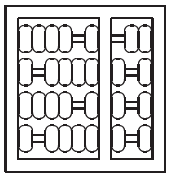

INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

Interval graphs with repeats and the DNA fragment assembly problem

J. Meidanis *P. Takaki*

Technical Report - IC-02-014 - Relatório Técnico

November - 2002 - Novembro

Interval graphs with repeats and the DNA fragment assembly problem

Joao Meidanis* Patricia Takaki†

Abstract

The DNA fragment assembly problem is described as follows. There is a DNA sequence $\texttt{Seq} : [1, L] \mapsto \{A, C, G, T\}$ which is unknown. Fragments are taken from Seq and their sequence is directly determined by sequencing machines. Each fragment corresponds to an interval $[i, j] \subseteq [1, L]$ and the sequencing machine outputs the substring $\texttt{Seq}[i, j] = \texttt{Seq}(i)\texttt{Seq}(i + 1) \ldots \texttt{Seq}(j)$. Typically the size of interval $[i, j]$, which is $j - i + 1$, lies in the range 500 to 1000, while L is much larger; in some cases like, $L = 50 \times 10^3$, but it may reach 3×10^9[Pev00, Gus97]. Given a number of fragments we want to determine the string Seq. Interval Graphs have been used to model this problem, but they do not account for repeats in the sequence Seq. In this note we introduce a new formalism, Interval Graphs with Repeats, to address this issue.

1 Introduction

The DNA fragment assembly problem is studied in many books [SM97, Pev00, Gus97, Wat95, Cer00, Li01] and can be described as follows. There is a DNA sequence $\texttt{Seq} : [1, L] \mapsto \{A, C, G, T\}$ which is unknown. Fragments are taken from Seq and their sequence is directly determined by sequencing machines. Each fragment corresponds to an interval $[i, j] \subseteq [1, L]$ and the sequencing machine outputs the substring $\texttt{Seq}[i, j] = \texttt{Seq}(i)\texttt{Seq}(i + 1) \ldots \texttt{Seq}(j)$. Typically the size of interval $[i, j]$, which is $j - i + 1$, lies in the range 500 to 1000, while L is much larger; in some cases like, $L = 50 \times 10^3$, but it may reach 3×10^9 [Pev00, Gus97]. In 1980 the sequencing of a 5,386-nucleotide virus received a Nobel Prize [Pev00, page 60]. In 1989 the Human Genome Project was launched and it is aimed at determining the approximately 100,000 human genes[1] that comprise the entire 3 billion nucleotides of the human genome[2]. Given a number of fragments we want to determine the string Seq. Interval Graphs have been used to model this problem, but they do not account for repeats in the sequence Seq.

*Institute of Computing, University of Campinas, 13081-970 Campinas, SP. Research supported in part by CNPq

†Institute of Computing, University of Campinas, 13081-970 Campinas, SP. Research supported in part by CNPq

[1] Nowadays this number is thought to be much smaller.

[2] The full sequence of the yeast genome *Saccharomyces cerevisiae* (length around 12 million base pairs) was released in spring 1996, the genome of *E. coli* (length 4.7 million base pairs) was finished on January 16, 1997.

The Shortest Superstring Problem is an overly simplified abstraction that does not capture the real fragment assembly problem, since it assumes perfect data and may collapse DNA repeats. The human genome contains many repeats; for example, a 300 bp *Alu* sequence is repeated (with 5-15 % variation) about a million times in the human genome [Pev00, EG01].

In this note we introduce a new formalism, Interval Graphs with Repeats, to address this issue.

2 The Fragment Assembly Problem

The information we have to assembly the fragments is the **overlap** between them. We say that two sequences s e t overlap by k characters when the last k characters of s are equal to the first k characters of t. We write $s \xrightarrow{k} t$ to indicate this fact. If the intervals corresponding to two fragments have a nonempty intersection, then the fragments overlap. But the converse is not necessarily true. Assuming a uniform distribution for `Seq`, however, we can say that, for larger values of k, the converse is statistically very likely true.

Let us use the notation $[i, j] \xrightarrow{k} [p, q]$ for an integer $k > 0$ when $i \leq p$, $p + k - 1 = j$, and $j \leq q$.

Theorem 1 *Let* $\mathtt{Seq} : [1, L] \mapsto \{A, C, G, T\}$ *be a uniformly distributed sequence and* $[i, j]$ *and* $[p, q]$ *intervals contained in* $[1, L]$. *Then:*

a) if $[i, j] \xrightarrow{k} [p, q]$ *then*

$$\mathtt{Seq}[i, j] \xrightarrow{k} \mathtt{Seq}[p, q]$$

b) if $[i, j] \cap [p, q] = \emptyset$ *then*

$$\Pr(\mathtt{Seq}[i, j] \xrightarrow{k} \mathtt{Seq}[p, q]) = \frac{1}{4^k}$$

provided that $j - i + 1 \geq k$ *and* $q - p + 1 \geq k$ *(otherwise the probability is zero).*

Proof: Part a) is immediate from the definitions. Part b) comes from the fact that `Seq` is uniformly distributed. If $[i, j] \cap [p, q] = \emptyset$, then each character of $[i, j]$ is chosen independently of the characters in $[p, q]$. If they agree on k positions, this happens with a chance of $\frac{1}{4}$ per position, or $(\frac{1}{4})^k$ overall probability. □

Remark In fact, b) can be generalized as:

b') if $[i, j] \not\xrightarrow{k} [p, q]$ then

$$\Pr(\mathtt{Seq}[i, j] \xrightarrow{k} \mathtt{Seq}[p, q]) = \frac{1}{4^k},$$

provided that $j - i + 1 \geq k$ and $q - p + 1 \geq k$.

The idea is to use overlap information to find the relative position of fragments. For that we need to choose a reasonably large value of k. What is a "reasonably large" value can be assessed by the following considerations.

An **interval model** over $[1, L]$ is a set I of intervals $[i, j] \subseteq [1, L]$. A **proper** interval model is an interval model where no interval is contained in another [Gol80, Har72]. From a proper interval model I we can construct its intersection graph which is of course an interval graph. In addition, given an integer $k \geq 1$, we can construct $G_k(I)$, the **interval graph of I with tolerance** k, as follows. For each $[i, j] \in I$ consider $[i, j - k]$, and take the intersection graph of those shortened intervals. Of course, this can only be done if $k \leq j - i + 1$, but in practice this always holds.

Given a set F of sequences we can construct the **graph of k-overlaps** of F as follows: the vertices are the sequences of F and two sequences s and t are adjacent when $s \xrightarrow{l} t$ for some $l \geq k$. We call this graph $O_k(F)$.

Theorem 2 *Let I be a proper interval model over $[1, L]$. Let F be the set $\{\texttt{Seq}[i, j] \mid [i, j] \in I\}$. For an integer $k \geq 1$ construct $G_k(I) = G$ and $H = O_k(F)$. Then:*

a) $G_k(I)$ is a generating subgraph of $O_k(F)$.

b) assuming a uniform distribution for `Seq`*, we have*

$$\Pr(G_k(I) = O_k(F)) \geq 1 - \frac{n^2}{4^k},$$

were $n = |I|$.

Proof: To prove Part a) we reason as follows. The vertices of $G_k(I)$ and $O_k(F)$ are in one-to-one correspondence, since both vertex sets correspond also to the intervals in I. If uv is an edge of $G_k(I)$, then u and v intersect. Since I is proper, we have, without loss of generality, that $u \xrightarrow{l} v$ for some $l \geq k$ and by Theorem 1, Part a) we conclude that u and v are adjacent in $O_k(F)$.

Part b) comes from the following reasoning. If $\texttt{Seq}[u]\texttt{Seq}[v]$ is an edge of $O_k(F)$ then without loss of generality $\texttt{Seq}[u] \xrightarrow{l} \texttt{Seq}[v]$ for some $l \geq k$. If furthermore uv are not adjacent in $G_k(I)$, then in particular $u \not\xrightarrow{l} v$. We conclude that each edge $e \in O_k(F) - G_k(I)$ occurs with probability $\frac{1}{4^l} \leq \frac{1}{4^k}$ by Theorem1, Part b) (actually, by the remark). Then

$$\Pr(O_k(F) \neq G_k(I)) = \Pr(e_1 \text{ occurs } \cup e_2 \text{ occurs } \ldots \cup e_l \text{ occurs}),$$

where $e_1, e_2, ..., e_l$ are the edges *not* in $G_k(I)$. Then,

$$\Pr(O_k(F) \neq G_k(I)) \leq \sum_{e \notin G_k(I)} \Pr(e \text{ occurs}) \leq \sum_{e \notin G_k(I)} \frac{1}{4^k} \leq \frac{n^2}{4^k}$$

□

Theorem 2, shows us that choosing a reasonably large k makes the probability of the two graphs $O_k(F)$ and $G_k(I)$ being equal very large. For instance, if $n = 1000$, choosing

$k = 20$ we have $\Pr(G_k(I) = O_k(F)) \geq 0.999999$. This is very important, because $O_k(F)$ can be constructed based on the fragments alone, while $G_k(I)$ is based upon information about their position. In fact, we can infer this positional information from the sequences alone, with high probability.

3 Practical issues

The preceding discussion assumes that the fragments have no errors, that there are no chimeric fragments, and that no repeats exist. Sequencing errors in the fragments will change a bit the probabilities, and we have to look for approximate overlaps instead of exacts overlap, but the main ideas still work. Regarding chimeric fragments, those are possible to spot and remove from consideration. Chimeric fragments are those that, by problems during the cloning phase, correspond to a sequence obtained by concatenation of two disjoint, non contiguous intervals.

However, repeats pose more serious problems [Cer00, Gus97, SM97, Pev00, Wat95] and to deal with them we need to study a new class of graphs: *interval graphs with repeats*.

4 Interval Graphs with Repeats

To help solve DNA fragments assembly problems we could apply the procedure described in Figure 1.

```
Assembly Model()
1  Input collection of fragments F
2  k ← 10
3  while 1
4  do compute the overlap graph O_k(F)
5      if O_k(F) is an proper interval graph
6         then break
7      k ← k + 1
8  Compute all interval models compatible with O_k(F)
9  Choose the best one
```

Figure 1: Finding Assembly Models.

This procedure tries successive values of k (starting with $k = 10$) until $O_k(F)$ is an interval graph. Then, all interval models compatible with $O_k(F)$ are examined in search of the best I for which $G_k(I) = O_k(F)$.

With regard to interval graphs, two algorithms are important here: (1) an algorithm that recognizes intervals graphs, and (2) an algorithm that generates all interval models compatible with a given interval graph.

We have argued that the practical issues of dealing with errors and chimeric fragments fit in this same framework, but not the issue of repeats. To deal with repeats, we need to

study a new class of graphs.

If there is a repeated region in the original DNA, it is not correct to assume a uniform distribution for $\texttt{Seq} : [1, L] \rightarrow \{A, C, G, T\}$. Instead, there are intervals $A = [a_1, a_2]$ and $B = [b_1, b_2]$, both contained in $[1, L]$, disjoint, with the same size, and such that $\texttt{Seq}[a_1, a_2] = \texttt{Seq}[b_1, b_2]$. In other words, the letters in $[b_1, b_2]$ are completely determined by the letters in $[a_1, a_2]$ and are therefore not independent of those. On the other hand, we can still assume that the letters outside A and B are uniformly distributed , and independent from the letters in A.

Theorem 1 is not true in this context. We need a modified result. First, we need a version of the interval intersection concept in the presence of repeats A and B. Let us say that two positions i and j **correspond** if either $i = j$ or ($a_1 \leq i \leq a_2$ and $j = b_1 + i - a_1$) or ($b_1 \leq i \leq b_2$ and $j = a_1 + i - b_1$). It is clear that if i and j correspond then $\Pr(\texttt{Seq}[i] = \texttt{Seq}[j]) = 1$ and if i and j do not correspond then $\Pr(\texttt{Seq}[i] = \texttt{Seq}[j]) = \frac{1}{4}$. Analogously, we say that two intervals $[i, j]$ and $[p, q]$ **correspond** if i corresponds to p, $i + 1$ corresponds to $p + 1$, ... , and j corresponds to q. Notice that the intervals must have the same size in order to satisfy this relationship. Again, if $[i, j]$ and $[p, q]$ correspond then $\Pr(\texttt{Seq}[i, j] = \texttt{Seq}[p, q]) = 1$ and $\Pr(\texttt{Seq}[i, j] = \texttt{Seq}[p, q]) = \frac{1}{4^k}$ otherwise where k is the size of $[i, j]$ and of $[p, q]$ (that is, $k = j - i + 1 = q - p + 1$). We are now ready to define the intersection relationship. We use the notation $[i, j] \overset{k}{\rightarrow}_{A,B} [p, q]$ when $[j - k + 1, j]$ corresponds to $[p, p + k - 1]$.

Theorem 3 *Let $A = [a_1, a_2]$ and $B = [b_1, b_2]$ be two disjoint intervals of the same size contained in $[1, L]$, and let $\texttt{Seq} : [1, L] \mapsto \{A, C, G, T\}$ be a uniformly distributed sequence in $[1, L] \setminus B$. Then:*

a) if $[i, j] \overset{k}{\rightarrow}_{A,B} [p, q]$ then

$$\texttt{Seq}[i, j] \overset{k}{\rightarrow} \texttt{Seq}[p, q]$$

b) if $[i, j] \overset{k}{\not\rightarrow}_{A,B} [p, q]$ then

$$\Pr(\texttt{Seq}[i, j] \overset{k}{\rightarrow} \texttt{Seq}[p, q]) = \frac{1}{4^k}.$$

Proof: Analogous to the proof of Theorem 1. □

Algorithm Assembly Model of Figure 1 is based on Theorem 2, but this result does not hold for DNA molecules with repeats. Theorem 3 replaces Theorem 1 in the context of repeats, but we also need a replacement of $G_k(I)$.

Let A and B be as in the statement of Theorem 3 and let I be a proper interval model over $[1, L]$. We define the graph $G_k^{A,B}(I)$ as follows: the vertices are the intervals of I and two intervals $[i, j]$ and $[p, q]$ are adjacent when $[i, j] \overset{l}{\rightarrow}_{A,B} [p, q]$ form some $l \geq k$.

An abstract graph G is called a **proper interval graph with repeats** when G is isomorphic to $G_k^{A,B}(I)$ for some proper interval model I, repeated regions A, B, and integer $k \geq 0$.

With these definitions, the algorithm in Figure 2 can be used in place of Assembly Model.

```
Assembly Model with Repeats()
1  Input collection of fragments F
2  k ← 10
3  while 1
4  do compute the overlap graph O_k(F)
5     if O_k(F) is a proper interval graph with repeats
6        then break
7     k ← k + 1
8  Compute all interval models compatible with O_k(F)
9  Choose the best one
```

Figure 2: Finding Assembly Models with Repeats.

5 Conclusion

Therefore, the main problems that need to be solved are:

- efficient recognition of proper interval graphs with repeats
- finding all models for a given such graph, satisfying the length constrains.

References

[Cer00] Fábio Ribeiro Cerqueira. Montagem de fragmentos de DNA. Master's thesis, Universidade Estadual de Campinas, Campinas - SP, Janeiro 2000.

[EG01] W. J. Ewens and G. R. Grant. *Statistical Methods in Bioinformatics: An Introduction.* Springer-Verlag New York, Inc, 2001.

[Gol80] Martin Charles Golumbic. *Graph Theory and Perfect Graphs.* Academic Press, 1980. Courant Institute of Mathematical Sciences - New York University.

[Gus97] Dan Gusfield. *Algorithms on Strings, Trees, and Sequence.* Cambridge University Press, 1997.

[Har72] Frank Harary. *Graph Theory.* Addison-Wesley Series in Mathematics. Addison-Wesley Publishing Company, third edition, 1972.

[Li01] Lin Tzy Li. Montagem de fragmentos de DNA pelo método "ordered shotgun sequencing" (OSS). Master's thesis, Universidade Estadual de Campinas, Campinas - SP, Fevereiro 2001.

[Pev00] Pavel Pevzner. *Computational Molecular Biology - An Algorithmic Approach.* The MIT Press, 2000.

[SM97] João C. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology.* PWS Publishing Company, 1997.

[Wat95] Michael S. Waterman. *Introduction to Computational Biology.* Chapman and Hall, 1995.