A STUDY ON ALGORITHMIC APPROACH ON DNA SEQUENCE USING GRAPH THEORY

M. Akalya

Assistant Professor
Department of Mathematics
KG College of Arts and Science, Coimbatore-641035, Tamil Nadu, India

ABSTRACT: This paper is a study on DNA sequencing by hybridization and reconstructing the DNA molecule based on fragments overlap via shortest common Superstring (SCS) problem. Some problems are solved with suitable graphical diagrams to justify the algorithm which was developed by Pranab Kalita and Bichitra Kalita to solve the combinatorial part of DNA sequencing by hybridization using graph theoretical concepts to find the Hamiltonian path. Here the spectrum is ideal one and the fragments are of equal length is to be assumed.

Keywords: DNA sequence, Fragments, Hamiltonian path, SCS problems, SBH problem, TSP problem.

INTRODUCTION

DNA or deoxyribonucleic acid is the hereditary material in humans and almost all other organisms. The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). DNA bases pair up with each other, A with T and C with G, to form units called base pairs.

The two DNA strands are termed as polynucleotides, it has a form of double helix since they are composed of simpler monomer units is called as nucleotides. Each nucleotide is composed of four nitrogen containing nucleobases - either cytosine (C), guanine (G), adenine (A), or thymine(T) and a sugar called deoxyribose and a phosphate group.

From the modern biology structure of DNA is reshaped using DNA Sequencing, DNA mapping and DNA assembling. DNA sequencing problem is used to determine the overlapping fragments. The spectrum may contain both positive and negative errors also spectrum without any error is said to be ideal one. There two types of methods for DNA sequencing are Sanger method and SBH method.

Here, SBH (Sequencing by Hybridization) method is used to reconstructing the DNA sequence, since the algorithm is approached for ideal spectrum ie., the spectrum without any errors.

LINK BETWEEN DNA SEQUENCING AND GRAPH THEORY

The first algorithm reconstructing the original sequence on the base of a spectrum was proposed the spectrum used there does not contain any errors. This branch-and-cut algorithm builds a search tree, where spectrum elements correspond to nodes and two nodes are connected by an arc if last l - 1 letters of the predecessor cover first l - 1 letters of the successor. The element once included into the current path cannot be visited the second time. As the root this oligonucleotide is taken, which begins the original sequence. If we do not know it, the algorithm must construct |S| search trees differing in their roots. The solution is the path from the root to a leaf, which contains all spectrum elements.

The second approach to the DNA sequencing problem refers to a well-known problem from graph theory. In a directed graph, built on the base of an ideal spectrum, the Hamiltonian path is looked for. Each vertex in the graph corresponds to other element of the spectrum. Two vertices u and v are connected by the arc (u, v) if last l - 1 letters of the label (oligonucleotide) of u cover first l- 1 letters of the label of v.

In an approach similar to the one from was presented, but it avoids excess steps. Sub paths between points of branching in the tree are combined into words of length l or greater. The search tree is built on the base of these words, and two nodes are

connected if they overlap on l - 1 letters. Similarly, the solution is the path from the root to a leaf, containing all oligonucleotides from the spectrum.

Example:

Let us assume that for the original sequence ACTCTGG, the hybridization experiment has been performed without errors and the ideal spectrum has been generated:

S = {ACT, CTC, CTG, TCT, TGG}. We see, that |S| = n - l + 1, where n = 7 and l = 3. Assume that we know the first oligonucleotide of the original sequence. The lower path traverses through all elements of the spectrum, so it is the solution for the problem. The original sequence can be reconstructed by reading the labels of the vertices from the root to the leaf.

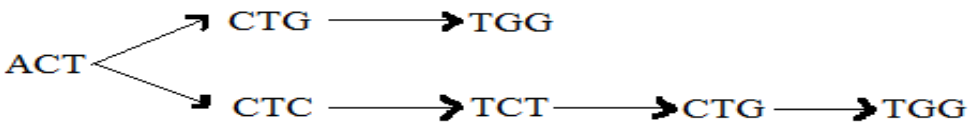


Figure 1.

In the graph exactly one Hamiltonian path exists, and it corresponds to a unique sequence of length n possible to build from the spectrum. The method from search a similar tree but sub paths between branching nodes are combined into vertices. The graph constructed by the method of Lysov et al. is the line graph of the graph by Pevzner for the same spectrum. For that pair of directed graphs the two problems of looking for the Hamiltonian and Eulerian paths are equivalent.

In this network we look for the flow of value m - 1 and of the minimum cost. If the cost appears to be equal n-l+1=|S|, the base graph will be completed by arcs (paths) composing the flow.

Sequencing by hybridization (SBH) is one of the most popular methods from the computational molecular biology domain. In SBH, assumptions are- spectrum is ideal one and fragments are of equal length *l* composing the original sequence. The spectrum is the set of all possible *(n-l+1) l-mers (lengths*) in a string *s* of length *n* and it may be denoted as , spectrum *(s,l)* .

The DNA sequencing problem can also be stated as the problem of constructing a string over $\sum$ = {A, C, G, T} from a given spectrum (not necessarily an ideal spectrum) S = {S_1, S_2,…,S_m}, so that the resulting string is the shortest string which contains as many of the fragments in the spectrum as possible.

If we consider, S = {S_1, S_2,…,S_m} over $\sum$ = {A, C, G, T} then the solution is of string that contains all s_i of S. Without loss of generality, we assume that there is no strings s_i, s_j ε S, i≠j such that s_i is a substring of s_j.

Now, we proposed a new algorithmic approach to solve DNA sequencing problem as the shortest common superstring (SCS) problem for ideal spectrum and fragments with constant length '*l*' composing the original sequence by means of graph theory.

ALGORITHM

Step I: For a spectrum we define a complete graph $K_{|S|}$ = (V, E, W) where,

V = S (One vertex v_i corresponds to one fragment s_i).

E = {(v_i, v_j) : (v_i, v_j) is an ordered pair}

W(v_i, v_j) = Overlap (s_i, s_j) = |w_{ij}|, where $s_i = xw_{ij}$, $s_j = w_{ij}y$ for i≠j and W(v_i, v_j) = 0 if i = j.

Step II: Define an overlap matrix M = (a_{ij}), where a_{ij} = W (v_i, v_j).

Step III: From matrix M = (a_{ij}), we develop an algorithmic for obtaining a parenthetical tree (T) as follows:

1. Starting vertex (root) = v_s, if Cv_s is minimum.
2. Ending vertex (leaf) = v_e, if Rv_e is minimum.
3. v_s follows v_i, s≠ i, if a_{si} is maximum in the sth
4. (v_s, v_i) follows v_j if a_{ij} is maximum in the ith row and so on.
5. Repeat the sub-step 4 until we get Ending vertex v_e.
6. SCS = (((($v_s v_i$)v_j)…v_e) provided each v_i ε V is taken for single time only.
7. Stop.

Here, $\sum Cv_i$ is the sum of the elements in the v_ith column, $\sum Rv_i$ is the sum of the elements in the v_ith row and $(v_iv_j) = (s_is_j) = xw_{ij}y$.

Now we look for SCS's by finding a Hamiltonian path (from root to leaf, which contains all fragments for single time only) of maximum overlap and $|s| = L - \sum |w_{ij}|$ ($|s|$ is minimum when $\sum |w_{ij}|$ is maximum), where L is the total length of the strings which is fixed by the problem, hence constant for all Hamiltonian paths and therefore it has been converted to Traveling Salesman Problem (TSP).

Example 1:

Step I: For a spectrum we define a complete graph $K_{|S|} = (V, E, W)$ where,

$V = S$ (One vertex v_i corresponds to one fragment s_i).

$E = \{(v_i, v_j) : (v_i, v_j)$ is an ordered pair$\}$

$W(v_i, v_j) =$ Overlap $(s_i, s_j) = |w_{ij}|$, where $s_i = xw_{ij}$, $s_j = w_{ij}y$ for $i \neq j$ and $W(v_i, v_j) = 0$ if $i = j$.

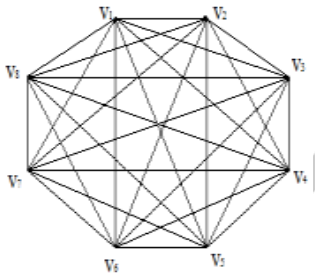
Suppose the ideal spectrum:

$$S = \{ATG, AGG, TGC, TCC, GTC, GGT, GCA, CAG\}$$

DNA sequencing and SCS problem corresponds to the set of vertices:

$V = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}$, then $K_8 = (V, E, W)$

The complete graph is



(OR)

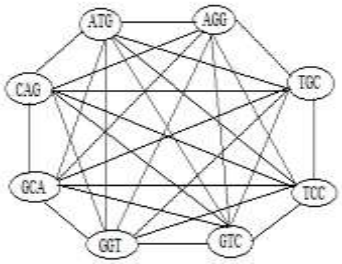


Figure 2.

Step II: Define an overlap matrix $M = (a_{ij})$, where $a_{ij} = W(v_i, v_j)$.

The overlap matrix:

M =

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	$\sum Rv_i$
V_1	0	0	2	0	1	1	1	0	5
V_2	0	0	0	0	1	2	1	0	4
V_3	0	0	0	0	0	0	2	1	3
V_4	0	0	0	0	0	0	0	1	1
V_5	0	0	0	2	0	0	0	1	3
V_6	0	0	1	1	2	0	0	0	4
V_7	1	1	0	0	0	0	0	2	4
V_8	0	2	0	0	1	1	1	0	5
$\sum Cv_i$	1	3	3	3	5	4	5	5	

Step III: From matrix $M = (a_{ij})$, we develop an algorithmic for obtaining a parenthetical tree (T) as follows:

1. Starting vertex (root) = v_s, if Cv_s is minimum.
2. Ending vertex (leaf) = v_e, if Rv_e is minimum.
3. v_s follows v_i, $s \neq i$, if a_{si} is maximum in the sth
4. (v_s, v_i) follows v_j if a_{ij} is maximum in the ith row and so on.
5. Repeat the sub-step 4 until we get Ending vertex v_e.
6. SCS = $(((v_sv_i)v_j)\ldots v_e)$ provided each $v_i \varepsilon V$ is taken for single time only.
7. Stop. And Parenthetical tree (T):

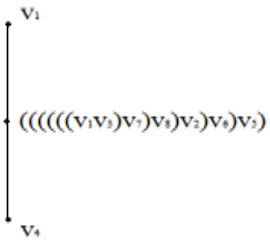


Figure 3.

Here, $\sum Cv_i$ is the sum of the elements in the v_ith column, $\sum Rv_i$ is the sum of the elements in the v_ith row and $(v_iv_j) = (s_is_j) = xw_{ij}y$.

Hence the SCSs are

$$S = v_1v_3v_7v_8v_2v_6v_5v_4 = ATGCAGGTCC$$

Now we look for SCS's by finding a Hamiltonian path (from root to leaf, which contains all fragments for single time only) of maximum overlap and $|s| = L - \sum |w_{ij}|$ ($|s|$ is minimum when $\sum |w_{ij}|$ is maximum), where L is the total length of the strings which is fixed by the problem,

$$|s| = L - \sum|w_{ij}| = 24 – (2+ 2+ 2+ 2+ 2+ 2+ 2) = 10.$$

Hence constant for all Hamiltonian paths and therefore it has been converted to Traveling Salesman Problem (TSP).

Example 2:

Step I: For a spectrum we define a complete graph $K_{|S|} = (V, E, W)$ where,

$V = S$ (One vertex v_i corresponds to one fragment s_i).

$E = \{(v_i, v_j) : (v_i, v_j)$ is an ordered pair$\}$

$W(v_i, v_j) = \text{Overlap}(s_i, s_j) = |w_{ij}|$, where $s_i = xw_{ij}$, $s_j = w_{ij}y$ for $i \neq j$ and $W(v_i, v_j) = 0$ if $i = j$.

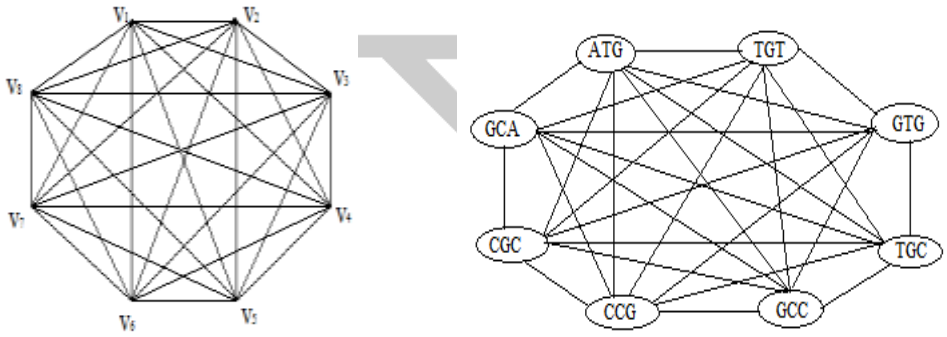
Suppose the ideal spectrum:

$$S = \{ATG, TGT, GTG, TGC, GCC, CCG, CGC, GCA\}$$

DNA sequencing and SCS problem corresponds to the set of vertices:

$V = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}$, then $K_8 = (V, E, W)$

The complete graph is



(OR)

Figure 4.

Step II: Define an overlap matrix $M = (a_{ij})$, where $a_{ij} = W(v_i, v_j)$.

the overlap matrix:

M =

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	$\sum Rv_i$
V_1	0	2	1	2	1	0	0	1	7
V_2	0	0	2	1	0	0	0	0	3
V_3	0	2	0	2	1	0	0	1	6
V_4	0	0	0	0	2	1	1	2	6
V_5	0	0	0	0	0	2	1	0	3
V_6	0	0	1	0	1	0	2	1	5
V_7	0	0	0	0	2	1	0	2	5
V_8	1	0	0	0	0	0	0	0	[1]
$\sum Cv_i$	[1]	4	4	5	7	4	4	7	

Step III: From matrix M = (a_{ij}), we develop an algorithmic for obtaining a parenthetical tree (T) as follows:

1. Starting vertex (root) = v_s, if Cv_s is minimum.
2. Ending vertex (leaf) = v_e, if Rv_e is minimum.
3. v_s follows v_i, $s \neq i$, if a_{si} is maximum in the sth
4. (v_s, v_i) follows v_j if a_{ij} is maximum in the ith row and so on.
5. Repeat the sub-step 4 until we get Ending vertex v_e.
6. SCS = $(((v_s v_i)v_j)\ldots v_e)$ provided each $v_i \varepsilon V$ is taken for single time only.
7. Stop. And

Parenthetical tree (T):

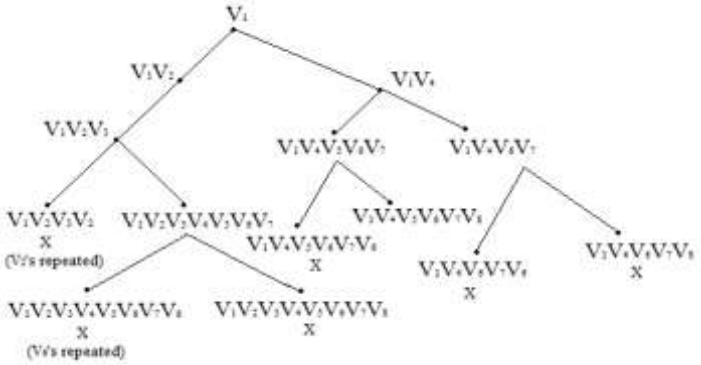


Figure 5.

Here, $\sum Cv_i$ is the sum of the elements in the v_ith column, $\sum Rv_i$ is the sum of the elements in the v_ith row and $(v_i v_j) = (s_i s_j) = x w_{ij} y$.

Hence the SCSs are

$$S = v_1v_2v_3v_4v_5v_6v_7v_8 = ATGTGCCGCA$$

Now we look for SCS's by finding a Hamiltonian path (from root to leaf, which contains all fragments for single time only) of maximum overlap and $|s| = L - \sum |w_{ij}|$ ($|s|$ is minimum when $\sum |w_{ij}|$ is maximum), where L is the total length of the strings which is fixed by the problem,

$$|s| = L - \sum |w_{ij}| = 24 - (2+ 2+ 2+ 2+ 2+ 2+ 2) = 10.$$

Applying TSP to DNA fragment assembly might be considered excessive, even though the expected results are obtained, since the fraction of real connections is very small. A graph theoretical approach using other algorithms might give better results.

CONCLUSION

This paper highlighted a new approach which is an application of graph theory in the field of DNA. The new approach mentioned in this paper is to reconstruct a DNA molecule via shortest common superstring (SCS) problem. Also, we solved some more problems according to the algorithmic approach.

REFERENCES

[1] Pranab Kalita and Bichitra Kalita, "A Graph Theoretical Algorithmic Approach for DNA Squencing", IOSR Journal of Mathematics, Vol.5, pages 40-46, February 2013

[2] M. Kasprzak, On the link between DNA sequencing and graph theory, Computational Methods in Science and Technology, 10(2004), pages 39-47

[3] Trajkovski, Lecture 3: DNA sequencing, a power point presentation

[4] J. Blazewicz and M. Kasprzak, Complexity of DNA sequencing by hybridization, Theorical Computer Science, 290(2003), 1459- 1473

[5] Ben Langmead, "Assembly & shortest common superstring", a PowerPoint Presentation