MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **67** (2012) 809-816

ISSN 0340 - 6253

3D-Dynamic Representation of DNA Sequences

Vahid Aram and Ali Iranmanesh*

Department of Mathematics, Faculty of Mathematical Sciences,
Tarbiat Modares University, P.O. Box : 14115-137,
Tehran, Iran

iranmanesh@modares.ac.ir

(Received May 16, 2011)

Abstract

In this paper, we introduce a new graphical representation of the DNA sequences, which we call 3D- Presentation graphs. The present method is based on the 2D dynamic representation developed in [1] and Hook Immanantal for polynomial of Laplacian matrix on DNA graph that have been used before.

1.Introduction

DNA (deoxyribonucleic acid) is a double stranded sequence of four nucleotides; the four nucleotides that compose a strand of DNA are as follows: adenine (A), guanine (G), cytosine (C), and thymine (T); they are often called bases. The chemical structure of DNA (the famous double- helix) was discovered by James Watson and Francis Crick in 1953[1]. It consists of a particular bond of two linear sequences of bases. This bond follows a property of complementarity: adenine bonds with thymine (A-T) and vice versa (T-A), cytosine bonds with guanine (C-G) and vice versa (G-C). This is known as Watson-Crick complementarity.

Each DNA strand has two different ends that determine its polarity: the 3.'end, and the 5' end. The double helix is an anti-parallel (two strands of opposite polarity) bonding of two complementary strands. The idea (due to Leonard Adleman) is to use strands of DNA to

*- Corresponding author (Ali Iranmanesh)

encode the (instance of the) problem and to manipulate them using techniques commonly available in any molecular biology laboratory, to simulate operations that select the solution of the problem, if it exists[1,2].

In the recent years, a rapid growth of sequence data in DNA databases has been observed. Some graphical representations of DNA sequences have been given by Nandy [3], and Guo et al. based on 2D graphical representation of DNA sequences. Guo and Nandy introduced a novel 2D graphical representation of DNA sequences of low degeneracy [4].

Consequently, designing mathematical tools that aim at a quick identification or for similarity studies between sequences have become an urgent necessity. Very useful are the methods based on graphical representations. In general, when the number of data is large, it can be easier to deal with mathematical descriptors that offer a numerical characterization of the graphs .

In general, many advances in 2D, 3D, and 4D DNA sequences representation appeared after the initial works [5-15] . Up to now , many papers published in DNA sequence. For example see [16-25].

Since our method is based on the 2D-dynamic graph which introduced in [1] , in this reference the masses of the degeneracy of the plots is important , but in our method, we need only the graph of DNA sequence and therefore we remove the mass of points .

In this paper, we introduce a new graphical representation of the DNA sequences, which we call 3D-presentation and we characterize a DNA graph by the second immanantal polynomial and give an example for HSHISAD sequence .

2. Theory

Let G= (V, E) be a graph with the set vertex of V = $\{v_1 , v_2, \dots , v_n\}$ and the set of edges of E . The adjacency matrix A(G) = (a_{ij}) of G is the n by n matrix defined by a_{ij}=1 , if (v_i , v_j) ϵ E and a_{ij}=0 otherwise . Of course, if G = (V, E) is a directed graph, then (v_i , v_j) is pair order .

If D(G) is a diagonal matrix of degree vertices of the graphs, then L(G) = D(G) – A(G) is a Laplacian matrix. The second immanant of an n by n matrix L(G) = (l_{ij}) is defined by

$$d_2\big(L(G)\big) = \textstyle\sum_{\sigma \in S_n} \chi_2(\sigma) \prod_{t=1}^{n} l_{t\sigma(t)} \text{ [26-27]}$$

where χ_2 is the irreducible character of S_n corresponding to the partition ($2,1^{n-2}$) . In particular, $\chi_2(\sigma) = (\sigma)[F(\sigma) - 1]$, where ε is the alternating character and F is the number of fixed points . Define the "d_2-polynomial" of G by

$$d_2\big(xI - L(G)\big) = \textstyle\sum_{k=0}^{n}(-1)^k c_k(G)x^{n-k} \quad [28].$$

The d_2 polynomial associated with the Laplacian matrix L(G) is:

$$d_2\big(xI - L(G)\big) = \sum_{k=0}^{n}(-1)^k\, c_k(G)x^{n-k}$$

The coefficients $c_0, \ldots, c_n$ can be computed as follows:[28]

$c_0(G) = n\text{-}1$

$c_1(G) = 2m\,(n\text{-}1)$

$c_k(G) = \sum_{X\in Q_{k,n}}(\sum_{i=1}^{n} l_{ii}\,\det(L(G)\{X\}(i) - \det(L(G)\{X\}))$.

$Q_{k,n}$ denotes the collection of $\binom{n}{k}$ k-element subsets of the set {1, 2, . . . n}. If we denote with L[X] the k × k principal sub matrix of M corresponding to X, where X ∈ $Q_{k,n}$, we can define the m × m matrix

$$L\{X\} = \begin{bmatrix} L[X] & 0_k \\ 0_k & I_{n-k} \end{bmatrix}$$

where I_{n-k} is the identity matrix of size n − k, 0_k is the null matrix of size k and L{X}(i) is the matrix obtained from L{X} by removing the n-th row and the n-th column.

In follow we show how it is possible to characterize a DNA graph by the second immanantal polynomial and how to embed the polynomial coefficients into a low dimensional vector space.

3. Result and discussion

Let us consider the sequence CGTCGA. The method for constructing the graph is shown in figure. 1. The plot of a sequence starts from the origin of the coordinate system (0,0) denoted as ‘start’ in the figure. Then, using the convention of a walk in 2D-space outlined in, the point is shifted, in succession, by one unit for each base in the sequence. The shifts are made by the following unit vectors: A = (-1,0), G = (1,0), C = (0,1) and T = (0,-1). The vectors C, T and A, G lie on the same lines. However, for pedagogical reasons, they are drawn next to each other in this paper.

As we explained in the introduction, we deal with the graph of DNA sequence which contains vertices without mass.

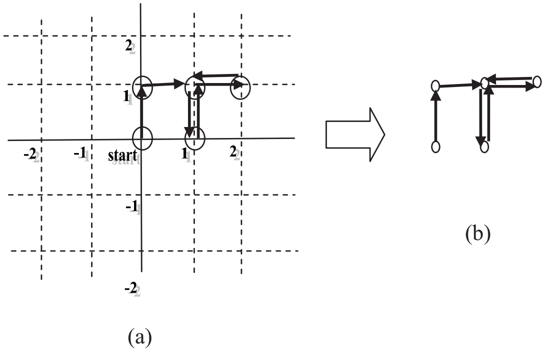

Figure 1. (a) The method for constructing the 2D-dynamic graphical representation of the CGTCGA sequence

(b) The graph of CGTCGA sequences according to figure (a)

At this point, we are able to partition the set of representative graphs into equivalence classes with the equivalence relation provided by sharing the same second immanantal polynomial, that is to say the same characteristic vector c = $(c_0, c_1, \ldots, c_n)$.

In this section, we are going to introduce a technique to get a better clustering of the characteristic set in order to provide a more reliable metric for the representative graphs, specifically designed for the graphs actually present in the database. In fact, our objective is to provide a metric that allows to make queries on a large collection of representative graphs, or better of their characteristic vectors c.

We used the first three most significant independent components e_1, e_2, e_3 to represent the characteristic vectors extracted from them. The coordinate system is spanned by the three independent components e_1, e_2, e_3 and the characteristic vectors c_j can be projected onto this pattern space by $x_{ij}= e_i \cdot c_j$, where i = 1, 2, 3, j = 1, 2, . . . ,N and N is the dimension of the database.

It follows immediately that the metric will be the standard one between vectors in this linear space (figure. 2).

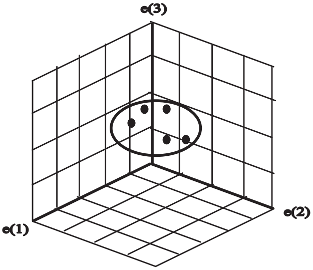

Figure 2. 3D-dynamic graphical representation of the CGTCGA sequence

We presented a DNA graph characterization through the second immanantal polynomial, which provides an invariant set of coefficients which completely characterizes a DNA graph. Then it is possible to embed the set of polynomial coefficients into three dimensional space where it is possible to give a graph metric tuned on the specific collection by means of the standard product. This space is spanned by the independent vectors provided by the Independent Component Analysis performed on all vectors belonging to the database

In the above part, we gave 3D- dynamic representation for a model CGTCGA sequence, now we give another example for obtaining 3D- dynamic representation. For this purpose, we used from table 2 in [1]. In this table, there is a DNA sequence with the name (HSHISAD). Using in [1], we have the following figure :

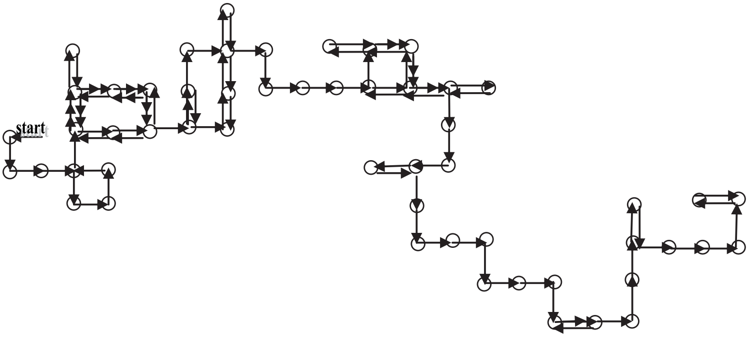

Figure 3. 2D – representation of human
ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAG
GTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG

The same as the previous model, we can show 3D - representation of HSHISAD sequence in the following figure.

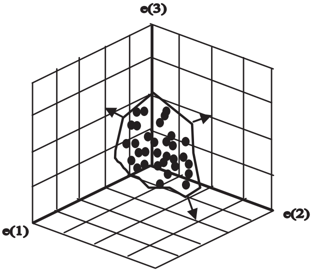

Figure 4. A part of 3D – representation of human ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAG GTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG.

Acknowledgement.

The authors would like to thank the referees for the valuable comments.

References

[1] D. B. Waz, T. Clark, P. Waz, W. Nowak, A. Nandy, 2D-dynamic representation of DNA sequences*, Chem. Phys. Lett.* **442** (2007) 140–144.

[2] L. M. Adleman, Computing with DNA, *Sci. Amer.* (1998) (8) 54–61.

[3] A. Nandy, On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation models, *Chem. Phys. Lett.* **368** (2003) 102–107.

[4] X. Guo, X. Nandy, Numerical characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy, *Chem. Phys. Lett.* **369** (2003) 361– 366.

[5] B. Liao, T. M. Wang, New 2D graphical representation of DNA sequences, *J. Comput. Chem.* **25** (2004) 1364–1368.

[6] B. Liao, W. Zhu, Y. Liu, 3D Graphical representation of DNA sequence without degeneracy and its applications in constructing phylogenic tree, *MATCH Commun. Math. Comput. Chem.* ***56*** (2006) 209–216.

[7] B. Liao, A 2D graphical representation of DNA sequence, *Chem. Phys. Lett.* **401** (2005) 196–199.

[8] M. Randić, M. Vračko, A. Nandy, S. C. Basak, On 3D-graphical representation of DNA primary sequences and their numerical characterization. *J. Chem. Inf. Comput. Sci.* **40** (2000) 1235–1244.

[9] J. F. Yu, J. H. Wang, X. Sun, Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation, *MATCH Commun. Math. Comput. Chem*. **63** (2010) 493–512.

[10] M. Randić, A. T. Balaban, On a four-dimensional representation of DNA primary sequences, *J. Chem. Inf. Comput. Sci*. **43** (2003) 532–539.

[11] E. Hamori, J. Ruskin, H-curves, a novel method of representation of nucleotide series especially suited for long DNA sequences*, J. Biol. Chem*. **258** (1983) 1318–1327.

[12] M. A. Gates, A simple way to look at DNA, *J. Theor. Biol*. **119** (1986) 319–328.

[13] A. Nandy, A new graphical representation and analysis of DNA sequence structure I. Methodology and application to globin genes, *Curr. Sci.* **66** (1994) 309–313.

[14] P. M. Leong, S. Morgenthaler, Random walk and gap plots of DNA sequences, *Comput. Appl. Biosci*. **11** (1995) 503–507.

[15] A. Nandy, Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons, *Curr. Sci*. **70** (1996) 661–668.

[16] R. Wu, Q. Hu, R. Li, G. Yue, A novel composition coding method of DNA sequence and its application, *MATCH Commun. Math. Comput. Chem*. **67** (2012) 269–276.

[17] X. Zhou, K. Li, M. Goodman, A. Sallam, A novel approach for the classical Ramsey number problem on DNA-based supercomputing, *MATCH Commun. Math. Comput. Chem*. **66** (2011) 347–370.

[18] Q. Zhang, B. Wang, On the bounds of DNA coding with H-distance, *MATCH Commun. Math. Comput. Chem*. **66** (2011) 371–380.

[19] Wang, T. Wang, Conditional LZ complexity and its application in mtDNA sequence analysis, *MATCH Commun. Math. Comput. Chem*. **66** (2011) 425–443.

[20] Q. Zhang, B. Wang, X. Wei, Evaluating the different combinatorial constraints in DNA computing based on minimum free energy, *MATCH Commun. Math. Comput. Chem*. **65** (2011) 291–308.

[21] Y. Zhang, W. Chen, A new measure for similarity searching in DNA sequences, *MATCH Commun. Math. Comput. Chem*. **65** (2011) 477–488.

[22] R. Wu, R. Li, B. Liao, G. Yue, A novel method for visualizing and analyzing DNA sequences , *MATCH Commun. Math. Comput. Chem*. **63** (2010) 679–690.

[23] W. Chen, B. Liao, Y. Liu, W. Zhu, Z. Su, A numerical representation of DNA sequences and its applications, *MATCH Commun. Math. Comput. Chem*. **60** (2008) 291–300.

[24] Pesek, J. Zerovnik, A numerical characterization of modified Hamori curve representation of DNA sequences, *MATCH Commun. Math. Comput. Chem*. **60** (2008) 301–312.

[25] Y. Zhang, W. Chen, New invariant of DNA sequences*, MATCH Commun. Math. Comput. Chem*. **58** (2007) 197–208.

[26] R. Merris , Immanantal invariants of graphs , *Lin. Algebra Appl.* **401** (2005) 67–75.

[27] C. M. Da Fonseca, The μ-permanent of a tridiagonal matrix, orthogonal polynomials, and chain sequences, *CMUC - Centro de Matem´atica da Universidade de* Coimbra, 2009.

[28] R. Merris, The second immanantal polynomial and the centroid of a graph*, Siam J . Alg. Discr . Math*. **7** (1986) 484–503.