -1)$	$x_a = -1$
cytosine (C)	$(x_c, y_c, z_c) = (-1, 1, -1)$	$(x_c, y_c) = (-1, 1)$	$x_c = -1$
thymine (T)	$(x_t, y_t, z_t) = (1, -1, -1)$	$(x_t, y_t) = (1, -1)$	$x_t = 1$
guanine (G)	$(x_g, y_g, z_g) = (1, 1, 1)$	$(x_g, y_g) = (1, 1)$	$x_g = 1$

Table 3. Coordinates of the Consecutive Bases in the DNA Sequence ATGGTGCACC...

j	seq	x_j	b/w	y_j	z_j
0	start	0.00000000		0.00000000	0.00000000
1	A	-0.50000000	b, -, or 0	-0.50000000	0.50000000
2	T	0.25000000	b, -, or 0	-0.75000000	-0.25000000
3	G	0.62500000	w, +, or 1	0.12500000	0.37500000
4	G	0.81250000	w, +, or 1	0.56250000	0.68750000
5	T	0.90625000	b, -, or 0	-0.21875000	-0.15625000
6	G	0.95312500	w, +, or 1	0.39062500	0.42187500
7	C	-0.02343750	w, +, or 1	0.69531250	-0.28906250
8	A	-0.51171875	b, -, or 0	-0.15234375	0.35546875
9	C	-0.75585938	w, +, or 1	0.42382813	-0.32226563
10	C	-0.87792969	w, +, or 1	0.71191406	-0.66113281

^a The first 10 bases of the first exon of human β -globin. The labels "b", "-", and "0" in the 4-th column refer always to the "A" or "T" bases, while labels "w", "+", and "1" labels to the C and G bases, respectively.

the 64 possible triplets from AAA, AAC, ... TTT can be coded using 1-D recursion (equation 1c), if the triplets' positions are initialized on the x axis starting with AAA at

$x = -32$, AAC at $x = -31$, AAG at $x = -30$ and ending with TTG at $x = +31$ and TTT with $x = +32$. The point $x = 0$ should be taken as the starting point. Of, course, any other distribution of points or order of points is allowed.

The loss of information in the 1-D representation, mentioned above, has two solutions. First, by associating each point in the 1-D code (column 4 of Table 2) that is obtained by the discussed algorithm with black/white labels (or by a binary digit "0" or "1") depending to the sign of the y -coordinate in the 2-D code (column 3 in Table 2). Points having a coordinate y_j negative get the opposite color than the points having y_j positive. The labels "b", "-", and "0" in the 4-th column in Table 3 refer always to the "A" or "T" bases, while labels "w", "+", and "1" refer to the C and G bases, respectively, depending on the sign of coordinate x !

This, in a way a cumbersome solution, ensures the 1-D representation full reversibility, i.e., the possibility to reconstruct any DNA sequence from the three coordinates of the last point of the representation, no matter how long the DNA sequence is. The corollary is valid for 3-D, 2-D, and with the described black/white coloring of the points, as well for the 1-D representation.

The proof of the reversibility is straightforward: At each step of the forward coding process the coordinate values are halved which requires doubling the distance between the corner (or the end point) symmetrically over the current point to obtain the former point.

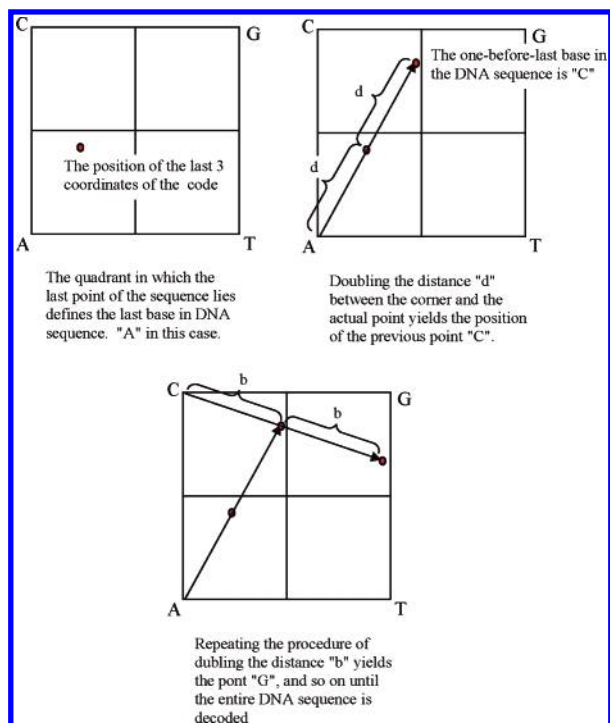


Figure 3. Decoding process of the 2-D representation. Providing the last point's precision to enough decimal places, the entire DNA sequence, no matter how long it is, can be decoded, from the last point's position. In principle the same procedure can be applied for 1-D coding. In the later case the supplementary binary number coding black/white points should be considered.

The decoding process starts with the triplet, doublet or singlet coordinate descriptors of the position of the last point depending on whether one has the 3-D, 2-D or 1-D code. The vicinity to the base acid corner point to the *position* of the coordinate triplet, doublet or singlet of the *last* point in determining the last base in the sequence. The direction defined by the two points, the base corner and the last coordinate values, determines the *direction* in which the one-before-last point is lying. Finally, the one-before-the-last base and its position in the coding sequence are quantitatively determined by *doubling* the distance from the corner point to the last point in the selected direction. This decoding procedure is schematically shown in Figure 3.

The second way to overcome the loss of information in the 1-D code is to expand the starting positions of four bases into four points on the x -axis. For example $(-2, -1, 1, \text{ and } 2)$ for A, C, G, and T, respectively, with the 0 point in the middle as the starting point for coding. Using the same recursion formula as before (equation 1c) one obtains a unique "spectrum-like" code for any DNA sequence (Figure 4). Such a "spectrum-like" representation of a certain DNA sequence segment can be further treated as any spectrum using the same assortment of all spectra-handling procedures such as Fast Fourier transformation, autocorrelation, peak-modulo search etc.

To show how the explained coding schemes can be exploited the hierarchical clustering of the codes of the first exons of 10 species (listed in Table 1) are shown in Figure 5. The dendrograms of the hierarchical clustering scheme have been obtained by the Ward method²¹ using the Euclidean distance as a similarity measure between two objects, i.e., between two "spectrum-like representations"

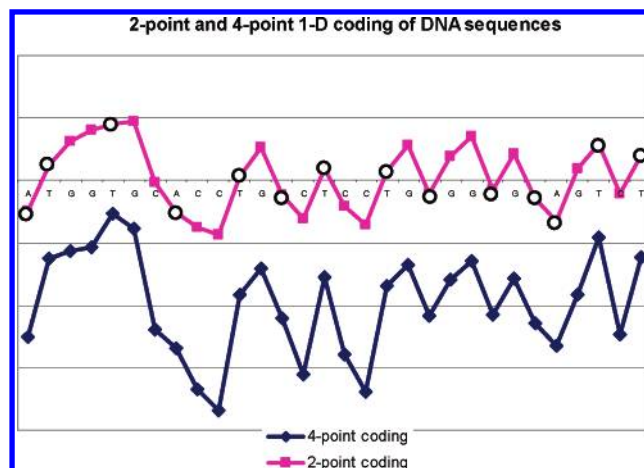


Figure 4. Coding of 25 base units of the first exon of human β -globin. The 4-point coding (below) takes the points $(x = -2, -1, 1, \text{ and } 2)$ for coding A, C, T, and G bases, while the 2-point one (above) takes for coding the four bases only 2 points $(-1, 1)$ for pairs A/C and T/G, respectively. Although the 4-point code seems to contain more information than the 2-point code, both need additional information for complete decoding. The open circles in the 2-point coding of the upper curve mark the negative "y" sign.

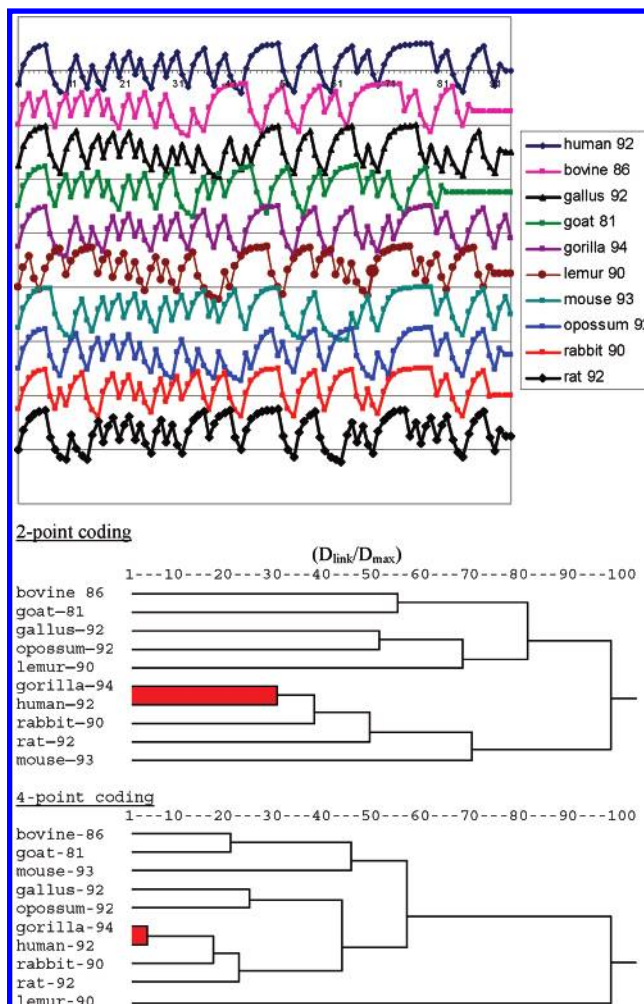


Figure 5. Ten DNA sequences given in Table 1 coded as 2-point "spectrum-like" representations (above). Clusters of 2- and 4-point "spectrum-like" representations were obtained using the Ward strategy (below).

obtained by the 2-point coding. All ten "spectra" as they result from the explained coding process (recursion formula

1c and DNA sequences given in Table 1) are shown in Figure 5. It has to be pointed out that the clusters were obtained actually on the “raw” data, i.e., with “spectrum-like” representations without any preprocessing like normalization, scaling, shifting or further transformation. Although the sequences of the first exons of different species are of different length, we did not attempt to cut them to the uniform length but have simply used the “zero-fill” up to the 96 points for all of them. The obtained clusters exhibit surprisingly meaningful groups. Human first exon of the β -globine is linked to the gorilla exon and both of them are in the same cluster together with the rat and rabbit. On the other hand the bovine's exon is linked to that of the goat. In general both the 2-point and 4-point coding schemes give very similar results.

CONCLUSION

In the present work the possibility offered by a simple recursion formula that transforms any DNA sequence to either 1-D, 2-D, or 3-D in visually more attractive, and from the data handling point of view more suitable representations are outlined. The fact that these representations can be decoded to the original string from their last point values only (with one supplemental figure for 1-D representation) strongly points to the fact that they contain the same amount of information as the original sequence. Therefore, if such a representation offers better handling and information extraction possibilities they are very valuable.

We are aware that the full decoding capability of the representations is due to the increase of the precision of the values with which the last point is given. The increase of the precision is about one decimal place after about three steps ($10 \sim 2^3$), i.e., it is identical to the binary coding of the amino acid pairs. In practice this means that the length of the mantissa of the final coordinate values yields the length of the DNA sequence. The decoding of the DNA sequence in this way is, hence, not a substantial gain or benefit. However, a different form which allows easier and uniform handling of sets of DNA sequences is very useful. The clusters shown in Figure 5 are used here only to stress this point. Such representations are able to reveal much valuable information if handled by standard spectroscopy dedicated procedures.

Another benefit of the recursion formula 1a–1c is that it is completely capable of transforming sequences consisting of any number of units (20 natural amino acids, for example) into 1-D, 2-D or 3-D “spectrum-like” or “zigzag” curves, respectively.

ACKNOWLEDGMENT

The financial support by the Research Program grant No. P1-0017 by the Ministry of Higher Education, Science, and Technology of Republic of Slovenia is gratefully acknowledged.

REFERENCES AND NOTES

- (1) Randić, M.; Nandy, A.; Basak, S. C.; Plavšić, D. On the numerical characterization of DNA primary sequences. *J. Math. Chem.* Submitted for publication.
- (2) Randić, M.; Vračko, M. On similarity of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 599–6064.
- (3) Randić, M.; Zupan, J. On Graphical Representations and Graph Theoretical Characterizations of DNA and Proteins. *Advances in Quantum Chemistry* (special issue on Chemical Graph Theory, D. J. Klein, guest editor), in press.
- (4) Randić, M.; Vračko, M.; Nandy, A.; Basak, S. C. On 3-D graphical representation of DNA primary sequences and their numerical characterization. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1235–1244.
- (5) X. Gou, X.; Randić, M.; Basak, S. C. A novel 2-D graphical representation of DNA sequences of low degeneracy. *Chem. Phys. Lett.* **2002**, *350*, 106–112.
- (6) Liu, Y.; Gou, X.; Xu, J.; Pan, L.; Wang, S. Some notes on 2-D graphical representation of DNA sequences. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 529–533.
- (7) Randić, M.; Lerš, N.; Plavšić, D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. *Chem. Phys. Lett.* **2003**, *368*, 1–6.
- (8) Randić, M.; Lerš, N.; Plavšić, D. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. *Chem. Phys. Lett.* **2003**, *371*, 202–207.
- (9) Randić, M.; Vračko, M.; Zupan, J.; Novič, M. Compact 2-D graphical representation of DNA. *Chem. Phys. Lett.* **2003**, *373*, 558–562.
- (10) Randić, M. Graphical representation of DNA as a 2-D map. *Chem. Phys. Lett.* **2004**, *386*, 468–471.
- (11) Jeffrey, H. I. Chaos game representation of gene structure. *Nucleic Acid Res.* **1990**, *18*, 2163–2170.
- (12) Goldman, N. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acid Res.* **1993**, *21*, 2487–2491.
- (13) Basu, S.; Pan, A.; Dutta, C.; Das, J. Chaos game representation of proteins. *J. Mol. Graphics Modelling* **1997**, *15*, 279–2897.
- (14) Randić, M.; Zupan, J.; Balaban, A. T. Unique graphical representation of protein sequences based on nucleotide triplet codons. *Chem. Phys. Lett.* **2004**, *397*, 247–252.
- (15) Randić, M.; Zupan, J. Highly compact 2-D graphical representation of DNA sequences. *SAR QSAR Environ. Res.* **2004**, *15*, 147–157.
- (16) Randić, M. 2-D Graphical representation of proteins based on virtual genetic code. *SAR QSAR Environ. Res.* **2004**, *15*, 191–205.
- (17) Randić, M. Condensed representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 50–56.
- (18) Randić, M. On characterization of DNA primary sequences by condensed matrix. *Chem. Phys. Lett.* **2000**, *317*, 29–342.
- (19) Randić, M.; Balaban, A. T. On a four-dimensional representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 532; errata: *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1724.
- (20) B. Liao, and T–M. Wang, General combinatorics of RNA hairpins and cloverleaves. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1138–1142.
- (21) Zupan, J. *Clustering of large data sets*; Research Studies Press J. Wiley: Chichester, 1982, or any other textbook on hierarchical clustering.

CI040104J