New spectral numerical characterization of DNA sequences

IGOR PESEK[a] and JANEZ ŽEROVNIK[a,b]

[a]Institute for mathematics, physics and mechanics,
Jadranska 19, 1000 Ljubljana, SLOVENIA
[b]University of Ljubljana, Faculty of Mechanical Engineering
Aškerčeva 6, 1000 Ljubljana, SLOVENIA
igor.pesek@uni-mb.si janez.zerovnik@imfm.uni-lj.si

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 was proposed in [14] 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. In this paper we compare two further numerical characterizations of the same type emphasizing the robustness of the approach.

Key-Words: numerical characterization, graph representation, graph invariant, DNA sequence

1 Introduction

Deoxyribonucleic acid (*DNA*) is the chemical inside the nucleus of all cells that carries the genetic instructions for making living organisms. A *DNA* molecule consists of two strands that wrap around each other to resemble a twisted ladder. The sides are made of sugar and phosphate molecules. The ”rungs” are made of nitrogen-containing chemicals called bases. Each strand is composed of one sugar molecule, one phosphate molecule, and a base. Four different bases are present in *DNA* - adenine (A), thymine (T), cytosine (C), and guanine (G). The particular order of the bases arranged along the sugar - phosphate backbone is called the *DNA* sequence; the sequence specifies the exact genetic instructions required to create a particular organism with its own unique traits. Each strand of the *DNA* molecule is held together at its base by a weak bond. The four bases pair in a set manner: Adenine (A) pairs with thymine (T), while cytosine (C) pairs with guanine (G).

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 [6] and later expanded by many others, see the review [24] and a number of more recent papers, for example [[10], [11], [12], [15], [16] [17][18],[24],[27], [3]] 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 [[15], [16], [17], [18], [24], [27]]. Similarly as topological indices used as molecular descriptors can dramatically improve the search for synthesis of compounds with a desired property [23], 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 accompanies any characterization based on invariants. The loss of the information, however, can in part be reduced by use of larger number of descriptors (invariants) [[19],[20]].

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 called 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}, $1 \le i < n$.

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 [[12], [18], [27]], 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 [[10], [11], [12], [15], [16], [17], [18], [24], [27]] and is taken from EMBL-EBI database [29]. This dataset is one of the primary tools for comparison of different graphical and numerical characterizations and was first used by Nandy [13] and later by other authors [10],[15],][16], [18], [24]]. 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 similar computations which are based on eigenvalues of the graphical representations [15], but are based on graphs, therefore our approach is using less computational effort.

For example in [15] one has to compute 12 different permutations of the graphical representation before the actual characterization, while our approach computes only one.

2 Modified Hamori curve representation

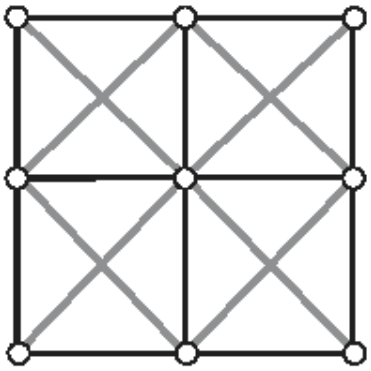
We based our research on DNA sequence representation introduced by Hamori [6]. In this method, the information content of a DNA sequence is mapped into a three-dimensional space function (H curve). The positive *x*-direction is used to count the number of bases in the sequence. At each point of *x* on 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.

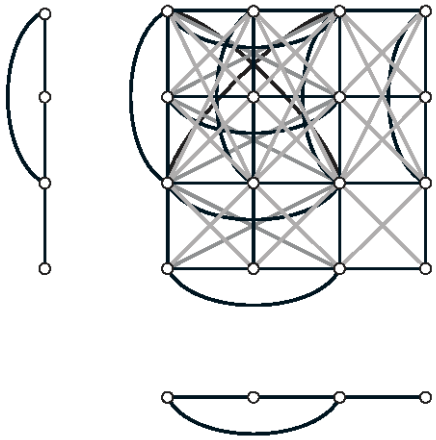
H-curve representation is sensitive to the directions chosen for four bases. For example representation with bases *ACGT* corresponding to four corners is different from *AGCT*, since the distance from base *A* to base *G* is different in this two cases.

We modified this approach by putting the corners of four bases on the complete graph 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 [9]. For example, the strong product of two edges (complete graphs on 2 vertices, K_2, is the complete graph on four vertices, K_4). Another example is the product of two paths of length 2



Below is depicted a product of two copies of a general graph together with the factors.



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:

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

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 final graph can be used [1]. 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.

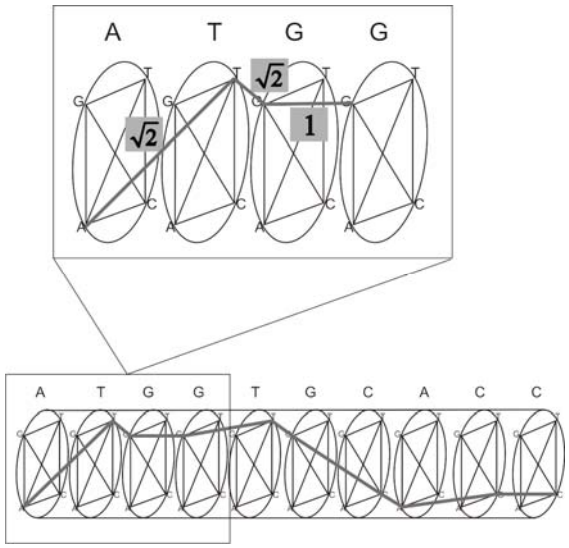


Fig. 1 Modified Hamori curve

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

Species & Coding sequence
Human (92 bases)
ATGGTGCACCTGACTCCTGAGGAGAAGTCT-GCCGTTACTGCCCTGTGGGGCAAGGTGAAC-GTGGAGTAAGTTGGTGGTGAGGCCCTGGGC-AG
Opossum (92 bases)
ATGGTGCACTTGACTTCTGAGGAGAAGAAC-TGCATCACTACCATCTGGTCTAAGGTGCAG-GTTGACCAGACTGGTGGTGAGGCCCTTGGC-AG
Gallus (92 bases)
ATGGTGCACTGGACTGCTGAGGAGAAGCAG-CTCATCACCGGCCTCTGGGGCAAGGTCAATG-TGGCCGAATGTGGGGCCGAAGCCCTGGCCA-G
Lemur (92 bases)
ATGACTTTGCTGAGTGCTGAGGAGAATGCTC-ATGTCACCTCTCTGTGGGGCAAGGTGGATGT-AG AGAAAGTTGGTGGCGAGGCCTTGGGCAG
Mouse (92 bases)
ATGGTGCACCTGACTGATGCTGAGAAGGCTG-CTGTCTCTTGCCTGTGGGGAAAGGTGAACTC-CGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Rabbit (90 bases)
ATGGTGCATCTGTCCAGTGAGGAGAAGTCTG-CGGTCACTGCCCTGTGGGGCAAGGTGAATGT-GGAAGAAGTTGGTGGTGAGGCCCTGGGC
Rat (92 bases)
ATGGTGCACCTAACTGATGCTGAGAAGGCTA-CTGTTAGTGGCCTGTGGGGAAAGGTGAACCC-TGATAATGTTGGCGCTGAGGCCCTGGGCAG
Gorilla (93 bases)
ATGGTGCACCTGACTCCTGAGGAGAAGTCT-GCCGTTACTGCCCTGTGGGGCAAGGTGAAC-GTGGATGAAGTTGGTGGTGAGGCCCTGGGC-AGG
Bovine (86 bases)
ATGCTGACTGCTGAGGAGAAGGCTGCCGTC-ACCGCCTTTTGGGGCAAGGTGAAAGTGGAT-GAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee (105 bases)
ATGGTGCACCTGACTCCTGAGGAGAAGTCT-GCCGTTACTGCCCTGTGGGGCAAGGTGAAC-GTGGATGAAGTTGGTGGTGAGGCCCTGGGC-AGGTTGGTATCAAGG

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 was proposed in [1] 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. Below we explain this and two further numerical characterizations of the same type.

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, Vračko, Lerš, and Plavšić [16]. 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 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 the ***L/L*** matrix as is shown in Table 1 where we used the first 6 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 $\frac{3}{\sqrt{2}+\sqrt{2}+1} \simeq 0.783$.

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

Base	A	T	G	G	T	G
A	*0*	*0.707*	*0.707*	*0.783*	*0.762*	*0.751*
T		*0*	*0.707*	*0.828*	*0.783*	*0.762*
G			*0*	*1.00*	*0.828*	*0.783*
G				*0*	*0.707*	*0.707*
T					*0*	*0.707*
G						*0*

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 [[4], [7]] and chemistry [8], 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 [19]. 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 [22].

Our characterizations are based eigenvalues of the matrix ***L/L***. In [14] the product of the 10 largest and 10 smallest eigenvalues was taken. Here we will compare this numerical characterization with the second which is the product of the five largest and five smallest eigenvalues. Species have different lengths of DNA sequence, shortest is DNA sequence of the bovine (*86* bases) and longest of the Chimpanzee (*105* bases).
It may be reasonable to consider ways to cancel out from comparison the influence of different lengths of sequences as much as possible. Therefore we also consider a normalized characterization, where we take the *n*-th root of the product of the eigenvalues.

$\Lambda_1(x)$	product of 10 largest and 10 smallest eigenvalues
$\Lambda_2(x)$	product of 5 largest and 5 smallest eigenvalues
$\Lambda_n(x)$	$\sqrt[n]{\Lambda_2(x)}$

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

We will illustrate a natural method for the 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 DNA sequences are similar if the corresponding difference between the value of the numerical characterization is small.

The values of the numerical characterizations are as follows:

species	$\Lambda_1(x)$	$\Lambda_2(x)$	$\Lambda_n(x)$
human	0.0121903	0.411371	0.990391
chimpanzee	0.0144888	0.597189	0.995102
gorilla	0.0128425	0.44091	0.991233
opossum	0.00357022	0.0997373	0.975255
gallus	0.00864563	0.285579	0.98647
lemur	0.00220456	0.0610567	0.970066
mouse	0.0533748	1.68273	1.00567
rabbit	0.00692271	0.243408	0.984422
rat	0.0192638	0.756461	0.996971
bovine	0.112602	2.89767	1.01245

Formally we can define similarity relations as:

$$S_\star(x,y) = |\Lambda_\star(x) - \Lambda_\star(y)|,$$

where x, y are sequences of the species.

In this way we obtain a matrix of mutual similarities among species. First we present a matrix with similarities (because of the size, we present them at the end of paper) and then we draw two graphs based on the values in the matrix. First graph represents nearest similarities relation and second graph represents widest dissimilarities relation. Similarities/dis-similarities matrix for $\Lambda_1(x)$ is in Table 3 and corresponding graphs are on Fig 2 and 3. Similarities/dissimilarities matrix for $\Lambda_2(x)$ is in Table 4 and corresponding graphs are on Fig 4 and 5 and finally, similarities/dissimilarities matrix for $\Lambda_n(x)$ is in Table 5 and corresponding graphs are on Fig 6 and 7.

While of course not surprisingly the three similarity measures give different numerical values, the overall results are not very much different. In particular, the smallest differences 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 [[10], [16]]. On the other hand, the largest entries in the similarity/dissimilarity matrix appear in rows belonging to bovine and opossum.

We may conclude that all presented numerical characterizations have captured some important features of the DNA sequences considered.

4 Conclusion

Our objective in [14] was to arrive at a numerical characterization of DNA sequences. This may be accomplished in a relatively simple algebraic manner and as such makes the proposed approach very attractive for the characterization of DNA sequences having 1,000 or more bases.

In this follow-up report we add results on related numerical characterizations showing that the approach is robust, hence the somewhat arbitrary choice of 5 or 10 eigenvalues taken does not severely influence the results of the method.

The preliminary results presented here support the intuition that some important structural information of the sequences is encoded in the spectrum, and in particular in the largest and smallest eigenvalues. We have provided a method that is computationally more efficient than some earlier approaches. Needles to say that the outlined approach may be 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 bioinformatics.

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. Vol. 4 1994 pp. 235-282.

[3] D. Bielinska-Waz, T. Clark, P. Waz, W. Nowak, A. Nandy, *2D-dynamic representation of DNA sequences*, Chem. Phys. Lett. Vol. 442 2007 pp. 140-144.

[4] N.G. Blas, E. Santos, M. A. Diaz, *Clustering over DNA Strings,* Proceedings of the 5th WSEAS Int. Conf. on Computational inteligence, man-machine systems and cybernetics. Venice, November 2006

[5] C. D. Godsil, D. A. Holton, B. D. McKay, *The spectrum of a graph, Combinatorial Mathematics V*, Lect. Notes. Math. Vol. 622 1977 pp. 91-117.

[6] E. Hamori, *Graphical representation of long DNA sequences by methods of H curves, current results and future aspects*. Biotechniques Vol. 7 1989 pp. 710-720.

[7] F. Harary, C. King, A. Mowshowitz, R.C. Read, *Cospectral graphs and digraphs*, Bull. London Math. Soc. Vol. 3 1971 pp. 321--328.

[8] W. C. Herndon, M. L. Ellzey, *Isospectral graphs and molecules*, Tetrahedron Vol. 31 1975 pp. 99-107.
[9] W. Imrich, S. Klavžar, *Product Graphs: Structure and Recognition*, John Wiley & Sons, New York, 2000.
[10] B. Liao, T. Wang, *Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation*,. Phys. Lett. Vol. 388 2004 pp. 195-200.
[11] B. Liao, T. Wang, *3-D graphical representation of DNA sequences and their numerical characterization*, J. Mol. Struc. - Theochem Vol. 681 2004 pp. 209-212.
[12] 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 Vol. 717 2005 pp. 199-203.
[13] A. Nandy, *A new graphical representation and analysis of DNA sequence structure*, Curr. Sci. India Vol. 66 (1994) pp. 309-314.
[14] I. Pesek, J. Žerovnik, *A numerical characterization of modified Hamori curve representation of DNA sequences*, MATCH Commun. Math. Comput. Chem. Vol. 60 2008 pp. 301-312.
[15] M. Randić, M. Vračko, N. Lerš, D. Plavšić, *Novel 2-D graphical representation of DNA sequences and their numerical characterization*, Chem. Phys. Lett. Vol. 368 2003 pp. 1-6.
[16] 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. Vol. 371 2003 pp. 202-207.
[17] M. Randić, M. Vračko, J. Zupan, M. Novič, *Compact 2-D graphical representation of DNA*, Chem. Phys. Lett. Vol. 373 2003 pp. 558-562.
[18] 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. Vol. 407 2005 pp. 205-208.
[19] M. Randić, M. Vračko, *On the similarity of DNA primary sequences*, J. Chem. Inf. Comput. Sci. Vol. 40 2000 pp. 599-606.
[20] M. Randić, M. Vračko, A. Nandy, S.C. Basak, *On 3D graphical representation of DNA primary sequences and their numerical characterization*,. Chem. Inf. Comput. Sci. Vol. 40 2000 pp. 1235-1244.
[21] M. Randić, *Condensed Representation of DNA Primary Sequences*, J. Chem. Inf. Comput. Sci. Vol. 40 2000 pp.50-66.
[22] M. Randić, S.C. Basak, *Characterization of DNA Primary Sequences Based on the Average Distances between Bases*, J. Chem. Inf. Model. Vol. 41 2001 pp. 561 - 568.
[23] D. Rouvray, *Predicting Chemistry from Topology*, Sci. Am. Vol. 254 1986 pp. 40-47.
[24] A. Roy, C. Raychaudhury, A. Nandy, Novel techniques of graphical representation and analysis of DNA sequences - A review, J. Bioscience Vol. 23 1998 pp. 55-71.
[25] M. Saeb, E. El-Abd, M. E. El-Zanaty, *DNA Steganography using DNA Recombinant and DNA Mutagenesis Techniques*, WSEAS TRANSACTIONS on COMPUTER RESEARCH, Vol. 2, 2007 pp 50 – 56.
[26] J. Zupan, M. Randić, *Algorithm for Coding DNA Sequences into "Spectrum-like" and "Zigzag" representations*, J. Chem. Inf. Model. Vol. 45 2005 pp. 309--313.
[27] Y. Yao, X. Nan, T. Wang, *Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation*, Chem. Phys. Lett. Vol. 411 2005 pp. 248-255.
[28] Y. Yamada, K. Satou, *Prediction of Genomic Methylation Status on CpG Islands Using DNA Sequence Features*, WSEAS TRANSACTIONS on BIOLOGY and BIOMEDICINE, Vol. 5, 2008, pp 153 – 162.
[29] http://www.ebi.ac.uk/

Table 3 Matrix of the $\Lambda_1(x)$ similarities

Human	Chimpanzee	Gorilla	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Bovine
0	0.00229851	0.000652253	0.00862008	0.00354466	0.00998573	0.0411845	0.00526758	0.0070735	0.100412
0.00229851	0	0.00164626	0.0109186	0.00584317	0.0122842	0.038886	0.00756609	0.00477499	0.0981132
0.000652253	0.00164626	0	0.00927233	0.00419692	0.010638	0.0405323	0.00591984	0.00642125	0.0997594
0.00862008	0.0109186	0.00927233	0	0.00507541	0.00136566	0.0498046	0.00335249	0.0156936	0.109032
0.00354466	0.00584317	0.00419692	0.00507541	0	0.00644107	0.0447292	0.00172292	0.0106182	0.103956
0.00998573	0.0122842	0.010638	0.00136566	0.00644107	0	0.0511703	0.00471815	0.0170592	0.110397
0.0411845	0.038886	0.0405323	0.0498046	0.0447292	0.0511703	0	0.0464521	0.034111	0.0592271
0.00526758	0.00756609	0.00591984	0.00335249	0.00172292	0.00471815	0.0464521	0	0.0123411	0.105679
0.0070735	0.00477499	0.00642125	0.0156936	0.0106182	0.0170592	0.034111	0.0123411	0	0.0933382
0.100412	0.0981132	0.0997594	0.109032	0.103956	0.110397	0.0592271	0.105679	0.0933382	0

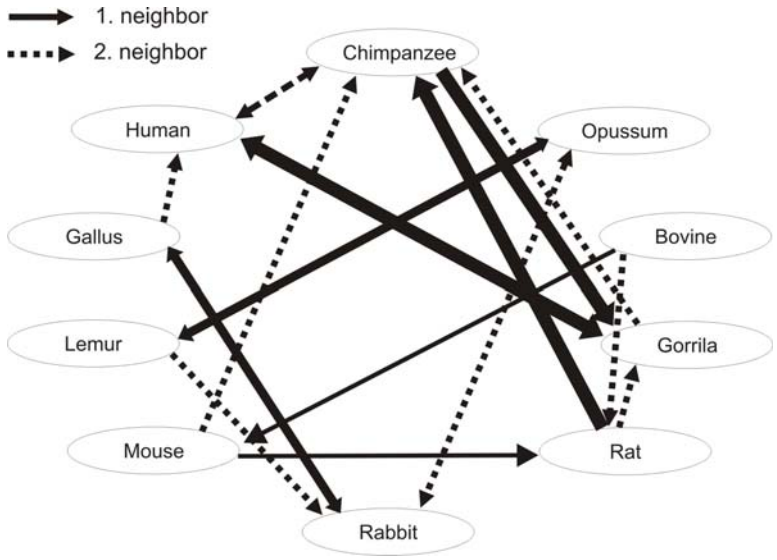


Fig. 2 Largest $\Lambda_1(x)$ similarities

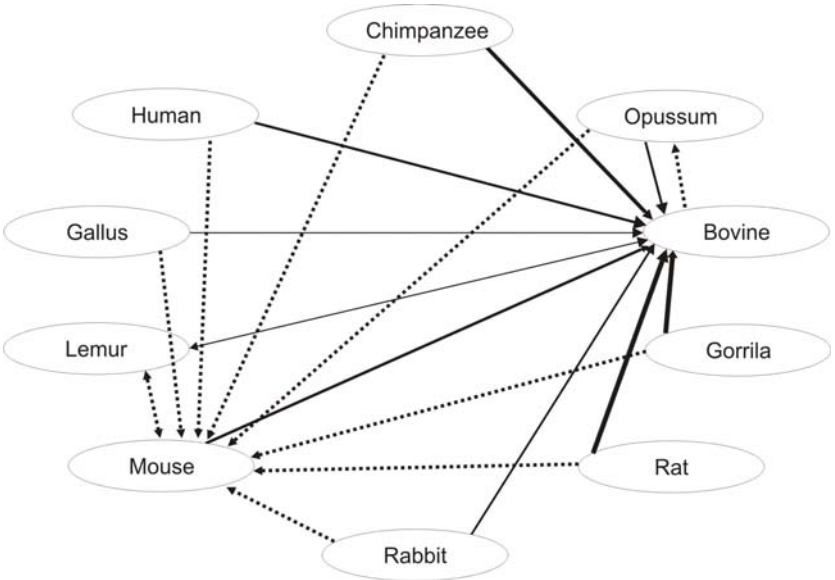


Fig. 3 Largest $\Lambda_1(x)$ dissimilarities

Table 4 Matrix of the $\Lambda_2(x)$ similarities

Human	Chimpanzee	Gorilla	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Bovine
0	0.185818	0.029539	0.311634	0.125792	0.350314	127.136	0.167963	0.34509	2.4863
0.185818	0	0.156279	0.497452	0.31161	0.536133	108.554	0.353782	0.159271	2.30048
0.029539	0.156279	0	0.341173	0.155331	0.379853	124.182	0.197502	0.315551	2.45676
0.311634	0.497452	0.341173	0	0.185842	0.0386806	1.583	0.14367	0.656723	2.79793
0.125792	0.31161	0.155331	0.185842	0	0.224523	139.715	0.0421717	0.470881	2.61209
0.350314	0.536133	0.379853	0.0386806	0.224523	0	162.168	0.182351	0.695404	2.83661
127.136	108.554	124.182	1.583	139.715	162.168	0	143.933	0.926273	1.21493
0.167963	0.353782	0.197502	0.14367	0.0421717	0.182351	143.933	0	0.513053	2.65426
0.34509	0.159271	0.315551	0.656723	0.470881	0.695404	0.926273	0.513053	0	2.14121
24.863	230.048	245.676	279.793	261.209	283.661	121.493	265.426	214.121	0

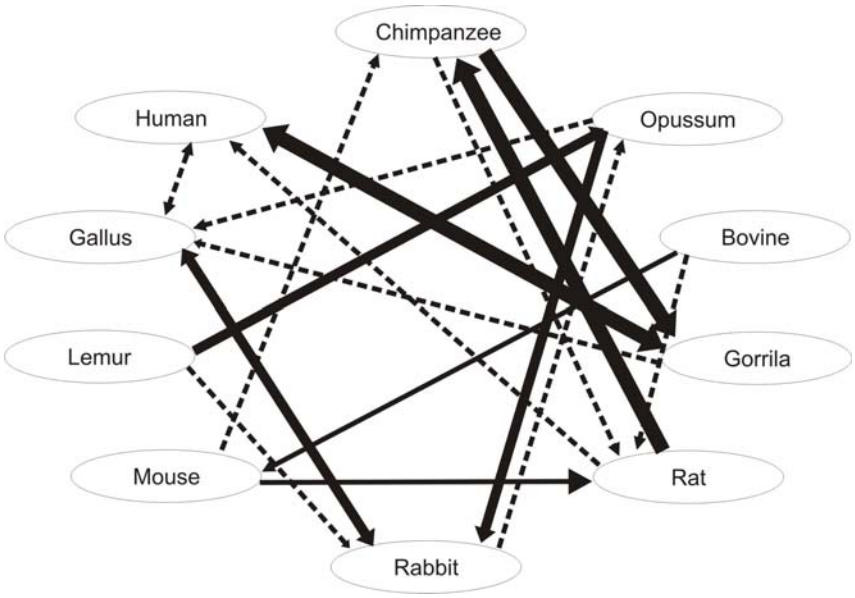


Fig. 4 Largest $\Lambda_2(x)$ similarities

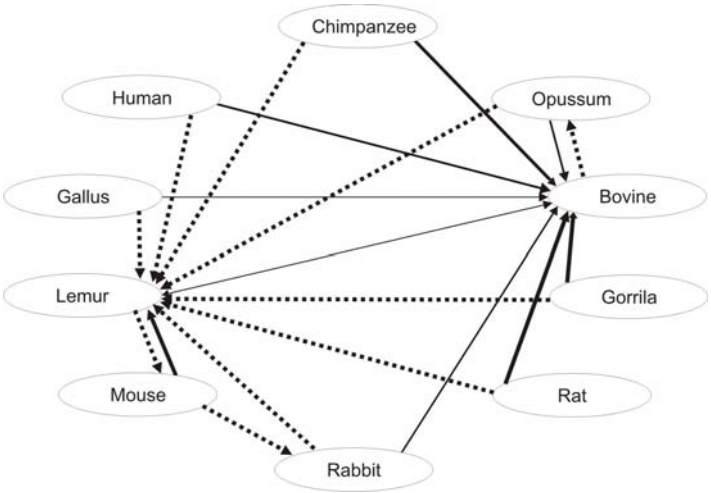


Fig. 5 Largest $\Lambda_2(x)$ dissimilarities

Table 5 Matrix of the $\Lambda_n(x)$ similarities

Human	Chimpanzee	Gorilla	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Bovine
0	0.004710	0.000841	0.015136	0.003921	0.020325	0.015281	0.005969	0.006579	0.022056
0.004710	0	0.003869	0.019847	0.008632	0.025035	0.010570	0.010679	0.001868	0.017345
0.000841	0.003869	0	0.015978	0.004762	0.021166	0.014439	0.006810	0.005737	0.021214
0.015136	0.019847	0.015978	0	0.011215	0.005188	0.030418	0.009167	0.021716	0.037193
0.003921	0.008632	0.004762	0.011215	0	0.016403	0.019202	0.002047	0.010500	0.025977
0.020325	0.025035	0.021167	0.005188	0.016403	0	0.035606	0.014356	0.026904	0.042381
0.015281	0.010570	0.014439	0.030418	0.019202	0.035606	0	0.021250	0.008701	0.006775
0.005969	0.010679	0.006810	0.009167	0.002047	0.014356	0.021250	0	0.012548	0.028025
0.006579	0.001868	0.005737	0.021716	0.010500	0.026904	0.008701	0.012548	0	0.015477
0.022056	0.017345	0.021214	0.037193	0.025977	0.042381	0.006775	0.028025	0.015477	0

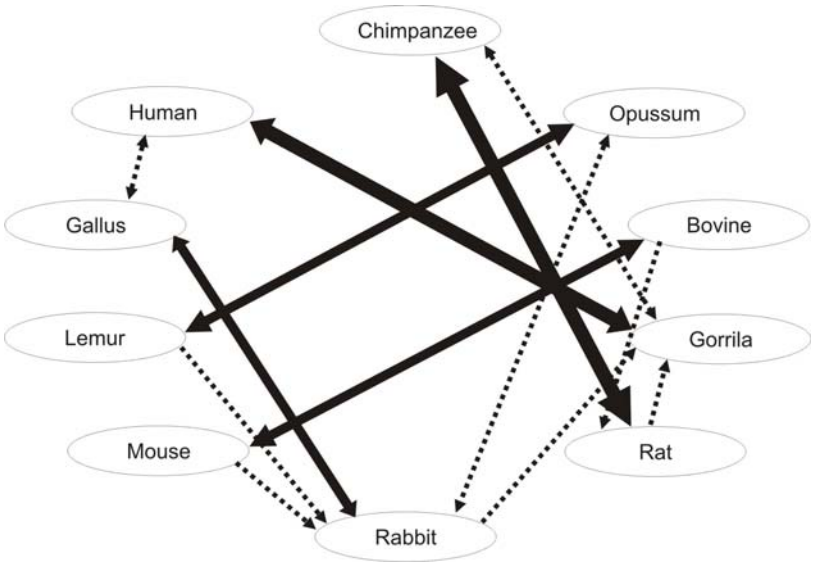


Fig. 6 Largest $\Lambda_n(x)$ similarities.

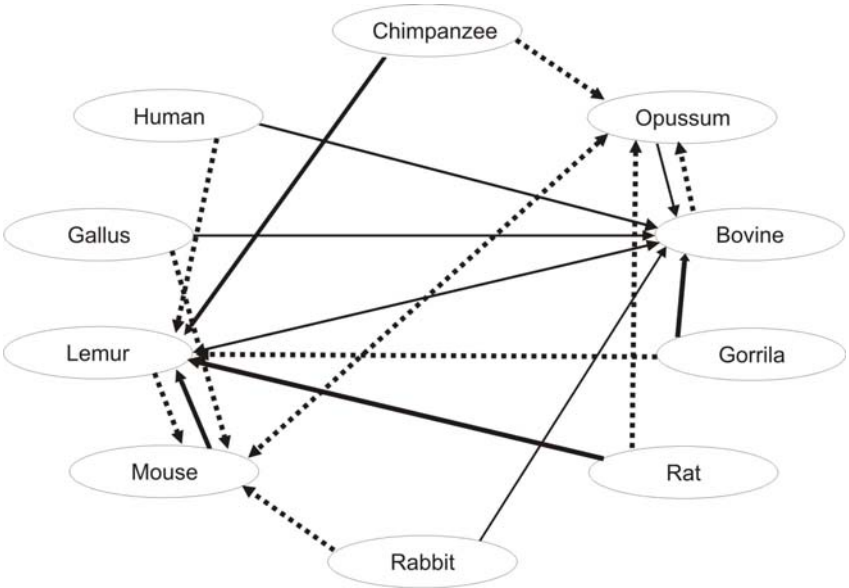


Fig. 7 Largest $\Lambda_n(x)$ dissimilarities.