ISSN:0975-9646

DNA Representation based on Ladder like Structure Property of DNA Sequence Double Helix

Archana Verma[1], R.K.Singh [2]

[1]Research Scholar, UTU, Dehradun
Astt.Prof. CSE,BTKIT,Dwarahat
Almora, Uttarakhand, India

[2]OSD, UTU, Dehradun
Uttarakhand, India

Abstract— **Biological sequences play a major role in molecular and computational biology.DNA sequence representation remains as one of the critical steps in the analysis of relationships between species. There are several 2-D and 3-D graphical representation methods for DNA sequences. We have introduced a tabular method for representing DNA sequences which will decreases the degeneracy problem occurred in earlier methods. Now we want to utilized our tabular approach for representation. This approach captures the essence of the base composition and distribution of the sequence. The choice of the tabular representation technique for a DNA sequence affects how well its biological properties can be reflected in the graphical domain for visualization and analysis of the characteristics of special regions of interest within the DNA sequence. In this contribution, we take the DNA sequence double helix into account, and make use of an earlier introduced tabular method for representing each DNA sequence based on the frequencies of bases it contains. Then we convert this tabular information to construct a ladder like structure. We are using the double helix property of DNA sequence and its resemblance with ladder like structure. Then an algorithm for preparing a DNA ladder is presented which will utilize the tabular representation of DNA sequences. The method is very simple and fast, and it can be used to analyze both short and long DNA sequences. The utility of this method is tested on Goat alanine β -globin 86 bases. This paper also presents a summary of various DNA graphical representation methods and their applications in envisaging and analyzing long DNA sequences. A discussion on the comparative merits and demerits of the various methods and conclusions has also been included.**

Keywords— DNA, base characters, visualization, gene, 2D and 3D graphical representations, RA method, biological sequences.

I. INTRODUCTION

DNA is a nucleic acid that contains the genetic instructions used in the development and functioning of all living organisms and viruses. The main role of DNA molecules is the long term storage of information regarding their biological processes. DNA consists of double stranded anti-parallel helix built by concatenating nucleotide blocks namely Adenine (A),Thymine (T), Cytosine (C) and Guanine (G). Biologists need to observe the useful features of long DNA sequences that include several thousands or several tens of thousands of bases. With the alphabetic representation of DNA sequences, it is difficult to observe meaningful features in the sequence. DNA sequence visualization would provide a simple, friendly, immediate and interactive graphical display such that the users can easily observe the global and location visual features in a long DNA sequence. DNA sequence visualization can be possible with the help of graphical representation methods, therefore several graphical methods for representing DNA sequences have been developed. DNA sequence visualization should also facilitate the analysis, comparison and identification of many DNA sequences. The problem of degeneracy arises in the graphical representation method, to solve that problem we have proposed a tabular method for representation which decreases the problem of degeneracy.

II. EXISTING REPRESENTATION METHODS

Existing methods of DNA sequence visualization can be classified into either 2D or 3D graphical representation. Examples of 3D graphical representation of DNA sequences include the H-Curve, the Chaos game display and the W-Curve. The 3D graphical representations can uniquely characterize a DNA sequence, but the disadvantage is that it is complicated, inconvenient, and requires the display of 2D projections or 3D stereo projections for visualization and analysis. Examples of 2D graphical representation of DNA sequences include the methods proposed by Gates, Nandy, Leong and Mogenthaler.

Various Representation of A DNA Sequence: A reduced two-dimensional graphical representation of DNA sequences was proposed by M. A. Gates, Nandy, Leong, and Mogenthaler. Their method is based on choosing the four cardinal directions in (x, y) coordinate system to represent the four bases in DNA sequences. The method essentially consists of plotting a point corresponding to a base by moving one unit in the positive or negative x− or y−axes depending on the defined association of a base with a cardinal direction. The cumulative plot of such points produces a graph that corresponds to the sequence. In the Gates axes system, one would move one unit in the positive x−direction for a cytosine (C), along the positive y-direction for a thymine (T), the negative x−direction for a guanine (G), the negative y−direction for an adenosine (A), implying a cumulative plot of the count of instantaneousC−Gagainst T − A. The Nandy axes system associates G with positive x−direction, C with positive y−direction, A with negative x−direction, and T with negative y−direction. In the Leong and Morgenthaler axes system, A is associated with positive x−direction, T with positive y−direction, C with negative x−direction, and G with negative y−direction.

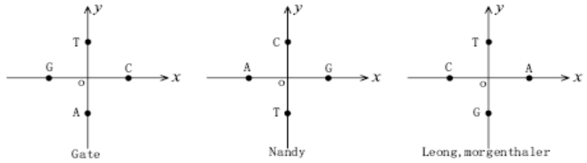

Fig 1: The three independent orthogonal axes systems

It is clear that the differences in the base composition and distribution of individual members of a homologous family will induce changes in the plots of the sequences in the graphical representation method outlined. In general, changes in base composition will result in changes in endpoints represented by the coordinates (G − A,C − T), Let C = x1x2 · · · xn be a DNA sequence which forms a circuit in its graphical representation, FA, FC, FG, FT be the frequencies of A, C, G, T in C, respectively. Let xm be the m th base in a sequence, and fA, fC, fG, fT be the frequencies of A, C, G, T in x1x2 · · · xm, respectively.

If a kind of graphical representation of DNA sequences has no circuit, then each DNA sequence can be uniquely determined by the graphical representation of it.

To reduce the degeneracy of Nandy's graphical representation, more than two of the four unit vectors that represent the corresponding bases must be deviated from their original cardinal axes directions. We design four special vectors in Cartesian (x, y) coordinate system to represent the four nucleic acid bases A, C, G, T , which the initial point of the vectors is the cardinal origin. A DNA sequence of four letters A, C, G, T with length n can be regarded as a successive vector sequence V1, V2, · · · , Vn of length n consisting of the four vector sequence corresponding to A, C, G, T . A vector sequence V1, V2, · · · , Vn is said to be a successive sequence if V1, V2, · · · , Vn are shifted parallel so that, for 2 ≤ i ≤n, the initial point of Vi is identical with the terminal point of Vi−1 step by step. The graphical representation of DNA sequence may be regarded as a directed walk in digraph. There are three methods of designing four special vectors in Cartesian (x, y) coordinate system to represent the four bases A, C, G, T as following figures, where m is an integer greater than 1.

It should be mentioned here that there are also three possible independent axes systems for the novel graphical representation of DNA sequences, which are respectively corresponding to axes systems of Nandy, Gates, and Leong and Morgenthaler. The axes systems of the graphical representation is shown in Fig. 2

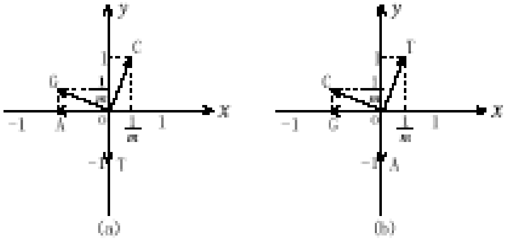

Fig 2: A Novel 2-D graphical representation of DNA sequence of low degeneracy

In the The Nandy axes system, the graphical representation of 8 β−globin are shown in Fig. 3

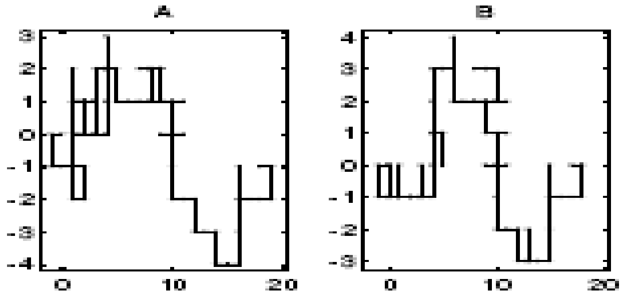

Fig 3: Graphical representation of 8 β−globin DNA sequence

We have proposed a new tabular method which uses a table to represent DNA sequence with the base characters and their relative position in the sequence. The algorithm for this method is as follows:

Steps of RA method
1. Create a table of four rows and two columns.
2. Initialize all rows of the base field as one of the base character i.e.{a, t, g, c}.
3. Read unprocessed character (k) in the sequence.
4. Find the match for the character (k) with the Base field value.
5. Then write the relative location of k in the relative position field.
6. Go to step 3 until EOF of the sequence.
7. End

III. Enhancing the previous representation

In the representation of DNA there is complementarity between DNA double-strands, as an A on one strand always binds with a T on the other strand, and similarly a C always binds with G. This property is used for representing only half side of a DNA sequence i.e. other half portion of DNA can be considered with the complemetarity property. When we want to represent the complete DNA, both the strands and all their base values must be visible in the representation pattern. The earlier representations display only one side of the DNA with the implicit other half portion of DNA. Our proposed method will display the complete information about a DNA with both its strands and all the base characters it contains. In this method we utilize our tabular representation method (i.e. RA method). We just translate the given sequence into a table and then the tabulated information is used for constructing a ladder like structure.

Generally DNA is helical in structure, if we just stretch the DNA helix this will construct a ladder like structure each side of which is used for the representation of DNA strands and bases are distributed on these sides. All the base characters are distributed on each side of the ladder according to their relative position. This idea is used for the new representation method in which we construct a ladder. Algorithm for the construction of ladder like structure for a DNA sequence:

LADDER ALGORITHM:

```
Let S be the given sequence which is represented in a table with the base characters and their relative positions.
Step 1: Initialize location (loc=1) and MAX is the end of the sequence.
Step 2: Find loc in the RA(Row Access) table and read the corresponding base character.
Step 3: if the character =A|T|G|C then
Put the base character and move one position right and put the relative base character
        loc++
move one position upward
Step 4: repeat step 2 and step 3 while loc<= MAX.
Step 5: END.
```

IV. EXPERIMENTAL RESULTS

We apply the RA method on the following DNA sequence –Goat alanine β -globin 86 bases ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGG CT TCTGGGGCAAGGTGAAAGTG GATGA AGTT GGTGC TGAGGCCCTGGGCAG.

Base	Relative Position
A	0,6,13,16,18,19,21,30,45,46,51,52,53,58,61,62,73,83
C	3,7,10,22,25,26,29,31,32,35,38,44,70,76,77,78
G	2,5,9,12,14,15,17,20,24,27,33,34,40,41,42,43,47,48,50,54, 56,57,60,63,66,67,72,74,75,80,81,82,84
T	1,4,8,11,23,28,36,37,39,49,55,59,64,65,68,69,71,79

Fig 4: representation of DNA using RA method

Applying our proposed ladder algorithm onto the above sequence we find the following ladder like structure (fig 5).

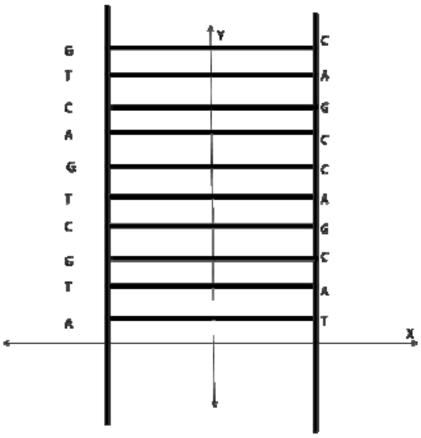

Fig 5: proposed ladder like structure for DNA

V. COMPARISONS & ANALYSES

The earlier graphical methods for DNA representations show degeneracy problem (Nandy, Gates, Leong and Mogenthalar). The tabular representation is an efficient method which overcomes the degeneracy problem and also easy to implement. In tabular method the sequence prediction and the results rely on mapping the nucleotides and their relative positions. This treatment has shown to be useful for locating periodicities and for locating potential gene sequences characterized by simple non uniform sequences. Studies indicate that usual representations show only half side(strand) of the DNA and it is not able to detect the features in some viral and bacterial organisms. It will be complicated while using previous one stranded view of DNA sequences due to the extra effort to add the information regarding the other half strand each and every time. The resemblance of a DNA helix with a simple ladder structure provides us a notion to represent the complete double strand in a better way so that the whole information can be viewed.

Obviously the unnecessary effort for loading the other half sequence will be eliminated with this approach. This will decrease the complicacy of previous representation methods and make the process of analysis easier and more comfortable. This can be viewed with our experimental results.

VI. CONCLUSIONS

Graphical representation of DNA sequence may provide a simple way of viewing, sorting and comparing various gene structures. The previous studies of the 2-D graphical representations of DNA sequence used varying approaches but they hold a high level of degeneracy. The problem of degeneracy was recovered with the RA method of representation. We recently presented a graphical representation for both the DNA strands which utilizes the RA method. Unlike previous representation methods this method displays the complete double stranded view of a DNA sequence. The study of DNA sequences with this method will be quicker and simpler. Also this approach is free from the problem of degeneracy. The complete information of a DNA can be viewed and retrieved very easily.

REFERENCES

[1] Swarna Bai Arniker and Hon Keung Kwan,"Graphical Representation of DNA sequences" , IEEE International Conference on Electro/ Information Technology, EIT 2009, Windsor, Ontario, Canada, June 7-9, 20092009.

[2] Rajendra Kumar Bharti, Archana Verma, Prof. R.K.Singh,"A New 2-D RA Method of Representation and Analysis of a DNA Sequence", International Conference on Networking and Information Technology (ICNIT), Philipines, 10.1109/ICNIT.2010.5508475.

[3] Liu Xikui, Li Yan, "Some Notes on 2-D Graphical Representation of DNA Sequence" Proceedings of the 27th Chinese Control ConferenceJuly 16–18, 2008, Kunming, Yunnan, China.

[4] HAMORI J, RUSKIN J. H curves - A novel method of representation of nucleotide series especially suited for long DNA sequences[J]. J. Bio.Chem, 1983, 258: 1318-1327.

[5] NANDY A. A new graphical representation and analysis of DNA sequence structure I. Methodology and application to globin genes[J]. Curr. Sci. 1994, 66: 309-313.

[6] HAMORI J, RUSKIN J. H curves - A novel method of representation of nucleotide series especially suited for long DNA sequences[J]. J. Bio.Chem, 1983, 258: 1318-1327.

[7] NANDY A. Graphical analysis of DNA sequence structure:III. Indications of evolutionary distinctions and characteristics of intros and exons[J]. Curr. Sci. 1996, 70: 611-668.

[8] LEONG P M, MORGENTHALER S. Random walk and gap plots of DNA sequences[J]. Comput. Applic. Biosci. 1995, 11: 503-507.

[9] Roy, C. Raychaudhury, A. Nandy, "Novel techniques of graphical representation and analysis of DNA sequences- A review," Journal of Biosciences, vol. 23, pp. 55-71, March 1998.

[10] NANDY A, NANDY P. Graphical analysis of DNA sequence structure II: Relative abundances of nucleotides in DNAs, gene evolution and duplication[J]. Curr. Sci. 1995, 68: 75-85.