International Journal of Scientific & Engineering Research, Volume 6, Issue 10, October-2015
ISSN 2229-5518

DNA Sequencing Similarity Analysis: Graph theory Application

Dr. Anjaneyulu G.S.G.N , Arush Kamboj, Somanshu Kalra

SCSE,VIT UNIVERSITY

kalra.somanshu@gmail.com, arushkamboj@gmail.com

Abstract:

Graph theory is study of mathematical structures called graphs which are represented by nodes (or vertices) and edges. As we know, during evolutionary history, not along with DNA mutation subsequent rearrangements also occurred for individuals. This study involves the use of graph theory to construct a mathematical descriptor for similarity analysis based on various mutation phenomena. As a DNA sequence can store considerable amount of computational data, a weighted directed graph will be set up for each DNA sequence. Each edge is assigned a weight in accordance with the distance to be travelled. This approach takes into account both ordering as well as the frequency of nucleotides so that more data is involved.

Keywords *- DNA, Quadratic hashing, bidirectional graph, Vigenere cipher, poly alphabetic cipher*

Introduction:

The number of DNA sequences is rapidly increasing in the DNA database. It is one of the major challenges for bio-scientists to analyze the large volume of genomic DNA sequence data. It has been a major challenge for the biologists that low time-complexity alignment free methods are needed for proper measurements of sequence similarity.

Graph theory is rapidly moving into the mainstream of mathematics mainly because of its applications in diverse fields which include biochemistry (DNA Sequencing

Similarity Analysis) and many others. The wide scope of these applications has been well-documented. The powerful combinatorial methods found in graph theory have also been used to prove significant and well-known results in a variety of areas in biochemistry itself.

In this paper, we try introduce a method which combines the results of compact representations and solution based on graphical representation of nucleotides, to represent DNA sequences mathematically for similarity analysis to give the solution more effectively for time constrained analysis problems.

Graphical Representation of DNA sequences:

Alphabets A, G, C, T represents a sequence of DNA. We represent more than once occurrence of a DNA base by A_n, C_n, T_n and G_n respectively. Let $S = s_1, s_2, s_3 \ldots s_n$ be a sequence of a DNA where $s_1, s_2, s_3 \ldots s_n$ is a pair of nucleotides. The graph is drawn between the values of s_i and s_j where i, j will vary from 0 to n where n being the length of the given DNA sequence. We calculate Z-transform using the given equations [1,2]

$a_n = (A_n + G_n) - (C_n + T_n)$

$b_n = (A_n + C_n) - (G_n + T_n)$ (n=1,2,…..N)

$c_n = (A_n + T_n) - (G_n + C_n)$

Now if represented in 2D-catesian coordinated, the four bases of DNA can be distributed in to four quadrants.

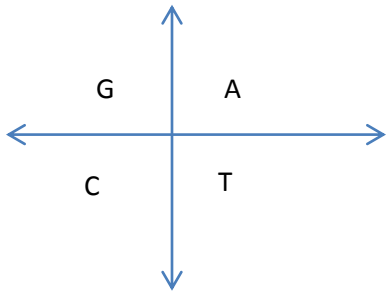


Figure 1: distribution of DNA bases into quadrants.

A Compact Graphical Representation of DNA sequences:

We derive a formula for assigning weights to the edges of graph based on the fact that the force of interaction between two is inversely proportional to the distance between them. Therefore if w_1 represents the weight of an edge connecting *A and G* spaces at a distance of d_1 and w_2 is another edge between the same nucleotides but spaces at a distance d_2 the values of w_1 will be

- $> w_2$ if d_1 is less than d_2
- $< w_2$ if d_1 is greater than d_2

We derive a general formula for assigning weights to different edges based on the results calculated above. Hence weights between s_i and s_j (s_i, s_j were defined earlier) can be calculated by the formula $w = (1/d^{\alpha})$ where $d=j-i$ and α is a constant from 0 to 1.

After calculating the weights for all the edges we attempt to draw a compact graph of the given sequence. But before that we transform each DNA into binary notation.

Note: DNA can be converted into binary by converting each nucleotide into its binary form. For example A becomes 00, G becomes 01, C becomes 10 and T becomes 11.

Table 1 (DNA Binary Representation)

Nucleotide	Binary representation
A	**00**
C	**01**
G	**10**
T	**11**

Hence a sequence ATCTCATACC becomes 00111011100011001010. Similarly any coding sequence can be converted into its binary equivalent. Some of the coding sequences of different species is given below.

THE CODING SEQUENCES OF THE FIRST EXON OF β-GLOBIN GENES OF 11 SPECIES.

Species	Coding sequence
Human	ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGT GGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGTGGATGA AGTTGGTGCTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAGGT TGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTCAATGT GGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGGATGT AGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Mouse	ATGGTGCACCTGACTGATGCTGAGAAGGCTGCTGTCTCTTGCCTGTGGGGAAAGGTGAACTC CGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGTGAATGT GGAAGAAGTTGGTGGTGAGGCCCTGGGC
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAACCC TGATAATGTTGGCGCTGAGGCCCTGGGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGT GGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAAGTGGATGA AGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGT GGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG

Figure 2: Coding sequence of first axon of β-Globin genes of 11 species.

An observation is the length of binary representation is twice the length of given DNA sequence. Using these values worm curve can be drawn [1, 2] which is basically a graphical representation of s given sequence drawn with a rule "after each step make a line perpendicular to the right of previously covered line if you haven't reached to your starting position else make the perpendicular to your left. The example curve for a tested data is shown below.

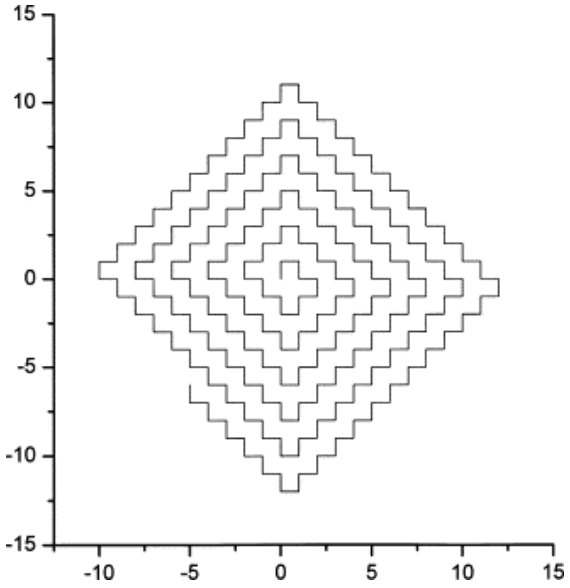


Figure 3: A sample worm graph of a testing data.

The above mentioned algorithm has a time a time complexity of $O(n^2)$. We modify this algorithm to reduce the time complexity for maximum possible sequences in order to get effective solution for the graph which is later to be examined for similarity analysis and hence.

New Algorithm for Compact Representation:

```
For i = 1 To N
 {
        P1 = Points(Points.Count - 1);
        P2 = Points(Points.Count);     //The
last point
        dir = Piont2 - Piontl ; //the direction
of moving (left or right)

       NextPiont = tum right, go one unit
and get next point;
        if NextPoint exist in WPoints then
```

```
            if w.Point > w.NextPoint  /*if
        the weight is point is lesser than the
        next point
                    NextPiont = turn left,
            go two units and get next
            point;
            end if
        end if
        Points.Add(NextPiont);
}
SPoints=GetSpottedPoint(Points, BSEQ);
/*If the i-th letter in BSEQ is '1', get the i-th
point in WPoints.*/
G = DrawGraph(Points, SPoints);
```

Why the above solution is better:

w.Point > w.NextPoint means that the distance between the nucleotides of given point "Point" is lesser than distance between the nucleotides of NextPoint. Hence if the edge with lesser weight is left for the end we can reduces the number of *left* turns because we will always find a node with weight lesser than the current weight making traversal of the edges of the graph having higher precedence in minimum number of turns. Hence before actually drawing the graph we can reduce the number of edges to be covered until a given specific moment.

The most important benefit of our modified algorithm is that it works best if the sequence was enforced upon a time constraint which is easily the case in most examples. Hence if similarity analysis is to be done on a sequence but the results were to declare the similarity only until the K_{th} nucleotide this modified algorithm is capable of most specific results. Also the time complexity reduces to O(n) for time specific compaction of a given sequence. Although the time complexity remains the same ($O(n^2)$) if the ultimate task is to compact the whole sequence for analysis.

Formation of Directed Acyclic Weighted Graph:

After reducing the number of turns we form the directed acyclic weighted graph for final analysis. As the DNA has four bases the graph will be drawn with four vertices. The weights of each edge will be assigned according to the formula derived earlier. Hence in the given graph let d_1, d_2, d_3 be the distances between AC, AT and AG the weights will be as follow:

For $\alpha = 1/2$

$w_1 = 1/\sqrt{d_1}$

$w_2 = 1/\sqrt{d_2}$

$w_3 = 1/\sqrt{d_3}$ and so on

The parameter α can different for different cases and will vary $0 < \alpha < 1$.

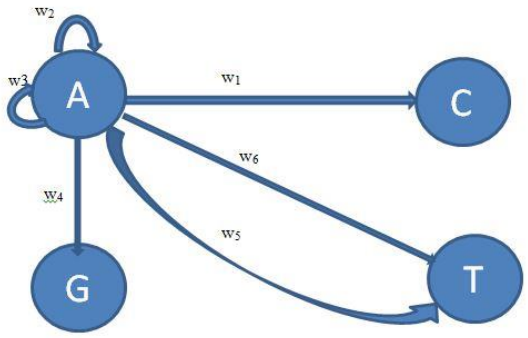


Figure 4: Directed Acyclic Graph for sequence ACTATG

The weights of the nucleotides directing at the same destination will be combined and a simplified DAG is drawn.

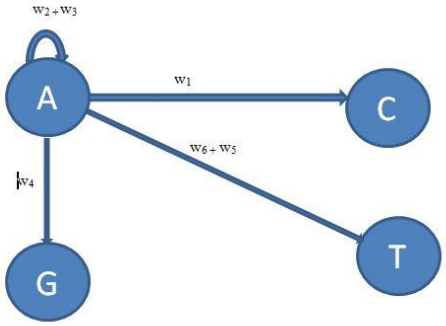
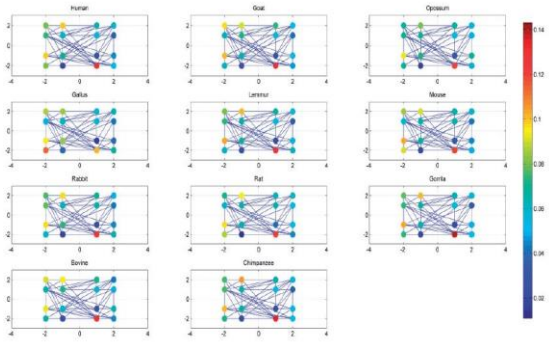


Figure 5: Simplified Directed Acyclic Graph for sequence ACTATG

Similarity Analysis Based on Compact Representation:

Graphical representation is a very useful and important method for similarity analysis. This method was proposed by Randi at al in 2003. The main area of this method was quantitative characterization of DNA sequences. Since this method many descriptors have been proposed. This field has been one main motivation of graphical representations. Researchers have done most of their works in the analysis of similarity in the DNA sequences of the first axon of β globin genes. In this section we study the similarity analysis by graphical representation of DNA.

As discussed earlier we can represent nucleotides of any DNA in a 2-D, 3-D or 4-D curve. We also presented the coding sequence of 11 species which shows the projections of D-curves of the presented 11 species. The snapshot of D-curves is given below



A close observation of the above figure shows that Human – Gorilla, Human – Chimpanzee, Gorilla – Chimpanzee are the most similar species.

Conclusion:

This paper combines the solutions of sequence analysis of DNA nucleotides given by graphical representation method and compact representation method. This paper modified the algorithm used in compact representation method. The proposed algorithm works best when the solutions are constrained upon time and runs with same complexity as other solutions for normal cases.

References:

[1]. A Novel Model for DNA Sequence Similarity Analysis Based on Graph Theory Xingqin Qi2, Qin Wu1, Yusen Zhang2, Eddie Fuller1 and Cun-Quan Zhang

[2]. Similarity analysis of DNA sequences based on a compact representation Zhujin Zhang #*1, Shuo Wang #2, Xingyi Zhang +3, Zheng Zhang #4

[3]. Analysis of Similarities/Dissimilarities of DNA Sequences Based on a Novel Graphica Representation Jia-Feng Yu *a*, *b*, Ji-Hua Wang *b*, Xiao Sun *a*

[4] . Potential of Graph Theory Algorithm Approach for DNA Sequenc Alignment and Comparison Syed Abdul Mutalib Al Junid, Nooritawati Md Tahir, Zulkifli Abd Majid, Mohd Faizul Md Idros

[5] . http://en.wikipedia.org/wiki/Bidirected_graph.

[6] . http://en.wikipedia.org/wiki/Cryptographic_hash_function.

[7] . https://en.wikipedia.org/wiki/DNA