DNA Sequence Representation methods

G. Santhosh Kumar
Member, ACM
Department of Computer Science
Cochin University of Science &
Technology, Cochin-22, Kerala, India
+914842862306
sancochin@acm.org

Shiji S H
Department of Computer Science
Cochin University of Science &
Technology,
Cochin-22, Kerala, India
+914842862316
shijish@gmail.com

ABSTRACT

DNA sequence representation methods are used to denote a gene structure effectively and help in similarities/dissimilarities analysis of coding sequences. Many different kinds of representations have been proposed in the literature. They can be broadly classified into Numerical, Graphical, Geometrical and Hybrid representation methods. DNA structure and function analysis are made easy with graphical and geometrical representation methods since it gives visual representation of a DNA structure. In numerical method, numerical values are assigned to a sequence and digital signal processing methods are used to analyze the sequence. Hybrid approaches are also reported in the literature to analyze DNA sequences. This paper reviews the latest developments in DNA Sequence representation methods. We also present a taxonomy of various methods. A comparison of these methods where ever possible is also done.

Categories and Subject Descriptors

A.1 [**Introductory And Survey**]**:** DNA Sequence Representation

General Terms

Documentation

Keywords

DNA sequence representation, Fractal method, DNA sequence analysis, Graphical representation, Numeric representation, and Geometric representation.

1. INTRODUCTION

DNA, deoxyribonucleic acid stores all the genetic information in a species. It is composed of long array of nucleotides which contains bases adenine (A), guanine (G), cytosine (C) and thymine (T). DNA sequence analysis helps to determine the patterns in these sequences, distinguish coding from noncoding sequences and find problems related to the classification and evolution of organisms. For rapid viewing and analysis of the data some representation methods are necessary. The graphical and geometrical approaches provide visual representation of DNA sequences, which play a good role in the research of DNA structure and its function. Graphical representation of DNA sequence provides a simple way of observing, sorting, analyzing and comparing different gene structures quickly. These techniques provide information about variations and repetition of the nucleotides along a sequence which are not as easily obtained by other methods. Various studies on DNA representations has started since 1980's. DNA sequences are stored in the database system in the form of character streams. It is difficult to observe and distinguish the differences and similarities among long DNA sequences. In order to observe, analyze and compare effective methods of representations are needed. Different methods in literature are categorized and a comparison study of these methods are done. Taxonomy of DNA sequence representation methods are given in Figure 1 and their year wise categorization is given in Table 1.

Table 1. Year wise Categorization of DNA sequence representation methods

Year	Representation methods
1983	Hamori's H Curve
1985	Gates 2D method
1990	Jeffry's CGR method
1992	Voss's numerical method, Dutta's CGR
1994	Nandy's 2D
1998	Buldrev's method using statistical physics, Lantin's H curve, Roy, Nandy's 2D method
1999	Rifaat's Fractal method
2000	Zu Guo Yu's fractal method, Ashlock's IFS
2002	Zu Guo's Fractal method
2003	Ashlock's Chaos Automata, Yau's 2D,Chang's 3D, Zhang's Z curve
2006	Joseph, Sasikumar's CGR, Wang's 2D, Ji, li's TB curve, Liao' Chakravarthy's AR model, Zhou's AR model, Randic's 1D, Randic's numeric method
2007	Ashlock's Fractal method , Akhtarl's numeric, Luo's 4D method
2008	Yu's CGR, Kang Dai's 2D, Randic's numeric, Cuttani's numeric, Rambally's Hybrid approach
2009	Wang's 3D, Jie's CGR

ISB'10, February 15-17, 2010, Calicut, India.

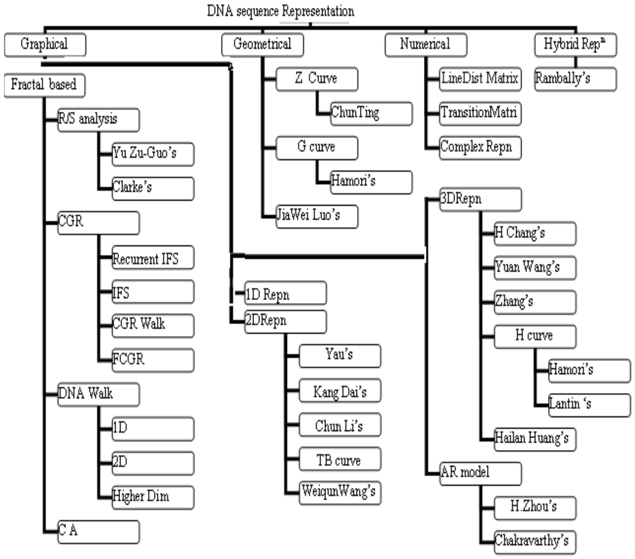

Figure 1 Taxonomy of DNA sequence representation

2. GRAPHICAL REPRESENTATION

2.1 Fractal representation methods

A fractal is generally a rough or fragmented geometric shape that can be split into parts, each of which is a reduced-size copy of the whole.. Fractal analysis is used to obtain the long range and short range correlations in the DNA sequences. Zu-Guo [23] reports a time series model of DNA sequences.

2.1.1 R/S (Rescaled range) analysis

A DNA sequence is a sequence over the alphabet {A, C, G, T}. These four bases are mapped to four distinct values. One can use {−2, −1, 1, 2} to replace {A, C, G, T} or other orders of A, G, C, T but distinguish A and G from purine, C and T from pyrimidine. Thus a number sequence obtained and it can be treated as a fractal records in time [25].

2.1.2 Chaos game representation (CGR)

The chaos game is described by an iterated function system (IFS) [5, 6]. Chaos game for DNA sequence is play according to the given DNA sequence [10]. Each corner of a square is written T at (0,0), A at (0,1), C at (1,1), and G at (1,0), and the initial point I_0 drawn at (0.5, 0,5). For example, for a sequence ATG.., the first game point I_1 is obtained by moving I_0(0.5, 0.5) to the midpoint between I_0 and A(0,1). Likewise, I_2 is drawn and so on. Jeffrey's method [13] permits the investigation of patterns in sequences, visually revealing previously unknown structures. In CGR-walk model [7], the distribution of positions in CGR walk is unique and the distance between positions may serve as a measure of similarity between the corresponding sequences. Randic put forward a method by knowing the Cartesian coordinates of the point representing the last base of a sequence. RIFS model can be used to simulate the measure representation of complete genomes but IFS model is only for protein sequences [26]. RIFS describes the scale invariance of a measure. The Frequency CGR (FCGR) shows the frequencies of oligonucleotides from a color scheme normalized to the distribution of frequency of occurrence of patterns. The FCGR is divided into regions of sub-pattern.

2.1.3 DNA walk representation:

For long-range correlations of a DNA sequence, a fractal landscape or DNA walk method uses the techniques of statistical physics [21]. In 1D random walk model, a walker moves either "up" +1 or "down" −1 one unit length for each step of the walk. 2D walk distinguish C from T in pyrimidines and A from G in purines. + X for A, - X for T, +Y for C and -Y for G. In higher dimensional walk the global fractal dimension of human DNA sequences treated as pseudo random walk. The global fractal dimension of coding sequences is different from that of non coding sequences and the sequences used are not random.

2.1.4 Chaos automata (CA)

A chaos automata is a finite state machine with a contraction map associated with each state [2]. Chaos automata can generate the fractal image of DNA sequences with iterated function system. They retain internal state information. A chaos automata is a 5-tuple (*S;C; A; t; i*) where *S* is a set of states, *C* is a set of contraction maps, *A* is an input alphabet, t is a transition function that maps a state and i is the input to a next state.

2.1.5 Comparison of fractal methods

Chaos game fractals or IFSs give useful visual displays of large, possibly complex data sets. 1D walk is highly compact, easy to implement and there is no loss of information. FCGR method has high computational efficiency and scale independence. Fractal methods take short running time, high accuracy, and can achieve as high as four orders of magnitude speedup compared to other methods, and is used to obtain the long and short range correlations in the DNA sequences.

2.2 1D and 2D representations

2.2.1 One dimensional representation

The widely used representation of a DNA sequence is the letter series representation. One dimensional method color code to represent and print DNA sequences as long as 100 kb in length on a single page. The representation uses color scheme to highlight local aggregate properties of large DNA segments.

2.2.2 Two dimensional graphical representations

2D graphical representation of DNA sequence is called the graph of the DNA sequence corresponding to ACGT axis system [4]. These methods choose the four cardinal directions in (x, y) coordinates to represent four bases. It was pointed out by Nandy [1] that there are three possible independent axes system to plot a 2D graph of DNA sequence. Yau et al. [21] demonstrated x and y-projection of any point on the graphical representation. A pyrimidine- purine graph was constructed on two quadrants of the cartesian coordinate system, with pyrimidines in the first quadrant and purines in the fourth quadrant. In Weiqun Wang's method [14] the number of A, T, G, and C from the starting point could be calculated based on an iterative comparison. Ji Hu give a 2-D ladder like graphical representation for the characteristic sequences of a DNA sequence, and then construct a 3-component vector. In Kang Dai's representation [15] DNA sequence is coded by assigning pyrimidine–purine to four real number pairs. Numerical coordinate was normalized so that all the values will be fallen into [0, 1]. TB curve [18] shows at most three essentially different curves representing the same DNA sequence. This curve displays two kinds of bases, the purine and pyrimidine, at a time on a plane, so called RY-TB (two kinds of bases) curve.

2.2.3 Comparison of 1D/2D representation

1D method is simple. Yau's and Wang's 2D methods provide a direct plotting method to denote DNA sequences without degeneracy. Kang Dai's method provides a straightforward comparison. TB curve allows numerical characterization. These are complex and require high ended computer graphics for visualization and are not yet quantitative enough to precisely characterize gene sequences. Most 2D methods are found to have loss of information.

2.3 3D and auto regressive representations

2.3.1 Three dimensional representations

Hsuan T. Chang et al [11] visualized DNA sequences by the use of 3-D trajectories (TDT). In the method, each of four nucleotides is assigned by three coordinates in the 3-D space. The directional distance can be used to represent two base pairs in the DNA sequence. TDT can be plotted using the vectors between two consecutive nucleotides. Yuan Wang et al [24] described a method based on the principle of symbolic dynamics, which maps DNA sequence into 3-D chaotic sequences of saw tooth function. Combined with the period-3 property of protein coding regions, and method based on extended Kalman filters (a minimum mean-square estimator) and autoregressive model is proposed to solve the problem of gene prediction. Zhang [22] assign the following vectors to the four nucleic bases. $(\mathrm{v}\sqrt{\mathrm{n}},0,\mathrm{u})\rightarrow\mathrm{A}$, $(0,\mathrm{v},\mathrm{w}\sqrt{\mathrm{n}}),\rightarrow\mathrm{G}$, $(w,u\sqrt{n},0)\rightarrow C$, $(u,w,v)\rightarrow T$ u, v and w are integer numbers and n is not perfect square. The curve connecting all plots of the characteristic plot set in turn is called 3D curve of DNA. In H curve, one unit along one of four directions representing four bases in xy-plane and one for each unit in the z-direction [16]. H-curves used to display entire genomes and to detect features in sequences such as a change in the DNA template-strand transcribed and overlapping genes.

2.3.2 Autoregressive (AR) model

The autoregressive model analyzes the spectrum of the DNA sequences [19, 10]. The AR modeling of a DNA sequence is done by mapping the sequence into the numerical domain and then calculate the AR parameters of the resulting numerical sequence. The AR modeling of sequences can be performed using linear prediction techniques and is robust to deletions, and insertions.

2.3.3 Comparison of 3D and AR methods

3D Trajectories can easily distinguish different DNA sequences directly. Multiple DNA sequences can be compared. This has low computational complexity and is easy to find the differences and the similarities among large sequences. 3D methods resolve degeneracy. AR model helps to distinguish coding and noncoding sequences and analyzes the spectral characteristics of repeating pattern of DNA sequences. It provides higher resolution for spectral estimation.

3. GEOMETRICAL REPRESENTATION

3.1 Z curve representation

Z curve [4] is one of the tool available for visualizing genomes. The three components of the Z-curve x y z represent three independent nucleotide distributions that completely describe the DNA sequence. Each of the three Z-curves generated a digital signal that has a clear biological interpretation. The components *x* y z display the distributions of purine versus pyrimidines, amino versus keto, and strong H-bond versus weak H bond bases along the sequence, respectively. One of the advantages of the Z curve is its originality. Luo, et al's Geometrical Representation [12] is a 4D representation of DNA sequence, which contains all the information in the DNA sequence and avoids overlapping.

3.2 G curve representation

The G-curves are generated in a virtual 5-D space whose orthogonal co-ordinates are assigned to the four nucleotides and to an integer characterizing the position of a nucleotide on a DNA chain. G-curves are useful only conceptually and not as a means of visual representation.

3.3 Comparison of geometrical methods

This method is useful in similarity studies. It has rich spatial folding structures that reflect the symmetry, periodicity, and the global features of the distribution of bases. Z curve method generally gives the observer a more intuitive impression. It is robust, independent and less redundant.

4. NUMERICAL REPRESENTATION

The numerical representation of a DNA sequence is given as a sequence of real numbers derived from a unique graphical representation of the standard genetic code. This representation is suitable for the quantitative analysis of the sequences.

4.1 LD matrix

LD matrix is the distance matrix for the points of the line defining line segments [17]. The list of coordinates along the line: x = 0, 1, 2, 6, 9, 10, 11, 12, 15 and 19 represent the first row of the LD matrix. The second row is obtained by subtracting the entry above the zero on the diagonal that is entry 1, from the corresponding elements of the first row, the main diagonal starting with zero: 0, 1, 5, 8, 9, 10, 11, 14 and 18. The third row of the LD matrix is obtained by subtracting again 1, the entry above the diagonal zero in the third row etc. Matrix elements adjacent to the main diagonal represent the length of the line segments making the line.

4.2 Transition matrix

Transition matrix is used for transitions from one kind of base to another [3]. For a given DNA sequence 's' it can construct a 4×4 matrix A = (tij), where tij means the number of times a given base being succeeded by another in the sequence. A is called the transition frequency matrix of s. We can construct a matrix P = (Pij) by dividing each element by the total of all entries in A. Such a matrix represents the relative frequency of all the possible types of transitions, and is called the transition proportion matrix of s. Voss's method is the earliest mapping of DNA to binary representation [20], which represents DNA with four binary indicator sequences showing the presence '1' and absence '0' of the respective nucleotides at locations 'n'.

4.3 Complex representation

The complex representation is based on the assumption that coefficients of the four 3-D tetrahedron vectors representing each DNA letter are either +1 or -1. The dimensionality of the resultant bipolar representation can be reduced to two by projecting the basic tetrahedron on an adequately chosen plane, resulting in a complex representation of each DNA base [3].

5. HYBRID APPROACH

A hybrid visualization method is for finding CG-islands in DNA sequences [9]. A CG island is a short stretch of DNA in which the frequency of the CG dinucleotide is 10 to 20 times higher than other regions. The presence of CG islands provide a way to assign the DNA sequence a unique integer and allowing the sequence to be mapped to a corresponding numeric sequence. This numeric sequence is then plotted in 3-D space from which regions with high frequencies of the CG dinucleotide are determined. The method has low computational complexity.

6. COMPARISON OF METHODS

Fractal method take short running time, high accuracy, suitable for any length sequence and provide visual output of complex data set [23,25,26]. 1D method is very simple but not suited for long sequences. Most 2D methods have information loss. 3D method can easily distinguish different sequences directly [16,22]. Geometrical methods are robust, independent, intuitive, and less redundant [4]. Numerical methods are suitable for quantitative analysis [3]. Hybrid method is suitable for finding CG islands [9].

7. CONCLUSION

A survey of various DNA representation methods and taxonomy based on different approaches are presented. Graphical representations are simpler to understand. Z curve and Fractal methods are suitable for visualizing and analyzing the sequences with any length and hence these methods are very suited for DNA sequence representation and analysis.

8. REFERENCES

[1] A. Nandy. Curr. Sci., 66, pp. 309-314 ,1994.

[2] Ashlock, James Golden "Chaos automata: Iterated function systems with memory," Phys.D, vol.181, pp. 274–285, 2003.

[3] Carlo Cattani, "Complex Representation of DNA Sequences" CCIS 13, pp. 528– 537, 2008.

[4] ChunTing Zhang, R Zhang and Hong-Yu Ou, "The Z curve database: a graphic representation, of genome sequences", Bioinformatics Vol. 19 no. 5, pp 593–599, 2003.

[5] D Ashlock, Justin Schonfeld, "A Fractal Representation for Real Optimization", IEEE, pp 87-94, 2007.

[6] Daniel Ashlock and James B. Golden. "Iterated function system fractals for the detection and display of dna reading frame". Proceedings of the Congress on Evolutionary Computation, vol. 2, pp. 1160–116, 2000.

[7] Gao Jie and Xu Zhen-Yuan, "Chaos game representation (CGR)-walk model for DNA sequences", Chin. Phys, Vol 18 No 1, Jan 2009.

[8] Gates M.A, "Simpler DNA sequence representations", Nature, 316, 219. 1985.

[9] Gerard Rambally, Rodney Rambally, "A hybrid Visualization Hidden Markov Model Approach to Identifying CG- Islands in DNA Sequences", IEEE 2008.

[10] Hongxia Zhou and Hong Yan, "Autoregressive Models for Spectral Analysis of Short Tandem Repeats in DNA Sequences" IEEE , pp 8-11 , October 2006.

[11] Hsuan T. Chang, Neng-Wen Lo, Wei C. Lu, & Chung J. Kuo, "Visualization and Comparison of DNA Sequences by Use of Three-Dimensional Trajectories", First Asia-Pacific Bioinformatics Conference proceeding, Vol. 19, 2003.

[12] J W Luo, Li Yang, Y C Zhou , "Gene Identification Based on Geometrical Representation of DNA Sequence", IEEE

[13] Jeffrey, "Chaos game representation of gene structure", Nucleic Acids Research, Vol.18, pp 2163-2171, 1990.

[14] Jiasong Wang and Weiqun Wang, "New 2-D graphical representation of DNA sequences", 2006.

[15] Kang Dai, "A Novel Two-Dimension Graphical Representation of DNA Sequences and Its Numerical Characterization", IEEE , 2008.

[16] Lantin, Carpendale, "Supporting Detail-in-Context for the DNA Representation, H-Curves" , IEEE Visualization , 98

[17] Milan Randic, J.Zupan, Tomaz Pisanski, "On representation of DNA by line distance matrix", Journal of Mathematical Chemistry, Vol. 43, No. 2, pp 674-692, 2008 .

[18] Ming Ji and Chun Li, "TB-curve, a new 2-D graphical representation of DNA sequences", Journal of Mathematical Chemistry, Vol. 40, No. 2, August 2006.

[19] N Chakravarthy, "Autoregressive Models for Spectral Analysis of Short Tandem Repeats in DNA Sequences" IEEE International Conference, pp 1286-1291, 2006.

[20] R. F Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences", Phy. Rev. Lett., vol. 68, no. 25, pp. 3805- 3808, June 1992.

[21] Stephen S. T. Yau, Jiasong Wang, Amir Niknejad, Chaoxiao Lu, Ning Jin and Yee-Kin Ho, "DNA sequence representation without degeneracy", Nucleic Acids Research, pp-3078-3080, Vol. 31, No. 12, 2003.

[22] Y. Zhang, B. Liao, "On 3DD-curves of DNA sequences", Molecular Simulation, Vol. 32, No. 1, 2006, pp. 29–34.

[23] Yu Zu-Guo, Vo Anha, Gong Zhi-Min and Long Shun-Chao, "Fractals in DNA sequence analysis", Chin. Phys(IOP) Vol. 11 No 12, December 2002.

[24] Yuan Wang, Feng-C Tian, X Liu, and J Wang "A Novel Representation Approach to DNA Sequence and Its Application", IEEE Sig Proc Letters, VOL.16, April 2009.

[25] Zu-Guo Yu , Guo-Yi Chen, "Rescaled range and transition matrix analysis of DNA sequences", Commun. Theor. Phys. 33, 2000, (4), pp. 673–678.

[26] Zu-Guo Yu, Long Shi, Qian-Jun Xiao, Vo Anh, "Chaos game representation of genomes and their simulation by recurrent iterated function systems", IEEE pp 41-47, 2008.