A Graph Theoretical Algorithmic Approach for DNA Sequencing

Pranab Kalita¹, Bichitra Kalita²

¹Pranab Kalita, Department of Mathematics, Gauhati University, Guwahati-781014, Assam, India
²Department of Computer Applications (M.C.A), Assam Engineering College, Guwahati-781013, Assam, India

Abstract: In this paper, an algorithmic approach is developed to solve the combinatorial part of DNA sequencing by hybridization using graph theoretical concepts. It is assumed here, that the spectrum is ideal one and fragments are of equal length. The goal is to reconstruct the DNA molecule based on the fragments overlap. Suitable examples are also provided to justify the algorithm.

Keywords: DNA sequencing, fragment, spectrum, SCS problem, SBH problem, TSP problem, overlap matrix, parenthetical tree.

I. Introduction and Preliminaries

Deoxyribonucleic acid (DNA) is the chemical in the cells of animals and plants that carries genetic information. It consists of two strands, each of which contains nucleotides obtained from the set $\Sigma = \{A, C, G, T\}$, where $A$, $C$, $G$, $T$ are for Adenine, Cytosine, Guanine and Thymine respectively. The two strands of the DNA are twisted together into the famous double helix structure and each nucleotide in a strand is connected to a complementary nucleotide in the other strand, where $A$ is complement of $T$ and $C$ is complement of $G$ and vice versa. It has been believed that a discovery of a DNA structure by Watson and Crick [10] has reshaped a structure of modern biology. Three main areas of interest can be distinguished in the field of DNA: DNA Sequencing, DNA Assembling and DNA mapping. The DNA sequencing problem is to determine a sequence (string) of nucleotides (symbols) drawn from the set $\Sigma = \{A, C, G, T\}$ [5, 4, 3]. Here the input data comes from a biochemical hybridization experiment, and can be viewed as a set (called spectrum) of words (fragments or oligonucleotides). These fragments usually have overlap. The spectrum may contain positive errors, i.e. fragments present in the spectrum but absent in the original sequence, and negative errors, i.e. fragments not present in the spectrum but possible to distinguish in the original sequence. Repetitions of fragments in the sequence are also treated as negative errors. The spectrum without any errors is called the ideal one [1]. The two most popular methods for DNA sequencing are the Sanger method and the Sequencing by Hybridization (SBH) method. The aim is to reconstruct the original DNA sequence of a known length ‘n’ on the basis of these overlapping words.

This paper is organized as follows:

The section 1 includes the introduction and preliminaries. Section 2 includes the motivation of this paper. In Section 3, we have shown the DNA sequencing problem as the shortest common superstring (SCS) problem and its optimization representation. Section 4 includes a short review about the popular method “Sequencing by Hybridization (SBH)” for DNA sequencing. In Section 5, we have proposed a new algorithmic approach to solve SCS problem for ideal spectrum and fragments with constant length. Some suitable examples are also included to justify the algorithm. Section 5 includes the conclusion of this paper.

II. The Motivation

This paper is mainly motivated from the paper [1] On the link between DNA sequencing and graph theory, Computational Methods in Science and Technology, 10(2004), 39-47. by M. Kasprzak and [11] DNA sequencing and the shortest superstring problem, Lecture 15: Computational Biology by S. Mneimneh. Several methods for DNA sequencing problem with constant length oligonucleotide library are drawn in [1] based on the overlapping words, the ones basing on approaches from graph theory.

In this paper, we have proposed a new algorithmic approach for ideal spectrum and fragments with constant length ‘l’ composing the original sequence by means of graph theory.
III. The shortest common superstring (SCS) problem

The DNA sequencing problem can also be stated as the problem of constructing a string over \( \Sigma = \{A, C, G, T\} \) from a given spectrum (not necessarily an ideal spectrum) \( S = \{s_1, s_2, \cdots, s_m\} \), so that the resulting string is the shortest string which contains as many of the fragments in the spectrum as possible. This problem is called shortest common superstring (SCS) problem [13].

3.1. SCS problem as an optimization Problem

If we consider, \( S = \{s_1, s_2, \cdots, s_m\} \) over \( \Sigma = \{A, C, G, T\} \) then Solution: Strings that contains all is of \( S \)

Cost: Length of a string.

Goal: Length is minimum.

Without loss of generality, we assume that \( S = \{s_1, s_2, \cdots, s_m\} \) is factor free, i.e. there are no strings \( s_i, s_j \in S, i \neq j \) such that \( s_i \) is a substring of \( s_j \).

IV. Sequencing by hybridization (SBH)

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 \( \text{spectrum}(s,l) \) [13].

4.1. The SBH Problem

Goal: Reconstruct a string \( s \) from its \( l \)-mer composition.

Input: A set \( S \), representing all \( l \)-mers from an unknown string \( s \).

Output: String \( s \) such that \( \text{spectrum}(s,l) = S \).

The computational complexity of various variants of the problem is already known. The variant with ideal spectrum is polynomially solvable [6] and the variants with error present in the spectrum are all strongly NP-hard [2]. Various methods for DNA sequencing problem with constant length fragment are discussed in [1].

V. An Algorithmic Approach

This algorithmic approach includes the following three steps.

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

\[
V = S \quad \text{(One vertex } v_i \text{ corresponds to one fragment } s_i). \\
E = \{ (v_i, v_j) : (v_i, v_j) \text{ is an ordered pair} \}
\]

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

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 algorithm for obtaining a parenthetical tree \( T \) as follows:

We consider the following sub-steps (1 to 6).

1. Starting vertex (root) = \( v_i \), if \( \Sigma C_i \) is minimum.
2. Ending vertex (leaf) = \( v_e \), if \( \Sigma R_i \) is minimum.
3. \( v_i \) follows \( v_j, s \neq i \), if \( a_{si} \) is maximum in the \( s \)th row.
4. \( v_i \) follows \( v_j \) if \( a_{ij} \) is maximum in the \( i \)th row and so on.
5. Repeat the sub-step 4 until we get the Ending vertex \( v_e \).
6. $SCS = \left( (v_i, v_j) \right) \cdots (v_r)$ provided each $v_i \in V$ is taken for single time only.

7. STOP.

Here, $\Sigma C_{v_i}$ is the sum of the elements in the $v_i$th column, $\Sigma R_{v_i}$ is the sum of the elements in the $v_i$th row and $\left( v_i, v_j \right) = (s_i, s_j) = x w_y , y$.

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 $|s| = L - \Sigma w_y$ (|s| is minimum when $\Sigma |w_y|$ 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 Travelling Salesman Problem (TSP) [12].

We take the following examples (1 to 3) to justify the algorithm. In Example 2, we observe that different spectrum may have same SCS. In Example 3, we observe that same spectrum may generate more than one SCS. Throughout the figures (2 to 6), $X$ gives the path (root to leaf) which violates the algorithm and hence does not give us SCS.

Example 1:
Suppose the original sequence be $s = ACTCTGG$ and an errorless hybridization experiment generates the ideal spectrum: $S = \{ACT, CTC, CTG, TGG, TCT\}$ [11]. Here $n = |s| = 7 , l = |s| = 3$ and $|S| = n - l + 1 = 7 - 3 + 1 = 5$. Suppose the ideal spectrum: $S = \{ACT, CTC, CTG, TGG, TCT\}$ corresponds to the set of vertices: $V = \{v_1, v_2, v_3, v_4, v_5\}$, then $K_s = (V, E, W)$ is the complete graph (figure 1).

![Figure 1.](image)

The overlap matrix:

\[
M = \begin{bmatrix}
0 & 2 & 2 & 1 & 1 & 6 \\
0 & 0 & 1 & 0 & 2 & 3 \\
0 & 0 & 0 & 2 & 0 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 2 & 2 & 1 & 0 & 5 \\
0 & 4 & 5 & 4 & 3
\end{bmatrix}
\]

And parenthetical tree $(T)$:
From figure, the SCS is 
\[ s = \left( \left( \left( v_1 v_2 \right) v_3 \right) v_4 \right) v_5 \] = ACTCTGG
and \[ |s| = L - \sum w_i = 15 - (2 + 2 + 2 + 2) = 7. \]

Example 2:
Suppose ideal spectrums are
\[ S_1 = \{TAT, ATG, TGG, GGT, GTG, TGC\}, S_2 = \{TGG, TGC, TAT, GTG, GGT, ATG\}, \]
\[ S_3 = \{ATG, GGT, GTG, TAT, TGC, TGG\} \] taken from [13] and each \( S_i (i = 1, 2, 3) \) corresponds to the
set of vertices: \( V = \{v_1, v_2, v_3, v_4, v_5, v_6\} \). For \( S_1 \), we have \( K_6 = (V, E, W) \) and the
overlap matrix:

\[
M = \begin{array}{cccccc}
\text{ } & v_1 & v_2 & v_3 & v_4 & v_5 & \text{ } \\
v_1 & 0 & 2 & 1 & 1 & 0 & 1 \\
v_2 & 0 & 0 & 2 & 1 & 1 & 2 \\
v_3 & 0 & 1 & 0 & 2 & 1 & 0 \\
v_4 & 1 & 0 & 1 & 0 & 2 & 1 \\
v_5 & 0 & 2 & 2 & 1 & 0 & 2 \\
v_6 & 0 & 0 & 0 & 0 & 0 & 0 \\
\sum C_{v_i} & 1 & 5 & 6 & 5 & 4 & 6
\end{array}
\]

And parenthetical tree \( (T_1) \):

---

\[ (((v_1 v_2) v_3) v_4) v_5 \]
From figure, the SCS \( s = (((v_1 v_2) v_3) v_4) v_5) v_6) = TATGGTGC. \)

For \( S_2 \), the overlap matrix:

\[
M = \begin{bmatrix}
    v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & \Sigma R_{v_i} \\
    v_1 & 0 & 0 & 0 & 1 & 2 & 0 & 3 \\
    v_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    v_3 & 1 & 1 & 0 & 0 & 0 & 2 & 4 \\
    v_4 & 2 & 2 & 0 & 0 & 1 & 0 & 5 \\
    v_5 & 1 & 1 & 1 & 2 & 0 & 0 & 5 \\
    v_6 & 2 & 2 & 0 & 1 & 1 & 0 & 6 \\
    \Sigma C_{v_i} & 6 & 6 & 1 & 4 & 4 & 2 
\end{bmatrix}
\]

And parenthetical tree \( (T_2) \):

From figure, the SCS \( s = (((v_3 v_6) v_1) v_5) v_4) v_2) = TATGGTGC. \)

For \( S_3 \), the overlap matrix:

\[
M = \begin{bmatrix}
    v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & \Sigma R_{v_i} \\
    v_1 & 0 & 1 & 1 & 0 & 2 & 2 & 6 \\
    v_2 & 0 & 0 & 2 & 1 & 1 & 1 & 5 \\
    v_3 & 0 & 1 & 0 & 0 & 2 & 2 & 5 \\
    v_4 & 2 & 0 & 0 & 0 & 1 & 1 & 4 \\
    v_5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    v_6 & 0 & 2 & 0 & 1 & 0 & 0 & 3 \\
    \Sigma C_{v_i} & 2 & 4 & 3 & 2 & 6 & 6
\end{bmatrix}
\]
And parenthetical tree diagram \( (T_3) \):

From figure, the SCS \( s = \left( \left( \left( \left( v_1v_6 \right) v_2 \right) v_3 \right) v_5 \right) \) = TATGGTGC.

For all \( S_i (i = 1, 2, 3), |S| = L - \sum w_{ij} = 18 - (2 + 2 + 2+ 2+ 2) = 8 \).

**Remark:**
The above Example shows that different spectrum may have same SCS. The following example shows it for \( SCS \ s = TATGGTGC \).

**Example 3:**
Suppose the ideal spectrum: \( S = \{TGG, GCG, TGC, GCA, GGC, ATG, GTG, CGT\} \) \[\text{[1]}\] corresponds to the set of vertices: \( V = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\} \), then \( K_8 = (V, E, W) \) and the overlap matrix:

\[
M = \begin{bmatrix}
v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 & v_8 & \Sigma R_{vi} \\
v_1 & 0 & 1 & 0 & 1 & 2 & 0 & 1 & 0 & 5 \\
v_2 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 2 & 5 \\
v_3 & 0 & 2 & 0 & 2 & 0 & 0 & 0 & 1 & 5 \\
v_4 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
v_5 & 0 & 2 & 0 & 2 & 0 & 0 & 0 & 1 & 5 \\
v_6 & 2 & 1 & 2 & 1 & 1 & 0 & 1 & 0 & 8 \\
v_7 & 2 & 1 & 2 & 1 & 1 & 0 & 0 & 0 & 7 \\
v_8 & 2 & 0 & 1 & 0 & 0 & 0 & 2 & 0 & 4 \\
\Sigma C_{vi} & 5 & 7 & 5 & 8 & 5 & | & 5 & 4 & \end{bmatrix}
\]

And parenthetical tree diagram \( (T) \):
Hence the SCSs are
\[ s = v_6v_1v_2v_3v_4v_5, \quad s = v_6v_1v_2v_3v_4v_5 = ATGGCGTGCA, \quad s = v_6v_1v_2v_3v_4v_5 = ATGCGTGGCA \]

And \[ |s| = L - \sum |v_i| = 24 - (2 + 2 + 2 + 2 + 2 + 2 + 2) = 10 \]

(Note: we have omitted the parenthetic in figure 6 to minimize the diagram in size)

**Remark:**
The above example shows that same spectrum may generate more than one SCS: \[ s = ATGGCGTGCA \] and \[ s = ATGCGTGGCA. \]

**VI. Conclusions**

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. For that, we find a Hamiltonian path in the complete graph with the help of *parenthetical tree* which may be converted to *Travelling Salesman Problem (TSP)*. The proposed approach may have some disadvantages as it cannot guarantee about the uniqueness of the solution, i.e. whether the result covers the original sequence or not. The problem of uniqueness was mentioned in various papers like [6, 7, 8], the surety of the result can be obtained properly with the help of hybridization experiments only.

**References**