Journal of Mathematical Chemistry, Vol. 43, No. 3, March 2008 (© 2007)
DOI: 10.1007/s10910-007-9267-1

On property of the invariant of graphical representations of DNA sequences

Chun-xin Yuan*, Li-wei Liu and Tian-ming Wang

Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, China
E-mail: chunxinyuan@yahoo.com.cn

Chun Li

Department of Mathematics, Bohai University, Jinzhou 121000, P. R. China

Received 18 November 2006; Revised 23 January 2007

In this article, we consider the influence of variation of DNA sequence on the leading eigenvalue of graphical representation of the biological sequences. The research interpret the rationality of the graphical representation method that compare different DNA sequences. And we show the result on two different models that presented before.

KEY WORDS: graphical representation, DNA sequences, similarity analysis, invariant, leading eigenvalue

AMS Classification: 15A18, 92E10

1. Introduction

The discovery and availability of new techniques has produced an incredible mass of data on biological sequences. This, in turn, produces new questions and makes it important to explore new ways to analyze these biological sequences.

In recent years, several researchers introduced an alternative way to compare the biological sequences, based on a set of invariants of biological sequences, rather than directly using sequence alignment. Graphical techniques have emerged as a powerful tool for the visualization and analysis of biological sequences. The researchers outlined or improved on different graphical representations of DNA sequences based on 2-D, 3-D or 4-D [1–10]. (In fact, 4-D representation is not graphical representation, but it follows the same procedure of characterization of DNA sequences). Similarly, graphical representations of RNA and proteins were also proposed in recent years [11–15]. The numerical characterization of biological sequences would render possible significant data reduction

*Corresponding author

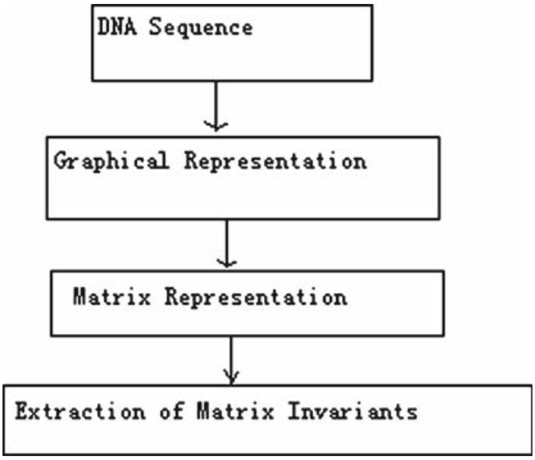

Figure 1. Underlying structure of the graphical methods for characterization of DNA sequence by invariants.

as well as a straightforward quantitative evaluation of the degree of similarity/dissimilarity between biological sequences.

In order to facilitate the comparison of different biological sequences, the researchers transformed the graphical representations of biological sequences into mathematical objects such as E matrix, D/D matrix, L/L matrix and their “high order” matrices, and they extract the invariants of matrices to numerically characterize the biological sequences. Then the invariants would be regarded as descriptors of the DNA primary sequence to facilitate comparison of different sequences. The basic procedure of these models is summarized in figure 1.

2. Methodology

Among sequence invariants, the leading eigenvalue of a matrix associated with a DNA sequence is an important invariant and is widely used for characterization of DNA sequences. The leading eigenvalue of the D/D matrix gives a measure of the degree of folding of long chains. The smaller the value of the leading eigenvalue, the more folded the corresponding graphical representation of DNA [16]. But the rationality of comparison of different DNA sequences with the leading eigenvalue is not fully explained yet. In this paper, we will study the influence of variation of DNA sequence on the leading eigenvalue of graphical representation of the sequence. We will consider this problem in 2-D and 3-D representation of DNA sequence. And this method is applicable to many representation models of biological sequence.

Table 1
DNA primary sequence for the first exon of Chimpanzee β-globin.

ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGA ACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG

2.1. The 3-D graphical representation of DNA sequences

The three-dimensional graphical representation of DNA sequences provides a visual inspection of DNA data. Several researchers have proposed different graphical representations of DNA sequence.

Our study uses the 3-D graphical representation of DNA sequence outlined in [7]. We present it briefly as follows. We assign A (adenine), G (guanine), T (thymine), and C (cytosine) to $-x, +x, -y$, and $+y$, respectively, while the corresponding curve extend along with z-axes. In detail, let $B = b_1b_2b_3 \ldots b_n$ be an arbitrary DNA sequence. Then we have a map Φ_1, which maps B into a plot set. Explicitly

$$\Phi_1(B) = \Phi_1(b_1)\Phi_1(b_2)\Phi_1(b_3)\ldots\Phi_1(b_n) \text{ where } \Phi_1(b_i) = \begin{cases} (-1, 0, i) \text{ if } b_i = A \\ (1, 0, i) \quad \text{if } b_i = G \\ (0, -1, i) \text{ if } b_i = T \\ (0, 1, i) \quad \text{if } b_i = C \end{cases}$$

Connecting adjacent points, we obtain a 3-D curve.

According to the method outlined in reference [7], the leading eigenvalue for the coding sequence of the first exon of Chimpanzee β-globin gene, which is shown in table 1, is 65.1398.

We select k bases randomly in the primary DNA sequence, and change them into an arbitrary base of the set {A, T, G, C}. We call the sequence a k-random sequence ($1 \leqslant k \leqslant n$). That is, in a k-random sequence, k bases are changed randomly according to the primary sequence. With the same method mentioned above, we could calculate the leading eigenvalue of the k-random sequence.

We show the average leading eigenvalue of some k-random sequences in figure 2(a) and table 2. And we define Skew-Variance, (S-Var for short) $Var_l(\xi) = E(\xi - lp)^2$, where lp is the leading eigenvalue of the primary DNA sequence, and ξ denote the leading eigenvalue of an random sequence. This definition follows the definition of variance, it indicates the deviation of ξ to lp. And these values are shown in figure 2(b) and table 2.

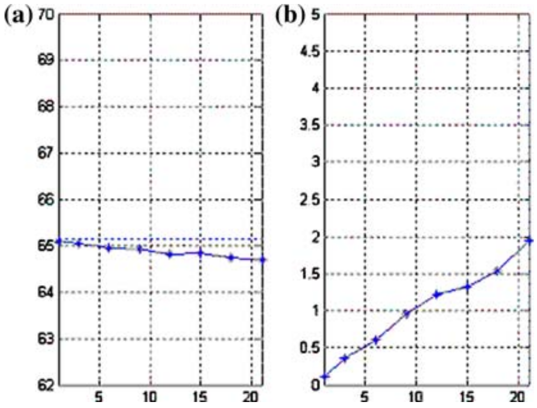

Figure 2. The average leading eigenvalue and S-Var based on the 3-D graphical representation.

Table 2
The average leading eigenvalue and S-Var based on the 3-D graphical representation.

k	1	3	6	9	12	15	18	21
Average leading eigenvalue	65.1014	65.0495	64.9635	64.9309	64.8195	64.8442	64.7449	64.703
$Var_l(\xi)$	0.109091	0.363349	0.611234	0.957657	1.21684	1.32068	1.53572	1.95151

2.2. The 2-D graphical representation of DNA sequences

In reference [5, 6], a 2-D graphical representation was outlined to analyze similarity/dissimilarity of DNA sequences. Assign A (adenine), G (guanine), T (thymine), and C (cytosine) to the four horizontal lines separated by unit distance. The consecutive bases are represented by dots and are placed along the horizontal axes at unit distance intervals. Then connect adjacent dots with lines, which results in a zigzag like curve of a definite geometrical form.

Based on this 2-D graphical representation, the leading eigenvalue for the coding sequence of the first exon of Chimpanzee β-globin gene, which is shown in table 1, is 66.62 in the case of labeling ATGC.

As what we have done in Section 2.1, we could calculate the average leading eigenvalue and S-Var of k-random sequences based on the 2-D graphical representation of DNA sequence. The detailed result is shown in figure 3 and table 3.

3. Discussion

From above result, we could find that the value of S-Var of k-random sequence, which indicates the deviation of the leading eigenvalue of k-random sequence from the leading eigenvalue of the primary DNA sequence, is of

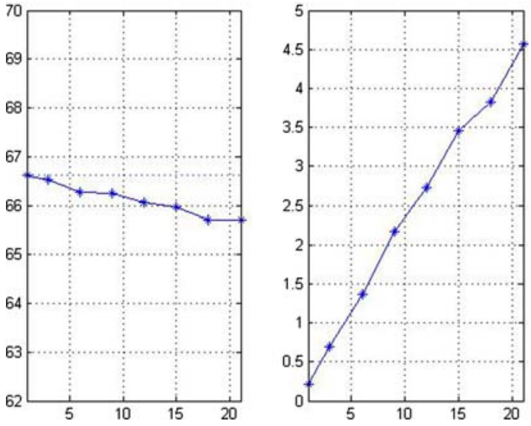

Figure 3. The average leading eigenvalue and S-Var based on the 2-D graphical representation.

Table 3
The average leading eigenvalue and S-Var based on the 2-D graphical representation.

k	1	3	6	9	12	15	18	21
Average leading eigenvalue	66.6086	66.533	66.2601	66.2512	66.0545	65.9648	65.7148	65.7099
$Var_l(\xi)$	0.214357	0.700608	1.36013	2.17614	2.71948	3.45377	3.82847	4.56795

augmentation with the increase of k in both the 2-D and the 3-D graphical representations. A k-random sequence results from a stochastic substitution of k bases in the primary sequence. So the larger is the value of k, the more dissimilarity is of the two sequences. That is to say, the more dissimilar are two sequences, the more deviant are the leading eigenvalues corresponding to the two sequences. This is where in lies the rationality of the invariant approach that compares different DNA sequences. Since the randomicity of the selection, the conclusion is statistically reliable. In figures 4 and 5, we present the equations of linear regression of the 2-D and 3-D graphical representation models, respectively.

We could also notice that the value of average leading eigenvalue of k-random sequence is degressive with the increasing of k in the both models. (Though there is an exception, for example, when k = 15 in table 2, the trend is degressive.) And all average leading eigenvalues of k-random sequence are less than the leading eigenvalue of the primary DNA sequence. The reason to this phenomenon might be that with the increase of k, the sequence is more and more stochastic, and the corresponding graphical representation of DNA sequence becomes more folded. Since the leading eigenvalue of the matrix gives a measure of the degree of folding of long chains, it is naturally present a trend of degression with the increase of k.

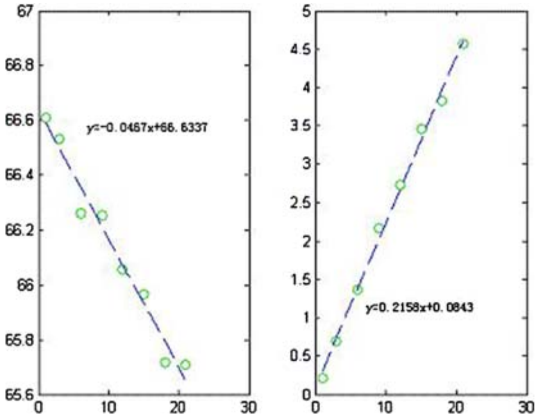

Figure 4. Equation of linear regression based on the 2-D graphical representation.

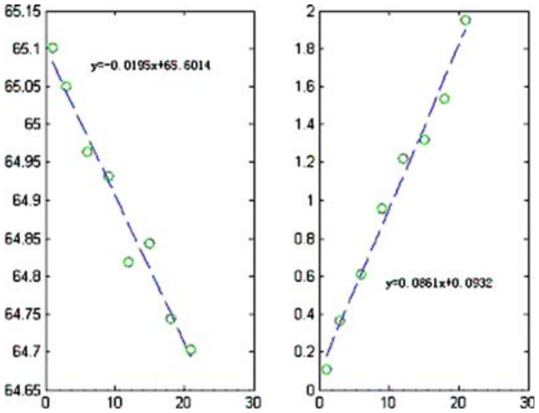

Figure 5. Equation of linear regression based on the 3-D graphical representation.

4. Conclusion

The graphical representation method has provided us a powerful tool for characterization and comparison of DNA sequences. Researchers have outlined many different models of graphical representation of biological sequence. However, the basic problems about this approach are still need our attention. In this paper, we consider the influence of variation of DNA sequence on the leading eigenvalue of graphical representation of the sequence. Based on the study on a 2-D and 3-D graphical representation models, the present work indicate the rationality of the graphical representation method that compare different DNA sequences.

Acknowledgement

This work was supported in part by National Natural Science Foundation of China (10571019).

References

[1] M. Randić, X.F. Guo and S.C. Basak, J. Chem. Inf. Comput. Sci. 41 (2001) 619.
[2] X.F. Guo, M. Randić and S.C. Bassk, Chem. Phys. Lett. 350 (2001) 106.
[3] M. Randić and A.T. Balaban, J. Chem. Inf. Comput. Sci. 43 (2003) 532.
[4] Y. Liu, X.F. Guo, J. Xu, L. Pan and S. Wang, J. Chem. Inf. Comput. Sci. 42 (2002) 529.
[5] M. Randić, M. Vračko, N. Lerš and D. Plavšić, Chem. Phys. Lett. 368 (2003) 1.
[6] M. Randić, M. Vračko, N. Lerš and D. Plavšić, Chem. Phys. Lett. 371 (2003) 202.
[7] C. Yuan, B. Liao and T.M. Wang, Chem. Phys. Lett. 379 (2003) 412.
[8] B. Liao and T.M. Wang, Chem. Phys. Lett. 388 (2004) 195.
[9] C. Li and J. Wang, Comb. Chem. High T. Scr. 6 (2003) 795.
[10] C. Li and J. Wang, Comb. Chem. High T. Scr. 7 (2004) 23.
[11] B. Liao and T.M. Wang, J. Biomol. Str. Dyn. 21 (2004) 827.
[12] Y.H. Yao, B. Liao and T.M. Wang, J. Mol. Struct.: Theochem. 755 (2005) 131.
[13] Y.H. Yao, X.Y. Nan and T.M. Wang, J. Comput. Chem. 26 (2005) 1339.
[14] W. Zhu, B. Liao and K.Q. Ding, J. Mol. Struct.: Theochem. 757 (2005) 193.
[15] F.L. Bai and T.M. Wang, J. Biomol. Str. Dyn. 23 (2006) 537.
[16] M. Randić and M. Vračko, J. Chem. Inf. Comput. Sci. 40 (2000) 599.