GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

A novel technique for analyzing the similarity and dissimilarity of DNA sequences

Y.W. Liu[1] and Y. Peng[2]

[1]College of Science, Hunan Agricultural University, Hunan, China
[2]Key Laboratory for Crop Germplasm Innovation and Utilization of Hunan Province, Hunan Agricultural University, Hunan, China

Corresponding author: Y. Peng
E-mail: pengyan3759@163.com

Genet. Mol. Res. 13 (1): 570-577 (2014)
Received November 7, 2012
Accepted January 29, 2013
Published January 28, 2014
DOI http://dx.doi.org/10.4238/2014.January.28.2

ABSTRACT. $l_{i,j}$ denotes the distance between the point (x_i, y_i) and the point (x_j, y_i) in graphical representation. By classifying $l_{i,j}$, $i, j = 1, 2, \ldots, N$ according to the number of points between (x_i, y_i) and (x_j, y_i), $N - 1$ types are obtained. The average and variance of every type are assembled by the novel invariant $v = (a_1, d_1, a_2, d_2, \ldots, a_N, d_N)$. Compared with the traditional invariants, the leading eigenvalue, the max-min (eigenvalue), the leading eigenvalue/N, the average matrix element, and the average row sum, this strategy complies with the rule of using the average, extracts more information about biological sequences, and reduces the amounts of computation. It is superior to the traditional invariants in predicting similarity and dissimilarity among different species.

Key words: Graphical representation; Sequence comparison; Invariant; Average and variance; DNA sequence

INTRODUCTION

The rapid growth of sequence data in the DNA sequence databanks has called for the development of suitable techniques for rapid viewing and analysis of the data. In particular, graphical representations of DNA sequence have emerged as a very powerful technology for viewing, sorting, and comparing various gene structures with an intuitive feel (Nandy et al., 2006).

Nandy (1994) first presented the 2-D graphical representation by assigning the four types of bases (adenine (A), guanine (G), thymine (T), and cytosine (C)) to the four directions of Cartesian coordinate axes. However, some loss of visual information that is associated with crossing and overlapping of the curve with itself accompanies this method. Randić presented a novel 2-D graphical representation, in which A, G, T, and C are assigned to four symmetric non-equivalent horizontal lines (Randić et al., 2003). This method resolves the degeneracy of DNA sequences and it is mathematically proven to eliminate circuit formation. Subsequently, Yau et al. (2003), Liao (2005), Huang et al. (2009) also proposed 2-D graphical representations without degeneracy in the Cartesian coordinate system. Some researches improved and applied the graphical representation (Liao and Wang, 2004a; Chi and Ding, 2005; Qi and Qi, 2007; Guo et al., 2008; Wu et al., 2011). The graphical representation of proteins was also studied by some researchers (He et al., 2011; Randić et al., 2011).

In order to quantitatively compare the DNA sequences and determine the similarity/dissimilarity among them, the graphical representation must be transformed into mathematical objects such as an E matrix, M/M matrix, L/L matrix, and their “high order” matrices. Then, the invariants of these matrices are extracted to numerically characterize the biological sequences, and the traditional invariants comprise the leading eigenvalue, the max -min (eigenvalue), the leading eigenvalue/N, the average matrix element, and the average row sum. However, when the length of the DNA sequence is very long, the size of the mathematical objects becomes astonishing. It will take computers much computing time and memory space to compute the eigenvalues. On the other hand, the idea of the average matrix element takes the average value of all distances of line segments that connect two points corresponding to two bases in graphical representation. However, the lengths of these distances change obviously. For example, lengths of distances connecting two adjacent points are small, but the lengths of distances connecting two points are big if the number of points between them is large. Thus, the average of these distances may lose some information.

In this study, we have outlined a procedure to give a novel invariant of a mathematical object by graphical representation. This can avoid the complex computing of eigenvalues, comply with the rule of using the average, and extract more information from biological sequences. This procedure is also superior to the traditional invariants for analyzing similarity/dissimilarity among sequences.

MATERIAL AND METHODS

2-D graphical representation

In this letter, we used the 2-D graphical representation proposed by Yau et al. (2003), which is shown in Figure 1. In the Cartesian coordinate system, the vectors $(\frac{1}{2},-\frac{\sqrt{3}}{2})$, $(\frac{\sqrt{3}}{2},-\frac{1}{2})$, $(\frac{\sqrt{3}}{2},\frac{1}{2})$, and $(\frac{1}{2},\frac{\sqrt{3}}{2})$ are used to represent four nucleotides A, G, C, and T. Each nucleotide of a DNA sequence walks as one of the above four vectors in a 2-D Cartesian coordinate system. Thus, a graphical curve corresponding to the DNA sequence is obtained, which is mathematically proved to eliminate circuit formation. For example, a DNA sequence s = CTGAGCTGCA is considered. The graphical representation of the DNA sequence is shown in Figure 2. $\left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right)$, $\left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right)$, $\left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right)$, and $\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$ were used to represent four nucleotides A, G, C, and T.

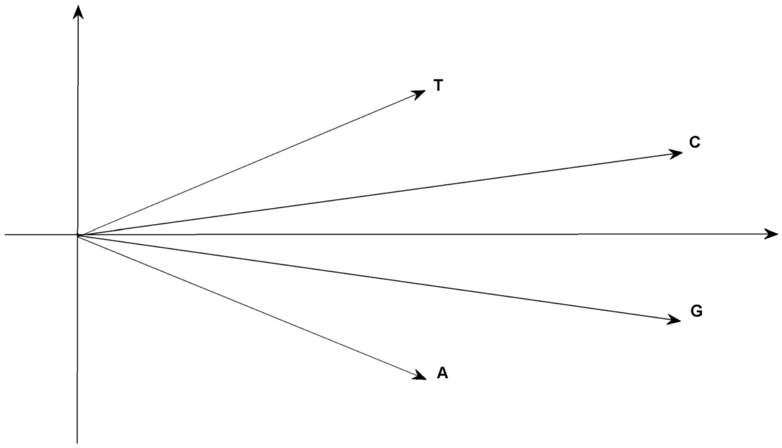

Figure 1. Unit vectors designed by Yau et al. (2003) in Cartesian coordinate plane.

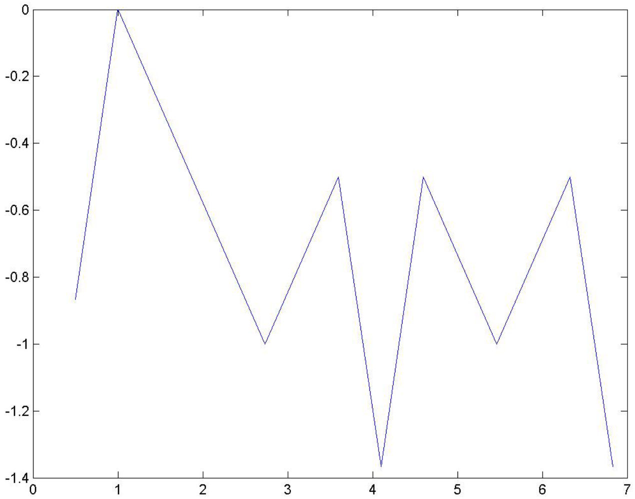

Figure 2. Graphical representation for the DNA sequences (ATGGCATGCA).

Mathematical objects

In order to obtain the numerical characterization of DNA sequences, many researchers associate the curves of graphical representations with mathematical objects. The E matrix, M/M matrix, and L/L matrix were introduced by Randić et al. (2003) The definitions of the E matrix, M/M matrix, and L/L matrix are as follows:

$$E_{i,j} = l_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \ ,$$

$$(M / M)_{i,j} = \frac{E_{i,j}}{|i-j|} \ ,$$

$$(L / L)_{i,j} = \begin{cases} \frac{E_{i,j}}{\sum_{k=i}^{j-1} E_{k,(k+1)}}, & \text{if } i \leq j \\ \frac{E_{ij}}{\sum_{k=j}^{i-1} E_{k,(k+1)}}, & \text{otherwise} \end{cases} .$$

Recently, these matrices were widely used to analyze biological sequences (Nandy et al., 1994; Randić et al., 2003; Liao and Wang, 2004b; Qi and Qi, 2007; Wu et al., 2011).

Traditional invariants

After the mathematical object of a graphical representation is constructed, the invariant of the matrix is extracted to numerically characterize the biological sequence. There are two categories of invariants. The first category is concerned with the eigenvalue, in which there are three types: the leading eigenvalue, the max (eigenvalue)-min (eigenvalue), and the leading eigenvalue/N. Consider the above E matrix with the size of N*N. The process of computing eigenvalues is as follows:

$$|E - \lambda I| = \begin{vmatrix} -\lambda & l_{1,2} & l_{1,3} & \cdots & l_{1,N} \\ l_{2,1} & -\lambda & l_{2,3} & \cdots & l_{2,N} \\ l_{3,1} & l_{3,2} & -\lambda & \cdots & l_{3,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{N,1} & l_{N,2} & l_{N,3} & \cdots & -\lambda \end{vmatrix} = 0$$

Unfortunately, no formula may be used to solve above N-degree (N > 5) polynomials. The approximate computation must be considered. However, increases in the degree of the polynomial augment the amount of computation and magnify the error of the computation.

The second category is concerned with the average, in which there are two types:
the average matrix element: $\bar{E} = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} l_{i,j}$ and
the average row sum: $\bar{E}_i = \frac{1}{N}\sum_{j=1}^{N} l_{i,j}$.

These methods can dramatically decrease the amount of computation compared with the computation of the eigenvalues, but the rule of using the average is breached. It is well known that the average is used when we measure a single parameter in a set. Examples include the average height of students in one class and the average mathematics grade in one class. However, in the E matrix, $l_{1,2}$ is the distance between the point (x_1, y_1) and the point (x_2, y_2), which are adjacent, $l_{1,3}$ is the distance between the point (x_1, y_1) and the point (x_3, y_3), between which there is one point, and $l_{1,N}$ is the distance between the point (x_1, y_1) and the point (x_N, y_N), between which there are N-2 points. Thus, $l_{1,2}$, $l_{1,3}$,…, $l_{1,N}$ are not the same type of measurement because there are different points among the distances, and using their average breaches the rule of averages. Therefore analyzing similarity/dissimilarity of DNA sequences by the average matrix element or the average row sum may lead to some errors.

Novel invariants

Consider, for example, the coordinates of the DNA sequence S = $s_1s_2s_3s_4s_5$ in a 2-D graphical representation are (x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), and (x_5, y_5). The distance is denoted as $l_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$, i, j = 1,2,…, 5. Then, the E matrix is as follows:

$$E = \begin{pmatrix} 0 & l_{1,2} & l_{1,3} & l_{1,4} & l_{1,5} \\ l_{2,1} & 0 & l_{2,3} & l_{2,4} & l_{2,5} \\ l_{3,1} & l_{3,2} & 0 & l_{3,4} & l_{3,5} \\ l_{4,1} & l_{4,2} & l_{4,3} & 0 & l_{4,5} \\ l_{5,1} & l_{5,2} & l_{5,3} & l_{5,4} & 0 \end{pmatrix}$$

Because the E matrix is symmetrical, the upper triangular matrix is considered. We classify $l_{i,j}$, i, j = 1,2,…, 5 according to the number of points between the point (x_1, y_1) and the point (x_j, y_j). Then, $l_{1,2}$, $l_{2,3}$, $l_{3,4,}$ and $l_{4,5}$ are sorted out as the first type because they are adjacent; $l_{1,3}$, $l_{2,4}$, and $l_{3,5}$ are sorted out as the second type because there is one point between each pair of points; $l_{1,4}$ and $l_{2,5}$ are sorted out as the third type because there are two points between each pair of points; and $l_{1,5}$ is sorted out as the fourth kind because there are three points between the pair of points. From the above E matrix, we can find that the four types of $l_{\mathrm{i,j}}$ are parallel to the leading diagonal, which is shown in the nether matrix:

$$E = \begin{pmatrix} 0 & l_{1,2} & l_{1,3} & l_{1,4} & l_{1,5} \\ l_{2,1} & 0 & l_{2,3} & l_{2,4} & l_{2,5} \\ l_{3,1} & l_{3,2} & 0 & l_{3,4} & l_{3,5} \\ l_{4,1} & l_{4,2} & l_{4,3} & 0 & l_{4,5} \\ l_{5,1} & l_{5,2} & l_{5,3} & l_{5,4} & 0 \end{pmatrix}.$$

In order to study the degree of the scatter of points in every type, we still consider the variance. We define the average a_i and variance d_i as follows:

The first type:

$$a_1 = \frac{1}{4}(l_{1,2} + l_{2,3} + l_{3,4} + l_{4,5}) \text{ and}$$

$$d_1 = \frac{1}{4}((l_{1,2} - a_1)^2 + (l_{2,3} - a_1)^2 + (l_{3,4} - a_1)^2 + (l_{4,5} - a_1)^2).$$

The second type:

$$a_2 = \frac{1}{3}(l_{1,3} + l_{2,4} + l_{3,5}) \text{ and}$$

$$d_2 = \frac{1}{3}((l_{1,3} - a_2)^2 + (l_{2,4} - a_2)^2 + (l_{3,5} - a_2)^2).$$

The third type:

$$a_3 = \frac{1}{2}(l_{1,4} + l_{2,5}) \text{ and}$$

$$d_3 = \frac{1}{2}((l_{1,4} - a_3)^2 + (l_{2,5} - a_3)^2).$$

The fourth type:

$$a_4 = l_{1,5} \text{ and}$$

$$d_4 = 0.$$

These averages and variances are assembled into a vector $v = (a_1, d_1, a_2, d_2, a_3, d_3, a_4, d_4)$, which is the novel invariant.

RESULTS

A graphical representation of DNA sequences gives us a simple way to numerically characterize the biological sequences. In this section, the graphical representation proposed by Yau et al. (2003) is used. We illustrate the use of this novel invariant with an examination of the similarities/dissimilarities among the coding sequences of the exon of β-globin genes of seven species: human, *Gallus*, opossum, lemur, mouse, rabbit, and rat. For simplicity, the first exon of β-globin genes of them are listed in Table 1.

Table 1. Coding sequences of the first exon of β-globin genes of 7 species.

Species	Coding sequences
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTG GGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTG GGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTG GTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTG GGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Mouse	ATGGTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTG GGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTG GGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTG GGGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAGG

For two DNA sequences S^i and S^j, two invariants $v^i = (a^i_1, d^i_1, a^i_2, d^i_2, \ldots, a^i_N, d^i_N)$ and $v^j = (a^j_1, d^j_1, a^j_2, d^j_2, \ldots, a^j_M, d^j_M)$ that correspond to S^i and S^j are obtained by this novel method. If $N > M$, v^j is amended to $v^j = (a^j_1, d^j_1, a^j_2, d^j_2, \ldots, a^j_M, d^j_M, 0, \ldots, 0)$, which has the same length as v^i. The same process is conducted for $N > M$. We define the distance by the following:

$$D_{ij}=\sqrt{\sum_{k=1}^{N}\left\{\left(a_k^i-a_k^j\right)^2+\left(d_k^i-d_k^j\right)^2\right\}}\ .$$

As indicated in Table 2, we found that the most similar species pair is (human, lemur), and similar species pairs include (rabbit, rat) and (lemur, mouse). On the other hand, the largest entries in the similarity/dissimilarity matrix appear in the rows belonging to opossum (the most remote species from the remaining mammals) and *Gallus* (the only non-mammalian representative).

Table 2. The similarity/dissimilarity matrix for the coding sequences of Table 1 based on the distance $D_{i,j}$.

Species	Human	*Gallus*	Opossum	Lemur	Mouse	Rabbit	Rat
Human	0	67.554	370.144	13.353	33.874	58.074	73.635
Gallus		0	417.388	73.139	80.758	113.313	126.915
Opossum			0	359.464	344.015	313.419	301.057
Lemur				0	22.841	47.893	61.920
Mouse					0	34.565	47.022
Rabbit						0	22.235
Rat							0

DISCUSSION

In order to demonstrate the advantage of this novel invariant, we compared it with the traditional five invariants. In Table 3, we listed the recently reported results of the degree of similarity/dissimilarity of the coding sequences of the exon of the β-globin gene of several species compared with that of the human β-globin gene by different methods. For an impartial comparison, all of these results were normalized to the human to *Gallus* ratio and were proposed by different researchers.

Table 3. Comparison of similarity/dissimilarity indexes for β-globin exon sequence differences between different species. All indexes are normalized to Human-*Gallus* ratio.

References	Normalized index from difference between human and					
	Gallus	Opossum	Lemur	Mouse	Rabbit	Rat
Table 4 (Nandy, 1994)	1.000	2.250	2.000	0.840	0.990	1.450
Table 15 (He and Wang, 2002)	1.000	1.340	1.061	0.763	0.646	1.232
Table 5 (Li and Wang, 2003)	1.000	0.768	0.583	0.000	0.000	0.806
Table 3 (Randić et al., 2003)	1.000	1.357	0.798	0.761	0.385	0.394
Table 7 (Liao and Wang, 2004c)	1.000	2.158	2.220	1.060	1.120	1.099
Table 6 (Chi and Ding, 2005)	1.000	4.518	3.330	0.834	0.609	0.586
Table 5 (Yao et al., 2005)	1.000	0.804	1.170	0.731	1.131	0.725
Table 4 (Qi and Qi, 2007)	1.000	1.078	0.666	0.162	0.258	0.591
Table 2 (Huang et al., 2009)	1.000	3.483	0.584	0.791	0.369	0.826

The average v^* = (1.0000, 1.9734, 1.3795, 0.6606, 0.6124, 0.8569) of these results was computed and was considered the ideal degree of similarity/dissimilarity of these species because so many results were presented by different scholars. The intervals between the average and results obtained by the traditional five methods and our novel method defined the errors of these approaches, which are shown in Table 4. It is noticeable that our novel method is superior to the others.

Table 4. Comparison of the errors of the six methods.

Methods	Errors
Leading eigenvalue	4.859
Max-min (eigenvalue)	5.992
Leading eigenvalue/N	4.859
Average matrix element	5.223
Average row sum	4.604
Our novel method	3.718

ACKNOWLEDGMENTS

The authors are grateful to the anonymous referees for many valuable suggestions that have improved the manuscript.

REFERENCES

Chi R and Ding KQ (2005). Novel 4D numerical representation of DNA sequences. *Chem. Phys. Lett.* 407: 63-67.

Guo Y and Wang TM (2008). A new method to analyze the similarity of the DNA sequences. *Comput. Theor. Chem.* 853: 62-76.

He PA, Li XF, Yang JL and Wang J (2011). A novel descriptor for protein similarity analysis. MATCH. *Commun. Math. Comput. Chem.* 65: 445-458.

He P and Wang J (2002). Numerical characterization of DNA primary sequence. *Internet Electron J. Mol. Des.* 1: 668-674.

Huang G, Liao B, Li Y and Yu Y (2009). Similarity studies of DNA sequences based on a new 2D graphical representation. *Biophys. Chem.* 143: 55-59.

Li C and Wang J (2003). Numerical characterization and similarity analysis of DNA sequences based on 2-D graphical representation of the characteristic sequences. *Comb. Chem. High Throughput Screen.* 6: 795-799.

Liao B (2005). A 2D graphical representation of DNA sequence. *Chem. Phys. Lett.* 401: 196-199.

Liao B and Wang TM (2004a). Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation. *Chem. Phys. Lett.* 388: 195-200.

Liao B and Wang TM (2004b). 3-D graphical representation of DNA sequences and their numerical characterization. *Comput. Theor. Chem.* 681:1-3: 209-212.

Liao B and Wang TM (2004c). New 2D graphical representation of DNA sequences. *J. Comput. Chem.* 25: 1364-1368.

Nandy A (1994). A new graphical representation and analysis of DNA sequence structure: I. Methodology and Application to Globin Genes. *Curr. Sci.* 66: 309-314.

Nandy A, Harle M and Basak SC (2006). Mathematical descriptors of DNA sequences: development and applications. *ARKIVOC* 2006: 211-238.

Qi ZH and Qi XQ (2007). PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization. *Chem. Phys. Lett.* 442: 434-440.

Randić M, Vracko M, Lers N and Plavsic D (2003). Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. *Chem. Phys. Lett.* 371: 202-207.

Randić M, Zupan J, Balaban AT, Vikic-Topic D, et al. (2011). Graphical representation of proteins. *Chem. Rev.* 111: 790-862.

Wu R, Hu Q, Li R and Yue G (2011). A novel composition coding method of DNA sequence and its application. *MATCH. Commun. Math. Comput. Chem.* 67: 269-276.

Yau SS, Wang J, Niknejad A, Lu C, et al. (2003). DNA sequence representation without degeneracy. *Nucleic Acids Res.* 31: 3078-3080.