On spectral numerical characterizations of DNA sequences based on Hamori curve representation

IGOR PESEK \(^a\)  
\(^a\)Institute of mathematics, physics and mechanics  
Jadranska 19, 1000 Ljubljana  
SLOVENIA  
igor.pesek@imfm.uni-lj.si

JANEZ ŽEROVNIK \(^a, b\)  
\(^b\)University of Ljubljana  
Faculty of Mechanical Engineering  
Aškerčeva 6, 1000 Ljubljana  
SLOVENIA  
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, \(K_4 \square P_n\), with weighted edges. Based on this representation, a novel numerical characterization was proposed in [Pesek and Žerovnik, MATCH Commun. Math. Comput. Chem. (Krag.), vol 60, 2008, 301-312] 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

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 [5] and later expanded by many others, see the review [23] and a number of more recent papers, for example [9, 10, 11, 14, 15, 16, 17, 24, 25, 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, 23, 25]. Similarly as topological indices used as molecular descriptors can dramatically improve the search for synthesis of compounds with a desired property [22], 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) [18, 19].

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 [10, 14, 17, 25], 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 $\beta$-globin gene for the 10 different species, a dataset shown in Table 1, that is used in many recent studies [9, 10, 11, 14, 15, 16, 17, 24, 25] and is taken from EMBL-EBI database [26]. This dataset is one of the primary tools for comparison of different graphical and numerical characterizations and was first used by Nandy [12] and later by other authors [9, 14, 15, 17, 24]. The reason why Nandy decided to use this gene lies in the fact that $\beta$ 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 [14], but are based on graphs, therefore our approach is using less computational effort. For example in [14] 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 [5]. 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 [8].

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:

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

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 [13] 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 et al. [15]. 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 \(L/L\) matrix as is shown in Table 2 where we used first 6 bases of the first exon of \(\beta\)-globin gene of human. For example, the first three entries of the first row are \(1/\sqrt{2} \approx 0.707\), \(2/\sqrt{2+\sqrt{2}} \approx 0.707\), and \(3/\sqrt{2+\sqrt{2+\sqrt{2}}} \approx 0.783\).

#### Table 1: The coding sequences of the first exon of \(\beta\)-globin gene of 10 different species

<table>
<thead>
<tr>
<th>Species</th>
<th>Coding sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human (92 bases)</td>
<td>ATGGTGCACTCTGCCCTGAGAAGAAATCTGCAG-</td>
</tr>
<tr>
<td>Opossum (92 bases)</td>
<td>ATGGTGCACTCTGCCCTGAGAAGAAATCTGCAG-</td>
</tr>
<tr>
<td>Gallus (92 bases)</td>
<td>ATGGTGCACTCTGCCCTGAGAAGAAATCTGCAG-</td>
</tr>
<tr>
<td>Lemur (92 bases)</td>
<td>ATGGTGCACTCTGCCCTGAGAAGAAATCTGCAG-</td>
</tr>
<tr>
<td>Mouse (92 bases)</td>
<td>ATGGTGCACTCTGCCCTGAGAAGAAATCTGCAG-</td>
</tr>
<tr>
<td>Rabbit (90 bases)</td>
<td>ATGGTGCACTCTGCCCTGAGAAGAAATCTGCAG-</td>
</tr>
<tr>
<td>Rat (92 bases)</td>
<td>ATGGTGCACTCTGCCCTGAGAAGAAATCTGCAG-</td>
</tr>
<tr>
<td>Bovine (86 bases)</td>
<td>ATGGTGCACTCTGCCCTGAGAAGAAATCTGCAG-</td>
</tr>
<tr>
<td>Chimpanzee (105 bases)</td>
<td>ATGGTGCACTCTGCCCTGAGAAGAAATCTGCAG-</td>
</tr>
</tbody>
</table>

#### Table 2: The upper triangle of the \(L/L\) matrix of the sequence ATGGTGCACT

<table>
<thead>
<tr>
<th>Base</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>G</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.707</td>
<td>0.707</td>
<td>0.783</td>
<td>0.762</td>
<td>0.751</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.707</td>
<td>0.828</td>
<td>0.783</td>
<td>0.762</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0.828</td>
<td>0.783</td>
<td>0.762</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0.707</td>
<td>0.707</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.707</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0.707</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Formally, we assign the matrix \(LL_x\) to the se-
4 Similarities/dissimilarities among the coding sequences of the first exon of β-globin gene of different species

We will illustrate the use 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:

<table>
<thead>
<tr>
<th>species</th>
<th>$\Lambda_1(x)$</th>
<th>$\Lambda_2(x)$</th>
<th>$\Lambda_n(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>0.0121903</td>
<td>0.411371</td>
<td>0.990391</td>
</tr>
<tr>
<td>shimpia</td>
<td>0.0144888</td>
<td>0.597189</td>
<td>0.991233</td>
</tr>
<tr>
<td>gorilla</td>
<td>0.0128425</td>
<td>0.44091</td>
<td>0.991233</td>
</tr>
<tr>
<td>oposum</td>
<td>0.00357022</td>
<td>0.0997373</td>
<td>0.975255</td>
</tr>
<tr>
<td>galus</td>
<td>0.00864563</td>
<td>0.285579</td>
<td>0.98647</td>
</tr>
<tr>
<td>lemur</td>
<td>0.00220456</td>
<td>0.0610567</td>
<td>0.970066</td>
</tr>
<tr>
<td>mouse</td>
<td>0.0533748</td>
<td>1.68273</td>
<td>1.00567</td>
</tr>
<tr>
<td>rabbit</td>
<td>0.00692271</td>
<td>0.243408</td>
<td>0.984422</td>
</tr>
<tr>
<td>rat</td>
<td>0.0192638</td>
<td>0.756461</td>
<td>0.996971</td>
</tr>
<tr>
<td>bovine</td>
<td>0.112602</td>
<td>2.89767</td>
<td>1.01245</td>
</tr>
</tbody>
</table>

Formally we can define similarity relations as:

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

where $x, y$ are sequences of the species. In this way we obtain a matrix of mutual similarities among species. Instead of writing up the matrices with similarities (see [13] for the $\Lambda_1(x)$ case) we rather draw graphs of largest similarities and dissimilarities for all three numerical characterizations defined above. The thickness of the arrows is proportional to the similarity.

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 [9, 15]. 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.
Figure 2: Largest $\Lambda_1(x)$ similarities.

Figure 3: Largest $\Lambda_1(x)$ dissimilarities.

Figure 4: Largest $\Lambda_2(x)$ similarities.

Figure 5: Largest $\Lambda_2(x)$ dissimilarities.

Figure 6: Largest $\Lambda_n(x)$ similarities.

Figure 7: Largest $\Lambda_n(x)$ dissimilarities.

5 Conclusion

Our objective in [13] 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 two 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 the bioinformatics.
References:


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