A Graph Theoretic Approach to Restriction Mapping

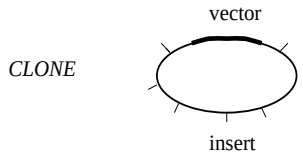
Tao Jiang
Univ of California, Riverside

Richard M. Karp
University of Washington
Seattle, Washington

Restriction Mapping

Target DNA (chromosome): 1 - 100 mb.

Library of cosmid clones: 40 - 50 kb each.



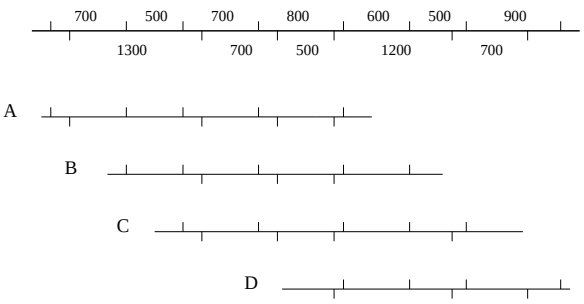
Restriction enzymes: *Eco*RI, *Hin*dIII, *Nsi*I, ...

Restriction fragment: 500 - 20 kb.

For each clone C and enzyme E, we measure the *sizes* of fragments in digest of C by E.

– often resulting in a *multiset* of card. 10.

Map: relative placement of clones and restriction sites.

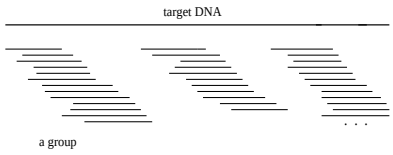


Objective: find a most "compact" map.

Experimental errors:
- Sizing error - 3%.
- Multiplicity error.
- Missing, spurious and vector fragments.
- Bad clones, and so on.

A Paradigm for Restriction Mapping

1. Compute pairwise clone overlap probabilities.
2. Edit data – screening out bad clones, spurious fragments and vector fragments, and identifying missing fragments.
3. Partition the clones into "contigs".



4. For each contig
   - 4.1. Order clones using overlap statistic.
   - 4.2. Identify fragments.
   - 4.3. Construct a map.
5. Merge the maps for all contigs.

We are interested in steps 4.2 and 4.3.

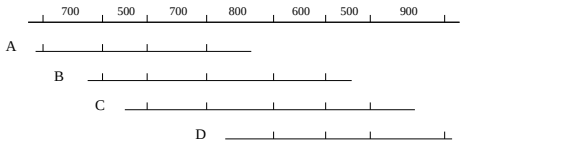
Single Complete Digest (SCD) Mapping

Only one restriction enzyme.

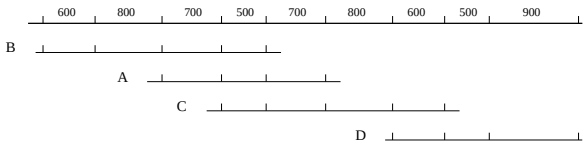
Clone: a *multiset* of integers.

Objective: minimize the number of fragments placed on the target DNA.

$A = \{500, 700, 700\}$ $B = \{500, 600, 700, 800\}$
$C = \{500, 600, 700, 800\}$ $D = \{500, 600, 900\}$



(a)



(b)

A Combinatorial Problem in SCD Mapping

Assumptions:

1. Data is error-free.

2. Subclones have been removed.

   Subclones can be detected and handled specially.

3. A single contig.

4. Ordering of clones on target DNA is given.

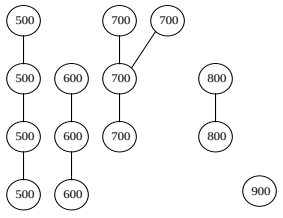
   One can divide the clones into contigs and infer the ordering by overlap statistic and TSP techniques.

Problem MOP: Find a most compact map consistent with the given ordering.
– *Fragment Identification Problem*

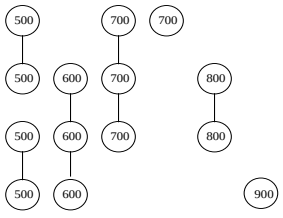
A Graph Formulation of MOP

Clone-Fragment Graph:
A multistage graph whose stages correspond to clones and edges link fragments of identical size in adjacent clones.

$A = \{500, 700, 700\}$
$B = \{500, 600, 700, 800\}$
$C = \{500, 600, 700, 800\}$
$D = \{500, 600, 900\}$



(a)

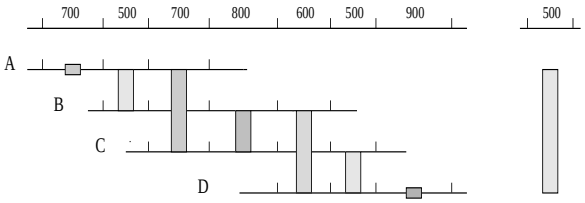


(b)

Path cover: decomposition of the vertices into disjoint paths.

Characterizing Map in Terms of Path Cover

Each map corresponds to a path cover, where each path represents a physical fragment.



Given path $800 \to 800$, we can't have path $500 \to 500 \to 500 \to 500$, and vice versa.

$800 \to 800$ *dominates* $500 \to 500 \to 500 \to 500$.

Lemma. Consistent map $\Leftrightarrow$ *dominance-free* path cover.

MOP $\Leftrightarrow$ Find a smallest dominance-free path cover for any clone-fragment graph.

The Hardness of MOP

Theorem. Approximating MOP with ratio $2-\epsilon$ is NP-hard, for any $\epsilon > 0$.

A problem useful in the reduction:

Multiset Cover by A Common Subsequence:
INSTANCE: Two sequences $A = a_1, \ldots, a_n$ and $B = b_1, \ldots, b_m$, and a multiset $S = \{s_1, \ldots, s_l\}$.
QUESTION: Do A and B have a common subsequence that covers S?

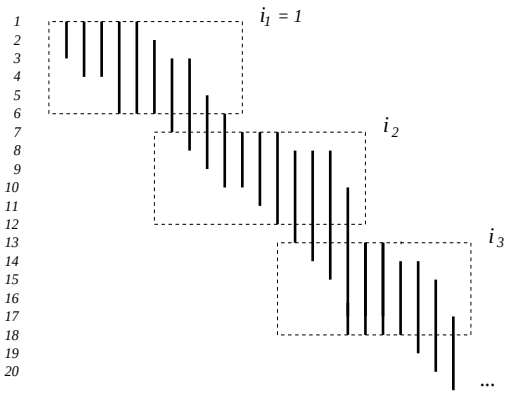
E.g. $A = 5, 6, 7, 5, B = 5, 5, 7, 6, S = \{5, 7\}$.

E.g. $A = 5, 6, 7, 5, B = 5, 5, 7, 6, S = \{5, 5, 7\}$.

Approximation Algorithms for MOP

Theorem. There exists a quadratic time algorithm that approximates MOP with ratio 2.

Partition the clone-fragment graph G w.r.t. an optimal dominance-free path cover:



Approximating MOP with Ratio 2

A path cover is *simple* if each path starts at the first stage or ends at the last stage.

Lemma.
1. Each such subgraph has a simple path cover.
2. The sum of the optimal simple path cover sizes of these subgraphs is at most 2opt(G).

Lemma. An optimal simple path cover can be found in linear time.

Partition unknown? Dynamic programming!

Remark. Some ideas of the algorithm are used in the software RMAP being developed at University of Washington.

MOP with Bounded Multiplicity

MOP-k: At most k elements are identical in each multiset.

In practice, often $k = 3$.

Open Question: Is MOP-k NP-hard, for any k?

In particular, is MOP-1 NP-hard?

Theorem. MOP-1 has a polynomial time approximation scheme (PTAS).

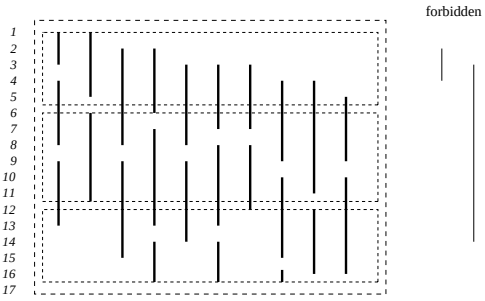
I.e. We can approximate MOP-1 with ratio $1+\epsilon$ for any $\epsilon > 0$ in polynomial time.

The degree of the polynomial depends on ϵ.

Approximating MOP-1

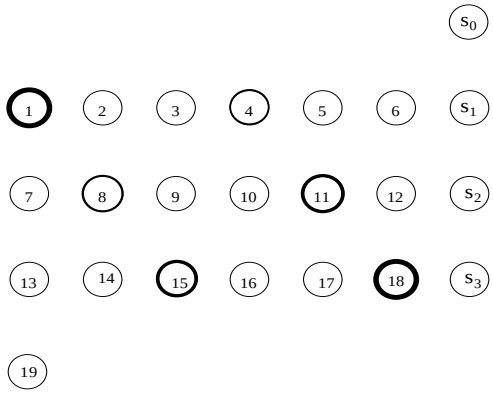
1. Each multiset is actually a set.
2. The clone-fragment graph is a collection of disjoint paths.
3. MOP-1 $\Leftrightarrow$ Given some intervals on the real line, divide them into the smallest number of intervals not properly containing each other.

Key of the PTAS: Given a partition $\mathcal{P}$ of a clone-fragment graph G into d components, find an optimal dominance-free path cover of G that *respects* $\mathcal{P}$.



Reducing to Shortest Paths with Planarity Constraint

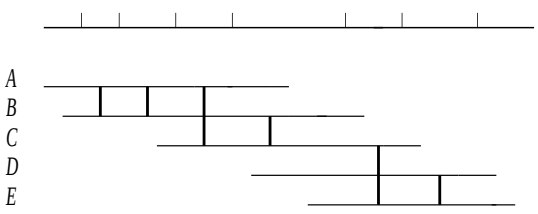
Shortest Non-crossing Paths:
Given a planar embedding of a complete d-stage graph and some pairs of vertices, connect the pairs with paths or pseudo paths using *non-crossing* edges to minimize the total length.



Does Knowing the Interleaving of Clones Help?

Interleaving: which clones overlap on target DNA and which don't.

The information and techniques (*e.g.* overlap statistic and TSP) for the inference of ordering may allow us to infer interleaving.

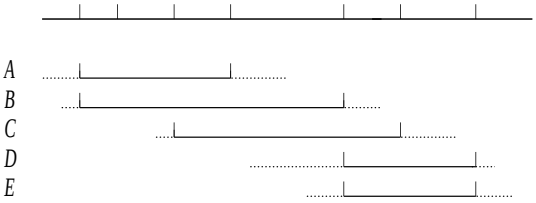


Problem MIP: Find a most compact map consistent with given interleaving.

Theorem. MIP-3 is NP-hard even if the coverage depth is 5.

A Tractable Variant of MIP

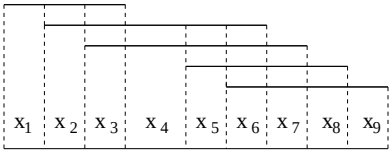
Suppose interleaving tells what clones share common fragments.



1. Overlap statistic based on common fragments.
2. Fragments are short oligonucleotides in hybridization mapping with non-unique probes.

Theorem. The variant is solvable in quadratic time.

We know exactly what kind of paths are legal. Thus, fragments of different sizes can be handled separately.



Each fragment is completely contained in some atomic interval.

Consider fragments of size a.

x_i: Number of a-paths in interval i.
a_j: Multiplicity of a in clone j.
l_j (r_j): Leftmost (rightmost) atomic interval on j.

For each clone j, we have

$$x_{l_j} + \cdots + x_{r_j} = a_j.$$

The goal is to minimize $\sum_{i=1}^{m} x_i$.

This can be transformed into a shortest path problem and solved using Bellman-Ford.

Open Questions:

1. Is MOP-k NP-hard for any k?
2. Is MOP-1 in P?
3. Is MIP-1 in P?