ALGORITHMS TO RECONSTRUCT THE TARGET DNA FROM ITS SPECTRUM CONNECTED AT SOME LEVEL

F. X. Wu* **and W. J. Zhang**
Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK, S7N 5A9

Abstract: In order to sequence a target DNA, it is first cleaved into many shorter overlapping fragments by restriction enzymes. Then each such a short fragment is identified as a letter string over the alphabet $\{A,C,G,T\}$ called a read fragment. We call the set of all read fragments which covers the target DNA a spectrum. It is believe that the shortest superstring of its spectrum outlines very well the target DNA. Unfortunately, the problem of finding the shortest superstring for any given set of strings S is NP-hard. However, in the biologically meaningful cases, the problem needn't be so hard. An observation is that it is not convincible that two read fragments consisting of several hundred letters, which come from consecutive locations on the target DNA, have only overlap of several letters. From this observation, one may reasonably assume that strings in the spectrum have enough overlap (connectivity). A class of important instances satisfying this assumption are those whose spectrum is from DNA array. Based on this assumption and another about repeat, the main result in the presented paper is: if the spectrum S of a target DNA is substring-free and connected at level t, and the target DNA has no repeats of size t or larger, then there exist an algorithm to reconstruct the target DNA in $O(|S|)$.

1. Introduction

The reconstruction of the target DNA from its read fragments is also called fragment assembly of DNA. In the context here, A DNA molecule is a very long polymer chain consisting typically of hundreds of thousands of letters over the alphabet $\{A,C,G,T\}$. At present, techniques and conditions only allow to identify about 500 bases out of a long single-stranded DNA molecule in an experiment. Therefore, reconstructing DNA is an important step to sequence a whole DNA molecule. It is biochemists' belief that the shortest superstring of its spectrum is the most reasonable outline of the target DNA. Such a shortest superstring representation is likely to be correct if the target DNA sequence does not have many long repeats (which is often the case in practice).

Since the problem is NP-hard a lot of effort has been taken to find good approximation algorithms with guaranteed performance [1-4, 7] or to find efficient and exact algorithms for some restricted problems [5-7]. In this paper, we try to develop an efficient and exact algorithm for a class of restricted but definitely biological problem. Our algorithm improves Setubal and Meideanis's one in [6] and may cover Pevzer's algorithm to some extent. We recall some basic definition and facts in section 2. In section3, we first review two existing algorithms related to ours, then present our algorithm. Finally we give some discussions and future works in section 4.

2. Preliminaries

Let $S = \{s_1, s_2, \cdots, s_m\}$ be a set of strings over some alphabet $\{A,C,G,T\}$. A common superstring, or simply superstring, of S is a string s such that each $s_i (i = 1,2,\cdots,m)$ in S is a substring of s. The shortest superstring problem is to find a superstring of the smallest possible length for any given set of strings S. The central concept of most existing algorithms for this problem is *distance gr*aph or *overlap* graph of the spectrum. Without loss of generality, assume that the set S is "*substring-free*" in that no string $s_i \in S$ is a substring of any other $s_i \in S$. For two string u and v, let y be the longest string such that $u = xy$ and $v = yz$ for some nonempty strings x and z. The string y is called the *overlap* of the ordered pair u and v, and denoted by $ov(u,v)$, and $|ov(u,v)| = |y|$ is denoted by $lov(u,v)$ (called the overlap function). The string x is called the *prefix* of u with respect to v, and denoted by $pref(u,v)$. Finally $|pref(u,v)| = |x|$ is called the *distance* from u to v, and denoted by $d(u,v)$ (called distance function). After these concepts, the distance graph of S is defined as a weighted digraph

* The corresponding author, faw341@mail.usask.ca

$G_S = (V, E, w)$, where the set of vertices $V = \{s_1, s_2, \cdots, s_m\}$, the set of edges $E = \{(s_i, s_j) \mid 1 \le i \ne j \le m\}$, and the weight function w is the distance function $d(,)$; and the overlap graph H_S is defined as a weighted digraph $H_S = (V, E, w)$, where the sets of vertices and edges in H_S are the same as those in G_S, but the weight function w in H_S is the overlap function $lov(s,t)$.

It was proved [6] that a superstring of S corresponds to a Hamiltonian path in the distance graph G_S or in the overlap graph H_S, and vice verse. Furthermore, either the shortest Hamiltonian path in G_S or the longest Hamiltonian path in H_S corresponds to the shortest superstring. Given a nonnegative integer t, we define a sub-graph $H_S(t)$ of H_S as the weighted digraph keeping only those edges of weight at least t in H_S. For a directed path $P = s_{i_1} \circ s_{i_2}, \cdots, s_{i_{r-1}} \circ s_{i_r}$ in the overlap graph H_S, we denote the string $pref(s_{i_1}, s_{i_s})pref(s_{i_2}, s_{i_3}) \cdots pref(s_{i_{r-1}}, s_{i_r})s_{i_r}$ by $s(P)$. Obviously, $s(P)$ is the shortest superstring generated along the path P. Finally we denote string $pref(u,v)v$ by $< u, v >$.

3. Algorithms

3.1 Two existing algorithms

Pevzner [5] presented an algorithm to reconstruct the target DNA whose spectrum come from DNA array consisting of all short known DNA fragments with length l called probes (its number of fragments is l^4). This kind of spectrums possesses the following properties: 1) All strings have the same length l, and 2) Any two consecutive l-length strings in the target DNA overlaps by $l-1$ letters. In addition, Pevzner assumed that each l-length string appears only once in the target DNA. Exploiting these properties and assumption, he reduce reconstructing the DNA to finding Eulerian path on a digraph F_s, whose set of vertices is the set of all $(l-1)$-tuples, and whose set of edge is the spectrum of the target. Although finding Eulerian path is simple, reconstructing the target DNA become complicate because Eulerian path is not unique when degree of vertices in F_s is bigger. Gusfield et. al proved in [7] that the optimal problem is NP-hard when degree of vertices in F_s is greater than 3.

In [6], Setubal and Meideanis viewed the target DNA as an integer interval. A sampling $\mathbf{A}$ of the target DNA consists of some sub-intervals of the interval. Two intervals α and β in $\mathbf{A}$ are said to be *linked at level* t if $|\alpha \cap \beta| \ge t$. $\mathbf{A}$ is said to be *connected at level* t if for every two intervals α and β in $\mathbf{A}$ there are a serials of intervals α_i for $0 \le i \le l$ such that $\alpha = \alpha_0$, $\beta = \alpha_l$ and α_i is linked to α_{i+1} at level t for $0 \le i \le l-1$. By using the properties of interval graph, Setubal and Meideanis obtained:

Thereom 1: Let $\mathbf{A}$ be a subinterval-free, connected at level t sampling of the target DNA s that covers s. If its spectrum S was generated by $\mathbf{A}$, and s has no repeats of size t or larger, then the digraph $H_S(t)$ has a unique Hamiltonian path P and $s(P) = s$.

3.2 Improved Algorithm

There is an unreasonable restriction on Setubal-Meidanis algorithm: In theorem 1, the spectrum S must be generated by a sampling $\mathbf{A}$ of the intervals that covers interval s. We can call S to be interval-dependent. This means that locations of all intervals in $\mathbf{A}$ were known before s is constructed by S. There are some controversies here. In this section, we try to improve theorem 1 such that it holds for interval-independent string collection.

We say that two read fragments s_i and s_j are *linked at level* t if $lov(s_i, s_j) \geq t$. The spectrum S is said to be *connected at level* t if for every two strings s_i and s_j in S there is either a directed path from s_i to s_j or one from s_j to s_i in H_S on which two consecutive fragments (vertices) are linked at level t. The connected at level t spectrum describes well the assumptions that the two read fragments from consecutive locations in s must have enough overlap and all read fragments cover the target DNA. It is obviously that these two assumptions are biologically reasonably. First we describe our main result as follows:

***Theorem 2*:** Assume that the spectrum S of the target DNA s is substring-free and connected at level t. If s has no repeats of size t or larger, then the digraph $H_S(t)$ has a unique Hamiltonian path P and $S(P) = s$.

Comparing theorem 1 with theorem 2, it is obviously that we cancelled the restriction in theorem 1 that the spectrum S must be generated by a sampling $\mathbf{A}$ of the intervals that covers interval s. Since the spectrum S is substring-free, the target DNA s is the shortest superstring of a Hamiltonian path in the digraph H_S. To prove this theorem, we need the following Lemmas.

Lemma 1: If s has no repeats of size t or larger, then there is no directed cycle in $H_S(t)$, that is, $H_S(t)$ is an acyclic graph.

Proof: By contradiction. Let $s_{i_1} \to s_{i_2} \to \cdots \to s_{i_l} \to s_{i_1}$ be a directed cycle in $H_S(t)$, so it is also a directed cycle in H_S. Since s corresponds the Hamiltonian path H_S, there exists some $k\ (1 \leq k \leq l)$ such that either string $< s_{i_{k-1}}, s_{i_k} >$ or string $< s_{i_k}, s_{i_{k+1}} >$, say $< s_{i_k}, s_{i_{k+1}} >$, doesn't appear in s. Thus the string $ov(s_{i_k}, s_{i_{k+1}})$ appears in two distinct locations in s. This means that there is a repeat $ov(s_{i_k}, s_{i_{k+1}})$ of size t or larger in s, so we get a contradiction. Therefore, we conclude Lemma 1.

Lemma 2: If s has no repeats of size t or larger, then every vertex in $H_S(t)$ has both in-degree at most one and out-degree at most one.

Proof: We first prove by contradiction that every vertex in $H_S(t)$ has in-degree at most one under the condition of Lemma. Assume that there exists some vertex s_i with in-degree at least two. Thus there exist two edges (s_j, s_i) and (s_k, s_i) in $H_S(t)$. Similarly to the argument in Lemma 1, out of two string $< s_j, s_i >$ and $< s_k, s_i >$, at least one, say $< s_j, s_i >$, is not in s, so string $ov(s_j, s_i)$ is a repeat whose size is t or larger. Thus we can obtain that the first conclusion. Similarly, we may prove that every vertex in $H_S(t)$ has out-degree at most one.

Lemma 3: If S is substring-free and connected at level $\boldsymbol{t}$, and if s has no repeats of size t or larger, then in $H_S(t)$ there exist exactly one vertex with in-degree zero and exactly one distinct vertex with out-degree zero. And the other vertices in $H_S(t)$ have exactly in-degree one and out-degree one.

Proof: First we notice that digraph $H_S(t)$ is connected since S is substring-free and connected at level t. Therefore every vertex in $H_S(t)$ has at least either in-degree one or out-degree one because otherwise there exist some isolated vertex. If there are two vertex s_i and s_j with in-degree zero, then there is neither a path from s_i to s_j nor one from s_j to s_i in $H_S(t)$. This contradicts with that S is substring-free and connected at level $\boldsymbol{t}$. Therefore there is at most one vertex with in-degree zero in $H_S(t)$. Furthermore, if there no vertex with in-degree zero, this means that all vertices have in-degree one by lemma 2. Thus there is a circle in $H_S(t)$, it contradicts with lemma 1. So there is exactly one vertex with in-degree zero in $H_S(t)$.

Similarly, we can obtain that there is exactly one vertex with out-degree zero in $H_S(t)$. Finally we can conclude that the other vertices in $H_S(t)$ have exactly in-degree one and out-degree one by employing lemma 2.

Proof of theorem 2 : By lemma 3, in the digraph $H_S(t)$ exactly two vertices are semi-balanced and all other vertices are balanced, so there exist a Eulerian path P in $H_S(t)$. Furthermore, every balanced vertex has exactly in-degree one and out-degree one and the two semi-balanced vertices have exactly in-degree one or out-degree one,

respectively, so this Eulerian path P is unique and visits all vertices in $H_S(t)$ exactly once. Therefore, the path P is also a unique Hamiltonian path in the digraph $H_S(t)$.

For simplicity, let $P = s_1 \square s_2, \cdots, \square s_{|S|}$. We prove that $S(P) = s$ by contradiction. Assume that $S(P) \neq s$, there exists an edge $e = s_i \square s_j$ on path P that doesn't occur on the Hamiltonian path corresponding s in H_S. This means that the string $ov(s_i, s_j)$ occurs two times in s. We have got a contradiction since $|ov(s_i, s_j)| \geq t$. Therefore, we have completed the proof of theorem 2.

Since for a given digraph finding a Eulerian path in it can be archived in the linear time $O(|V|)$ [8], where V is the set of vertex of the digraph, we conclude by theorem 2 and its proof:

Theorem 3 Assume that the spectrum S of the target DNA s is substring-free and connected at level t. If s has no repeats of size t or larger, then there exists an algorithm that can exact reconstruct s from its spectrum S in linear time $O(|S|)$ when the digraph $H_S(t)$ is given.

4. Some discussions and future works

Comparing Setubal and Meideanis's algorithm to our algorithm, it is obvious that we have improved the former in that the interval -dependent condition of the spectrum S is removed. Therefore, our algorithm is more useful. On the other hand, although Pevzner's algorithm can efficiently find a Eulerian path in the digraph F_s which can be candidate target DNAs, it faces difficulties when degree of vertices in the digraph is great than one. When degree of all vertices in the digraph F_s is equal to one, our algorithm is more useful than Pevzner's algorithm in that our algorithm may be applied to those problems whose spectrum need not come from DNA array.

When degree of vertices in F_s is greater than one, Pevzner's algorithm will find a number of the candidate target DNA strings. To choose out one string closest to the target DNA from them, the problem is reduced to find the optimal Euler path in F_s by using some additional experimental information. In a recent paper [7], authors proved that if the degree of every vertex in F_s is exactly two, there exists an algorithm, which find an optimal target DNA in the certain meaning in linear time, and that if the degree of vertex in F_s is equal to or larger than four, finding optimal target DNA is NP-hard again.

The structure of repeats in the target DNA plays a crucial rule in designing algorithms to reconstruct it from its spectrum. Our algorithm here requires no repeats with some size in the target DNA. Although this condition seems to be somewhat harsh, algorithm may provide some indications when one design algorithm to solve some practical problems. One future work is to design exact or good approximate algorithms to solve those problems in which the structure of repeats in the target DNA is known. This information may often be obtained before the target DNA is reconstructed

References

1. Blum A., Jiang T., Li M., Tromp J., and Yannakakis M, Limear approximation of shortest superstring, Journal of the Association for Computing Machinery, 41(4): 630-647, 1994.
2. Breslauer D., Jiang T., and Jiang Z., Rotations of periodic strings and short superstring, Journal of Algorithms, 24: 340-353, 1997.
3. Armen C. and Stein C., A $2\frac{2}{3}$ superstring approximation algorithm, Discrete Applied Mathematics, 88:29-57, 1998.
4. Sweedyk Z., A $2\frac{1}{2}$ - approximation algorithm for shortest superstring, SIAM J. Comput., 29(3), 954-986
5. Pevzer P. A., Computational Molecular Biology: An Algorithm Approach, The MIT Press, 65-82, 2000.
6. Setubal J. and Meidanis J. Introduction to Computational Molecular Biology, International Thomson Publishing, 105-141, 1997.
7. Gusfield D., Karp R., Wang L., and Stelling P., Graph traversals, genes and matroids: an efficient case of the traveling salesman problem, Discrete Applied Mathematics, 88:167-180, 1998.
8. Fleischner H., Eulerian Graphs and Related Topics, Elsevier Scince Piblishers, 1990.