

MA3259 Lecture 14

Genome Mapping, Assembly and Sequencing Part III

LX Zhang
Department of Mathematics
National University of Singapore
matzlx@nus.edu.sg

Build other physical maps

- Amplify the DNA you want to map, e.g. a chromosome or a smaller region.
- Break many copies of DNA into pieces (50kb - 1Mb clones). The breaking of the DNA is often done by physical means so it is random and give overlapping pieces, instead of using restriction enzymes (except for restriction site mapping).
- Determining which of these cloned fragments overlap by generating fingerprints of the fragments.

*A **fingerprint** should describe part of the information contained in a DNA fragment (in a unique way).*

- Reconstruct the order and distances between cloned fragments and markers.

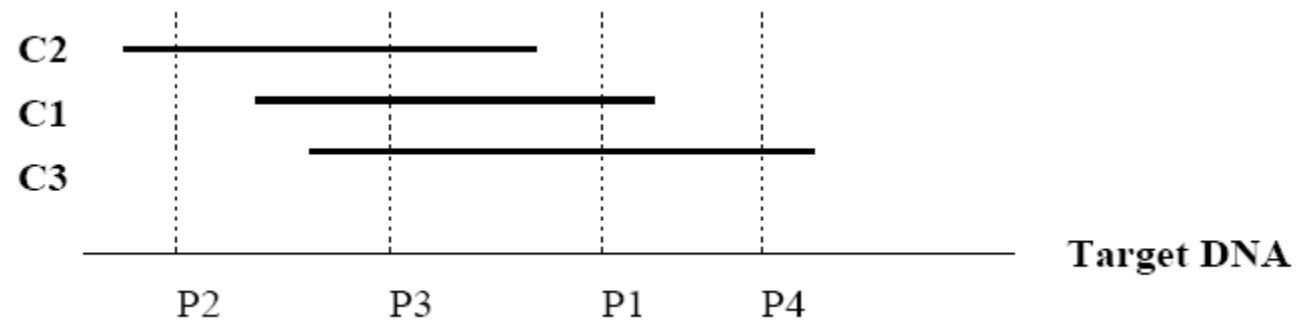
Hybridization mapping

- All clones are exposed to a number of STSs (called *probes*) by hybridization and then which of these STS's appear in the clone is determined. (The experimental output data is given in a matrix $M = (m_{ij})$, where $m_{ij} = 1$ if clone C_i contains probe P_j .)
- The matrix M is used to locate clones in the target DNA sequence.

In the mapping approach, Sequence Tagged Site (STS) markers are used to give information about each fragment. STSs are extracted from the DNA strands being studied and are about 200 bps long. Since each STS is sufficient long, it is unlikely to occur a second time on the target DNA sequence; hence, it identifies a unique site along the DNA strand.

Example Consider the following experimental result

	P_1	P_2	P_3	P_4
C_1	1	0	1	0
C_2	0	1	1	0
C_3	1	0	1	1



By rearranging the probes in the same order as they appear in the target DNA, we obtain

	P_2	P_3	P_1	P_4
C_1	0	1	1	0
C_2	1	1	0	0
C_3	0	1	1	1

A binary matrix has the *consecutive one property* if its columns can be permuted in such a way that 1s in each row occur in consecutive positions.

CONSECUTIVE ONE PROPERTY PROBLEM

Instance: A $(0, 1)$ -matrix $M = (m_{ij})$.

Question: Verify whether M has the consecutive one property.

Theorem 5.4.1 *There is polynomial time algorithm for verifying whether a binary matrix has the property and then finding a desired permutation.*

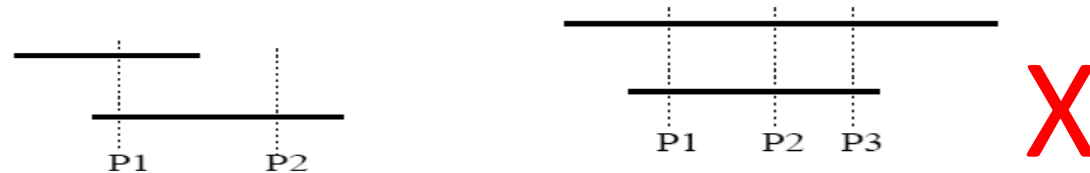
Hybridization can be done easier if the experiment data are error-free.

Remark Hybridization and cloning are imperfect. The experiment has fairly high error rates – roughly 2% false positive (a probe is reported when the fragment does not contain it) and 10% to 20% false negative (a probe is not reported when it should be). With real data, mapping becomes very complicated.

Consider each clone as a subset of probes.

Without loss of generality, we assume that the hybridization experimental data satisfy the following conditions:

- *Non-inclusion.* There is no clone containing another clone. Here, a clone X *contains* a clone Y if all the probes in Y appear also in X .

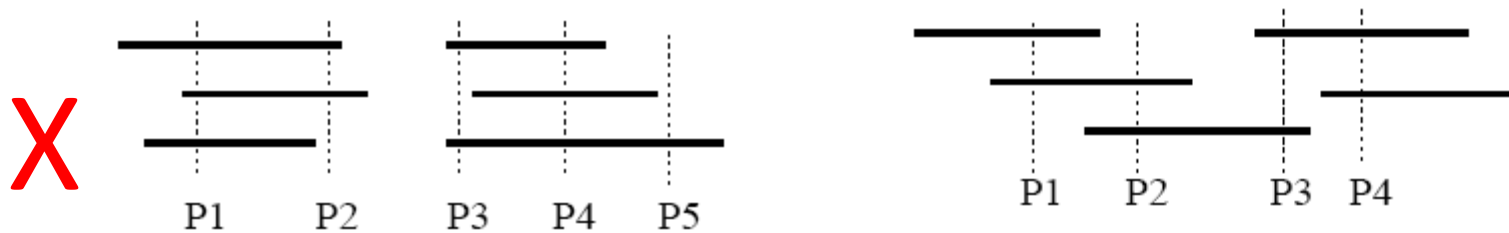


Assume $\{P_1, P_2, \dots, P_m\}$ is a set of STS probes. Let \mathcal{S}_i be the set of clones in the hybridization data containing probe P_i .

- *Distinguishability.* $\mathcal{S}_i \neq \mathcal{S}_j$ for different probes P_i and P_j .



- *Connectedness.* For every partition of the set of probes into two non-empty sets A and B , there are probes $P_i \in A$ and $P_j \in B$ such that $\mathcal{S}_i \cap \mathcal{S}_j \neq \phi$.



Proposition 5.4.1 *Let P_{i-1}, P_i, P_{i+1} appear consecutively in the correct map.*

$$\max_k |\mathcal{S}_i \cap \mathcal{S}_k| = |\mathcal{S}_i \cap \mathcal{S}_{i+1}| \text{ or } |\mathcal{S}_i \cap \mathcal{S}_{i-1}|.$$

Consecutive probes appear in more clones than other pairs of probes.

- (1) If $\max_{k \neq i} |\mathcal{S}_i \cap \mathcal{S}_k|$ is achieved only at probe P_j , then P_j is adjacent to P_i in the correct map.
- (2) If $\max_{k \neq i} |\mathcal{S}_i \cap \mathcal{S}_k|$ is achieved at several probes, then, one of them with minimal $|\mathcal{S}_j|$ is adjacent to P_i in the correct map. (Why?)

These two facts lead to the following algorithm:

Starting with an arbitrary probe P_i , we extend an already found block of k consecutive probes at the right or left end using the above two facts at step $k + 1$, for each $k = 1, 2, \dots$.

Example. Consider the following hybridization data

	P_1	P_2	P_3	P_4	P_5	P_6	P_7
C1	0	1	0	1	0	0	0
C2	0	1	0	0	1	0	0
C3	1	0	0	0	1	1	0
C4	1	0	1	0	0	1	0
C5	0	0	1	0	0	1	1

1. Calculate \mathcal{S}_i 's

2. Verify the three conditions

2.1 Non-inclusion

Consider each clone as a subset of probes.

Non-inclusion is to check if one clone contains another.

2.2 Distinguishability

Check if for any i, j $S_i \neq S_j$

2.3 Connectedness

Transform to check if the following graph $G=(V, E)$ is connected or not:

V = The set of clones, i.e. each node represents a clone.

There is an edge (v_1, v_2) if and only if the corresponding clones contains some probe simultaneously.

3. Apply the algorithm to order probes

For each probe P_i , we let $N(P_i)$ to denote the set of probes that occur together with P_i on at least one clone.

Step 1. Pick P_1 to start the algorithm.

Step 2. $N(1) = \{P_5, P_6, P_3\}$. Since

$$|\mathcal{S}_1 \cap \mathcal{S}_3| = |\mathcal{S}_1 \cap \mathcal{S}_5| < |\mathcal{S}_1 \cap \mathcal{S}_6|,$$

we put P_6 after P_1 in step 2.

Step 3. We have a block B of 2 probes P_1P_6 . $N(6) - \{P_1\} = \{P_3, P_5, P_7\}$. Furthermore,

$$|\mathcal{S}_6 \cap \mathcal{S}_3| > |\mathcal{S}_6 \cap \mathcal{S}_5| = |\mathcal{S}_6 \cap \mathcal{S}_7|.$$

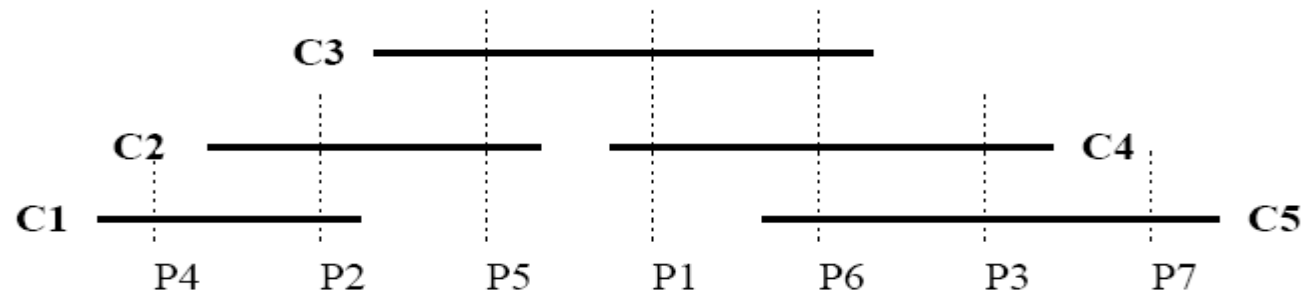
Since $|\mathcal{S}_6 \cap \mathcal{S}_3| = 2 > |\mathcal{S}_1 \cap \mathcal{S}_3|$ we add P_3 to the end of the block B before we start step 4.

Step 4. Now we have a block B of 3 consecutive probes $P_1P_6P_3$. $N(3) - \{P_1, P_6\} = \{P_7\}$. Since $|\mathcal{S}_7 \cap \mathcal{S}_3| = 1$ and $P_1 \notin N(7)$, we obtain a block of 4 consecutive probes $P_1P_6P_3P_7$ by adding P_7 to the end of B .

Step 5. Since all the probes in $N(7) = \{P_3, P_6\}$ lie in the block. We replace the block B by reversal. Now, $B = P_7P_3P_6P_1$.

Step 6. We have that $N(1) - \{P_3, P_6, P_7\} = \{P_5\}$. Noting that $P_7 \notin N(5)$, we add P_5 to the end of B .

This implies the following map of the given clones.

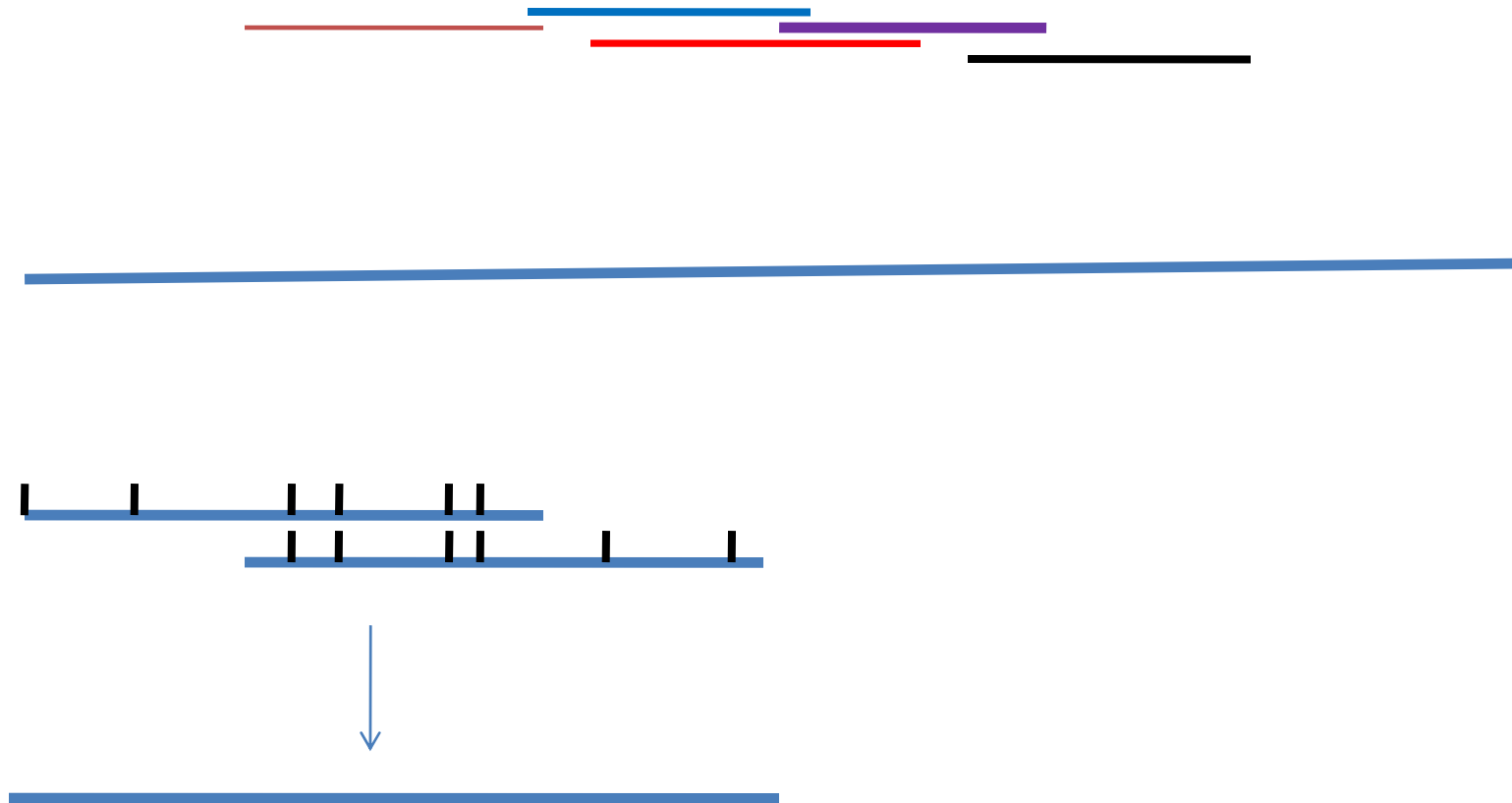


Map assembly using restriction fragments sizes

Double digest technique allows a scientist to construct restriction maps of small DNA molecules, such as mitochondrial DNA. However, this technique does not work for huge DNA molecules. To study a large DNA, scientists break it into small pieces, map each piece, and then assemble the pieces to determine the map of the entire molecule (Olson et al, 1986, and Kohara et al, 1987).

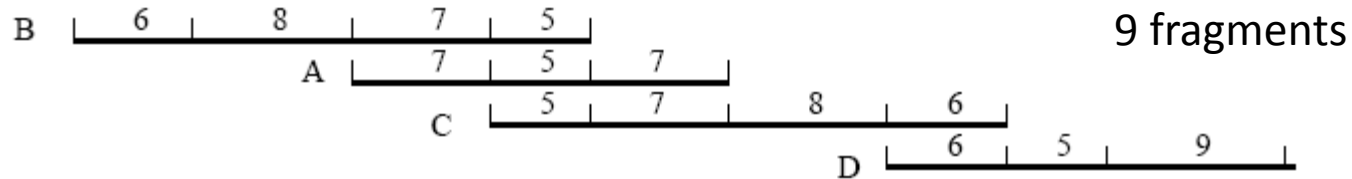
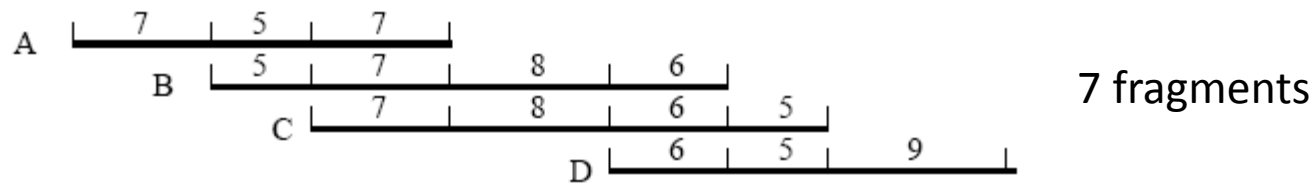
To study individual pieces, one obtains many identical copies of each piece by cloning them. A copy reproduced in this way is called a **clone**. As a result, a **clone library** is obtained which consists of clones from the DNA molecule.

Clones from the clone library may overlap. After the clone library is constructed, scientists want to order the clones, i.e, to reconstruct the relative placement of the clones along the DNA. (This information is lost in the construction of the clone library.)



Example Assume there are four clones A , B , C and D and their fingerprint lengths are

$$A = \{5, 7, 7, \}, B = \{5, 6, 7, 8\}, C = \{5, 6, 7, 8\} D = \{5, 6, 9\}$$



Single Complete Digest (SCD) Mapping Problem

INSTANCE: *A set of clones and for each clone, a multiset of restriction fragment lengths.*

QUESTION: *Find a most compact map that has the minimum number of restriction fragments.*

Another NP-hard problem!

Greedy Approach

Any time there are restriction fragment lengths shared by two clones, they are put together. (Backtrack where there is discrepancy.)

Example Given the following experiment data:

Clones	Restriction fragment lengths
C1	2, 2, 3, 3, 4, 5, 6, 7
C2	1, 2, 3, 3, 4, 8, 9
C3	1, 2, 2, 3, 4, 6, 8

Step 1

C1	3,5,7	2,2,3,4,6		
C3		2,2,3,4,6	1,8	

Step 2

C1	3,5,7	2,6	2,3,4		
C3		2,6	2,3,4	1,8	
C2			2,3,4	1,8	3,9

Explain the hybridization mapping in detail.