SCS AND TSP IN DNA SEQUENCING

PRANAB KALITA*

BICHITRA KALITA**

Abstract:

In this paper, we have taken the problem of DNA sequencing as an optimization problem and also proposed a combinatorial approach to get the original DNA sequence. For that, we consider the path in a weighted graph to maximize the travelling cost in solving the TSP having different intercity cost. This is a Hamiltonian path which gives the optimal solution to the DNA sequencing problem.

Keywords: fragment, spectrum, SCS problem, SBH problem, TSP problem, weighted graph, weighted matrix.

* Department of Mathematics, Gauhati University, Guwahati, Assam, India.

** Department of Computer Applications (M.C.A), Assam Engineering College, Guwahati, Assam, India.

1. Introduction

Deoxyribonucleic acid (DNA) contains the genetic codes that are passed from generation to generation. It consists of two strands connected by hydrogen bonds, each of which contains *nucleotides* from $\sum = \{A, C, G, T\}$ called *alphabet,* where A, C, G, T are for Adenine, Cytosine, Guanine and Thymine respectively. Each nucleotide in a strand is connected to a complementary nucleotide in the other strand, where $A \equiv \bar{T}$, $C \equiv \bar{G}$ and *vice versa*. There are three main areas in the field of DNA: DNA Sequencing, DNA Assembling and DNA mapping. It is said that the discovery of a DNA structure by Watson and Crick [14] has restructured the modern biology. DNA sequencing technologies have been available since the 1970s and are still evolving. Sanger and Gilbert receive the **1980 Nobel Prize** for DNA sequencing methods.

The DNA sequencing problem is to determine a sequence (string) of nucleotides (symbols) drawn from the set $\sum = \{A, C, G, T\}$ [3, 4, 5]. Here the input data can be viewed as a set (called *spectrum*) of words (called *fragments or oligonucleotides*) that comes from a biochemical *hybridization experiment*. These fragments usually have overlap as well as of varying length. A spectrum is said to have *positive errors/negative error*, if fragments present/absent in the spectrum but absent/present in the original sequence. Repetitions of fragments in the sequence are also treated as *negative errors*.

Error occurring during the hybridization experiment shows an important role in reconstructing the original sequence. The computational complexity of various variants of the problem is already known. The variant with no error (*ideal spectrum*) is polynomially solvable [7] and the variants with error present in the spectrum become NP-hard [3]. 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 the overlapping words.

The second section of this paper contains a short review of SBH along with two methods which are mentioned in [2], the ones basing on approaches from graph theory. A short review of DNA

sequencing problem as *shortest common superstring* (***SCS***) problem and its representation as an optimization problem is discussed in the third section. In the forth section, we propose a new approach to solve SCS problem for a spectrum with varying length *fragments* which may be converted to Travelling Salesman Problem (***TSP***) with maximum cost. Finally, the fifth section includes conclusions.

2 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 all l-mers (that is of equal length l) composing the original sequence.

Different sequences may have same spectrum: that is, we may have $s_1 \neq s_2$ such that

$$\text{spectrum}\,(s_1,\ l)\ =\ \text{spectrum}\,(s_2,\ l).$$

In case of no errors in the spectrum S, each fragment intersects with another in exactly $l-1$ positions and the total length $'n'$ of the final sequence can be calculated as $n = |S| + l - 1$.

2.1 Methods

Several methods for DNA sequencing problem with constant length oligonucleotide library are drawn in [2], each of them base on approaches from graph theory.

Method 2.1.1 This method refers to a well-known problem from graph theory mentioned in [13]. In this method, the Hamiltonian path is searched for in a directed graph. We construct a directed graph having each vertex in the graph corresponds to each element of the spectrum. Two vertices u and v are connected by the arc (u,v) if last $l-1$ letters of the label (fragment) of u overlap the first $l-1$ letters of the label of v. This method may correspond to more than one Hamiltonian path which represent the same length sequences (example 1).

Example 1 Suppose a hybridization experiment generates the ideal spectrum S = {ATG, TGG, TGC, GTG, GGC, GCA, GCG, CGT}.The graph from method 2.1.1 is presented in Fig.1.

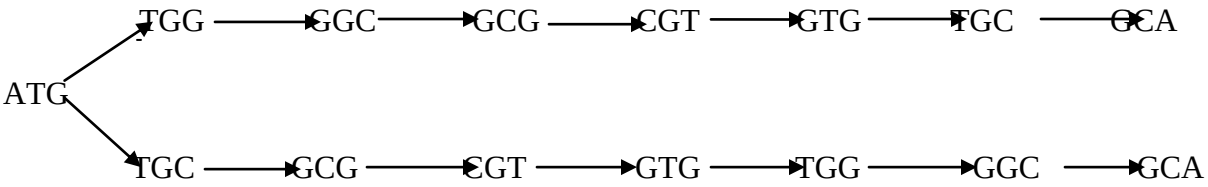

Fig.1. A graph from method 2.1.1.

In this graph we have two Hamiltonian paths:

Path 1: (ATG-TGG-GGC-GCG-CGT-GTG-TGC-GCA) which results ATGCGTGGCA and Path 2: (ATG-TGC-GCG-CGT-GTG-TGG-GGC-GCA) which results ATGGCGTGCA.

The length of the final sequence can be obtained as

$$\begin{aligned} n &= |S| + l - 1 \\ &= 8 + 3 - 1 \\ &= 10 \end{aligned}$$

This method accept ideal spectrum as input data and leads to an exponential-time algorithm. The first and only polynomial-time algorithm in solving constant length ideal spectrum was presented in [10]. In this method 2.1.2 (below), the Eulerian path is searched for in a directed graph based on the spectrum.

Method 2.1.2 Suppose the ideal spectrum S has l-mer fragments. We build a directed graph for S as follows: Vertices corresponds to all $(l-1)$ mers/ $(l-1)$ tuples. For each l-mer in spectrum add edge from vertex representing first $(l-1)$ characters to vertex representing last $(l-1)$ characters i.e. edges correspond to the fragments in S. A fragment $a\omega b$ (where a, b are letters) forms an edge from the prefix node $a\omega$ to the suffix node ωb. This method may also correspond to more than one Eulerian path which represents the same length sequences (example 2).

Example 2 Suppose a hybridization experiment generates the ideal spectrum S = {ATG, TGG, TGC, GTG, GGC, GCA, GCG, CGT} as in example 1. The graph from method 2.1.2 is presented in Fig.2.

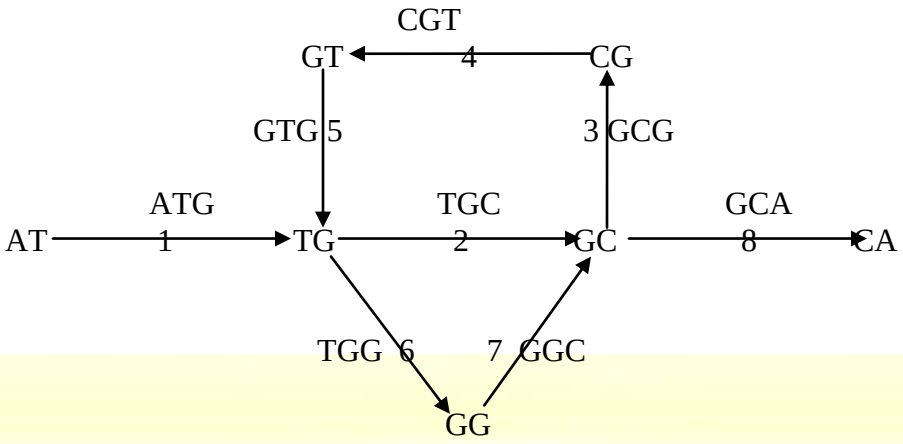

Fig.2.a A graph obtained from method 2.1.2 and solution is ATGCGTGGCA.

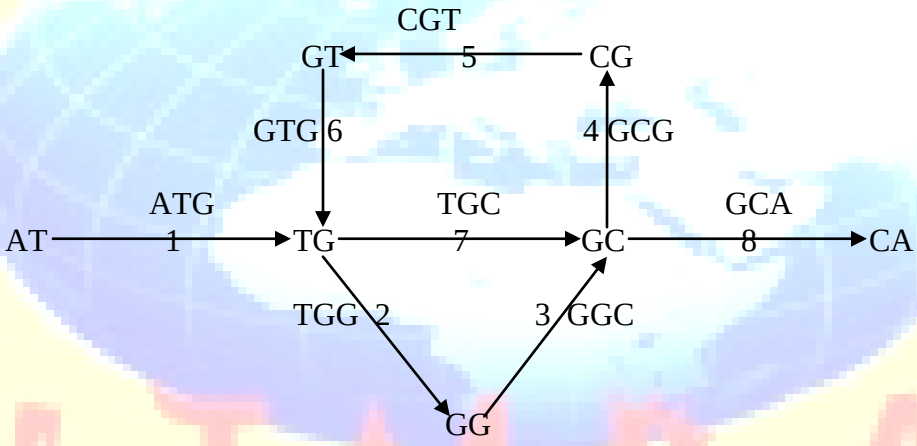

Fig.2.b A graph obtained from method 2.1.2 and solution is ATGGCGTGCA.

In this graph, we have two Eulerian paths:

Path 1: (1-2-3-4-5-6-7-8) = (AT.TG.GC.CG.GT.TG.GG.GC.CA) which results

ATGCGTGGCA

and

Path 1: (1-2-3-4-5-6-7-8) = (AT.TG.GG.GC.CG.GT.TG.GC.CA) which results

ATGGCGTGCA.

The length of the final sequence can be obtained as

$$n = |S| + l - 1$$
$$= 8 + 3 - 1$$
$$= 10$$

This method reduces the complexity of the algorithm in solving the DNA sequencing problem because finding an Eulerian path can be done in polynomial time.

3 DNA sequencing problem as *SCS* problem

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*) $S = \{s_1, s_2, s_3, \cdots, s_n\}$, 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 the *shortest common superstring* (***SCS***) problem.

3.1 *SCS* problem as an optimization Problem

If we consider, $S = \{s_1, s_2, s_3, \cdots, s_n\}$ over $\sum = \{A, C, G, T\}$ then

Solution: Strings that contains all s_1 of S.

Cost: Length of a string.

Goal: Length is minimum.

Without loss of generality we assume that $S = \{s_1, s_2, s_3, \cdots, s_n\}$ 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.

4 *SCS* problem to solve *TSP*

The Travelling Salesman Problem (***TSP***) with maximum cost can be explained as follows: Suppose a traveller salesman wants to travel among n cities in such a way that:

1. He visits all the cities only once.
2. He visits each of the cities only once.
3. He starts from the city C_1 and ends in the city C_n.

4. He wants to maximize his total cost.

Considering these conditions, we may relate the DNA sequencing problem and the *TSP* along with graph theory using the following approach:

4.1 A NEW APPROACH

This approach includes three steps.

Step I: For (given) a spectrum we define a complete weighted graph $K=(V,E,W)$ where,

$V=C$ (One vertex per city C_i and each city C_i is assigned with a *fragment* s_i).

$E=$ Edges between all vertices (a complete graph).

$$W(C_iC_j)=\text{overlap}(s_i,s_j)=\begin{cases}|w_{ij}| & \text{if} \quad i\neq j\\ 0 & \text{if} \quad i=j\end{cases} \quad \text{where} \quad s_i=xw_{ij}, s_j=w_{ij}y$$

In $K=(V,E,W)$, every Hamiltonian path determines a superstring and *vice versa*. Now, The length (L) of SCS (defined by the Hamiltonian path) = total length of the strings (T) – sum of the weights of the edges (W), i.e.

$$|SCS|=\sum|s_i|-\sum|w_{ij}|$$
$$\Rightarrow L=T-W$$
$$\Rightarrow L_{\min}=T-W_{\max} \qquad \text{..... (1)}$$

Step II: To find $W_{\max}$, we define an weighted matrix $M=(a_{ij})$ where $a_{ij}=W(C_iC_j)$.

Step III: Using the matrix $M=(a_{ij})$, we develop an *algorithm* having the following steps(1 to 8).

1. Find the entry a_{ij} with maximum value in the matrix $M=(a_{ij})$.
2. Merge the two vertices C_i and C_j into a single vertex C_iC_j and also merge the corresponding strings s_i and s_j into a single string $s_is_j=xw_{ij}y$.
3. $W_{\max}=\sum a_{ij}$.
4. Construct the matrix $M=(a_{ij})$ for the graph with vertices $C_1,C_2,\cdots,C_iC_j,\cdots,C_n$ as in step 2.
5. Find the size $n\times n$ of the matrix M.

6. If the size of the matrix M is $n \times n = 1 \times 1$, go to step 7, otherwise go to step II.
7. The maximum cost is $W_{\max}$ and the $SCS = s_1 \cdots s_i s_j \cdots s_n$ for the path $C_1 \cdots C_i C_j \cdots C_n$.
8. STOP.

The step 7 gives the optimal value $W_{\max}$ and the optimal path (Hamiltonian path) $C_1 \cdots C_i C_j \cdots C_n$ for the travelling salesman problem with maximum cost. The path $C_1 \cdots C_i C_j \cdots C_n$ results the shortest common superstring $SCS = s_1 \cdots s_i s_j \cdots s_n$. Since T is the total length of the strings which is fixed by the problem, hence constant for all Hamiltonian path and hence from (1), we obtain the length of the *SCS*.

Example 3: Suppose a hybridization experiment generates the spectrum $S = \{CATGC, CTAAGT, GCTA, TTCA, ATGCATC\}$ with varying length. We need to find the *SCS* having these fragments: Assume that the vertex set $V = \{C_1, C_2, C_3, C_4, C_5\}$ where C_i represents a city, is labelled by the respective element of the spectrum $S = \{CATGC, CTAAGT, GCTA, TTCA, ATGCATC\}$. Now, using the algorithm, we have the complete weighted graph K_5:

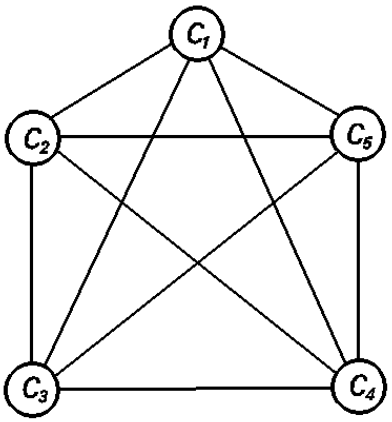

Fig. 3 K_5

The weighted matrix for K_5 is

$M =$

	C_1	C_2	C_3	C_4	C_5
C_1	0	1	2	0	4
C_2	0	0	0	1	0
C_3	0	3	0	0	1
C_4	2	0	0	0	0
C_5	3	1	0	0	0

Next, the vertex set becomes $V = \{C_1C_5, C_2, C_3, C_4\}$ which is labelled by the spectrum $S = \{CATGCATC, CTAAGT, GCTA, TTCA\}$.

For that, we have the complete weighted graph K_4:

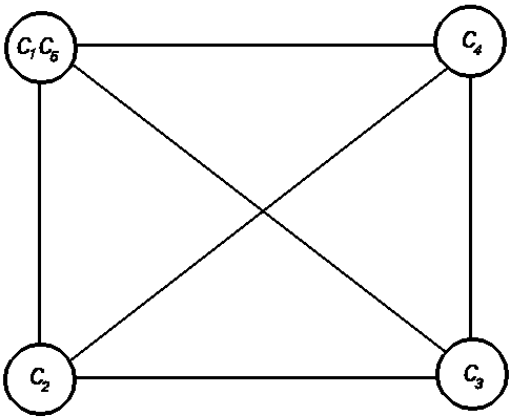

Fig. 4 K_4

The weighted matrix K_4

$M =$

	C_1C_5	C_2	C_3	C_4
C_1C_5	0	1	0	0
C_2	0	0	0	1
C_3	0	$\boxed{3}$	0	0
C_4	2	0	0	0

Next, the vertex set becomes $V = \{C_1C_5, C_4, C_3C_2\}$ which is labelled by the spectrum $S = \{CATGCATC, TTCA, GCTAAGT\}$. For that, we have the complete weighted graph K_3:

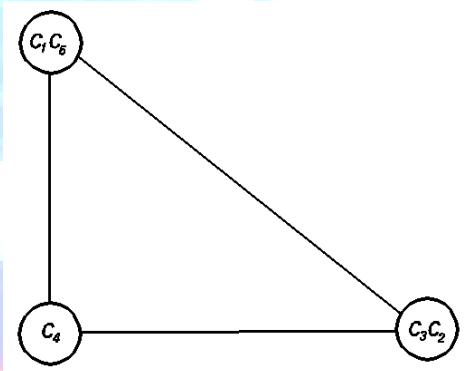

Fig. 5 K_3

The weighted matrix K_3

$M =$

	C_1C_5	C_3C_2	C_4
C_1C_5	0	1	0
C_3C_2	0	0	1
C_4	$\boxed{2}$	0	0

Next, the vertex set becomes $V = \{C_4C_1C_5, C_3C_2\}$ which is labelled by the spectrum $S = \{TTCATGCATC, GCTAAGT\}$. For that, we have the complete weighted graph K_2:

Fig. 6 K_2

The weighted matrix K_2

	$C_4C_1C_5$	C_3C_2
$C_4C_1C_5$	0	0
C_3C_2	$\boxed{1}$	0

Thus the optimal path (Hamiltonian path) is $C_3C_2C_4C_1C_5$ for the travelling salesman problem with the optimal value (maximum cost) $W_{\max} = 4+3+2+1=10$. The path $C_3C_2C_4C_1C_5$ results the shortest common superstring $SCS = GCTAAGTTCATGCATC$.

Here $T = 26$ and hence from (1), we obtain the length of the *SCS*.

$$L_{\min} = T - W_{\max}$$
$$= 26 - 10$$
$$= 16$$

Thus, a solution to the SCS problem translates into finding a Hamiltonian path of maximum weight. Also the total length of the strings is fixed by the problem, hence constant for all Hamiltonian paths and therefore it has been converted to Travelling Salesman Problem (***TSP***).

5 CONCLUSIONS

There are also other methods for DNA sequencing by hybridization developed by various researchers taking positive or negative spectrum with constant or variable length fragments. In this paper, the aim of the proposed method is searching for a path with maximum weight in the weighted graph. The path can be easily translated to a DNA sequence and is also viewed via *shortest common superstring* (***SCS***) problem by finding a Hamiltonian path which may be converted to *Travelling Salesman Problem* (***TSP***) with maximum cost. The proposed method has some disadvantages as it cannot tell about the uniqueness of the solution, i.e. whether the obtained result covers the original sequence or not. This problem of the uniqueness of the solution was disclosed earlier in a number of papers like [7, 8, 9].

References

[1]K. Mehdizadeh, M.A. Nekoui, K. Sabahi, and A. Akbarimajd, A modified DNA-Computing Algorithm To Solve TSP, 1-4244-9713-4/06/$20.00/2006 IEEE.

[2] Marta Kasprzak, *On The Link Between DNA Sequencing and Graph Theory*, Computational Methods in Science and Technology **10**, 39-47 (2004).

[3] J. Blazewicz and M. Kasprzak, *complexity of DNA sequencing by hybridization*, Theorical computer science **290**, 1459-1473 (2003).

[4] P.A. Pevzner, *Computational Molecular Biology: an Algorithmic Approach*, MIT Press, Cambridge (2000).

[5] J. Setubal and J. *Meidanis, Introduction to Computational Molecular Biology*, PWS Publishing Company, Boston (1997).

[6] M.S. Waterman, *Introduction to Computational Biology, Maps, Sequences and Genomes*, Chapman & Hall, London (1995).

[7] M. Dyer, A. Frieze, and S. Suen, *The probability of unique solution of sequencing by hybridization*, Journal of Computational Biology **1**, 105-110 (1994).

[8] P.A. Pevzner and R.J. Lipshutz, *Towards DNA sequencing chips*, Lecture Notes in Computer Science **841**, 143-158 (1994).

[9] E.M. Southern, U. Maskos, and J. K. Elder, *Analyzing and comparing nucleic acid sequences by hybridization to array of oligonucleotides: Evaluation using experimental models*, Genomics **13**, 1008-1017(1992).

[10] P.A. Pevzner, *l-tuple DNA squencing: computer analysis*, Journal of Biomolecular, Structure and Dynamics **7**, 63-73 (1989).

[11] R. Drmanac, I. Labat, I. Brukner, and R. Crkvenjakov, *Sequencing of megabase plus DNA by hybridization: theory of the method*, Genomics **4**, 114-128 (1989).

[12] W. Bains and G.C. Smith, *A novel method for nucleic acid sequence determination*, Journal of Theoretical Biology **135**, 303-307 (1988).

[13] Yu. P. Lysov, V. L. Florentiev, A. A. Khorlin, K. R. Khrapko, V. V. Shik, and A. D. Mirzabekov, *Determination of the nucleotide sequence of DNA using hybridization with oligonucleotides. A new method*, Doklady Akademii Nauk SSSR **303**, 1508-1511 (1988).

[14] J.D. Watson, F.H.C. Crick, *A structure for deoxyribose nucleic Acid*, Nature **173** (1953) 737-738.

[15] Saad Mneimneh, *Lecture 15: DNA sequencing and the shortest superstring problem*, Computational Biology.

(Retrived from www.cs.hunter.cuny.edu/~saad/courses/.../lectures/lecture15.pdf).

[16] Chapter 3: Reconstructing DNA

(Retrived from www.liacs.nl/~hoogeboo/mcb/mapp.pdf).

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

(Retrived from www.time.mk/trajkovski/teaching/eurm/bio/lecture3.pdf).