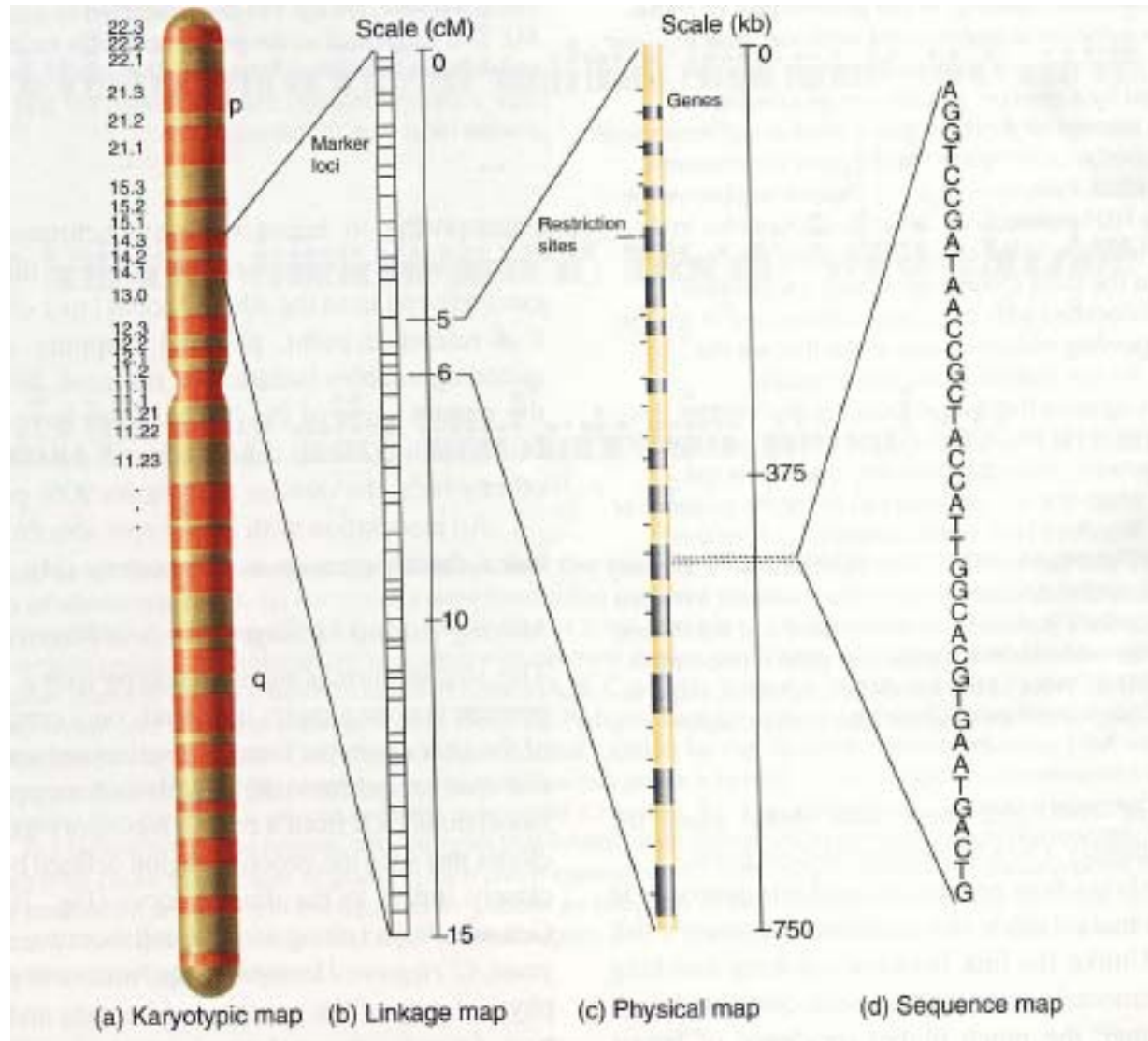


# MA3259 Lecture 13

## Genome Mapping, Assembly & Sequencing Part II: Restriction sites mapping

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

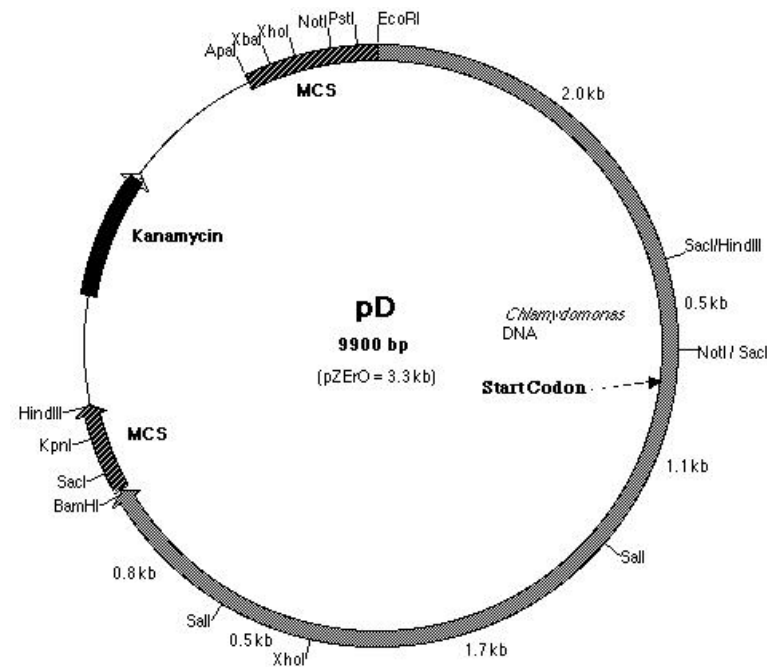
Different maps have different resolution.



(Genetics book by Hartwell et al.)

# Restriction sites mapping

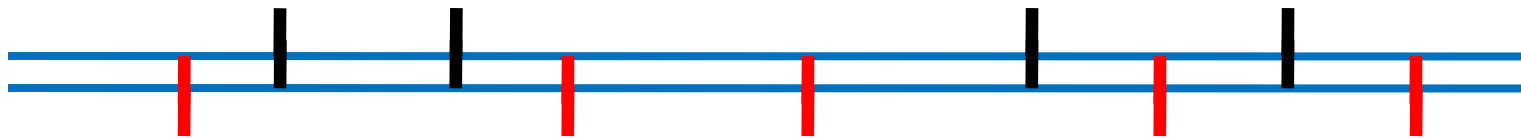
2. **Restriction site map.** It describes the order and distance between enzyme cutting (cleavage) sites. It can be obtained in top-down mapping: A single chromosome is cut into large pieces, which are ordered and subdivided; the smaller pieces are then mapped further. The resolution of this map is about 100,000 to 1Mb.



# Double Digest Mapping

---

One simple technique for restriction site mapping is called **double digest**. In this approach, biologists use two different enzymes. They measure the fragment lengths (not the order) from a complete digestion of the DNA by each of the two enzymes singly, and then by the two enzymes applied together. The positions of restriction sites are determined from these fragment length data.

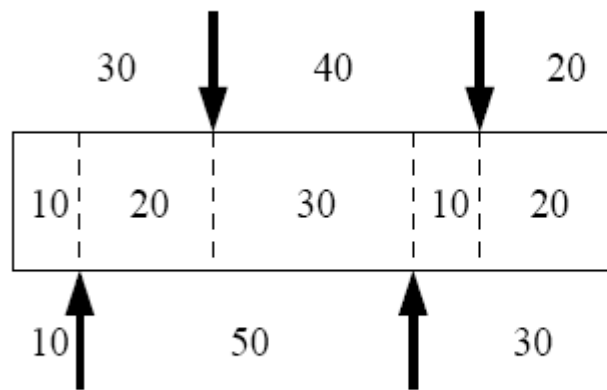


Order information is lost in these experiments!

How to recover the positions of these fragments?

**Example 1.** Using the gel electrophoresis experiments with two restriction enzymes A and B, a biologist obtains information about the lengths of restriction fragments 20, 30, 40 for A, and 10, 30, 50 for B. If the biologist also obtain the lengths of restriction fragments 10, 10, 20, 20, 30 by using the double digest A + B, then the correct map is (a) in Figure 5.1.

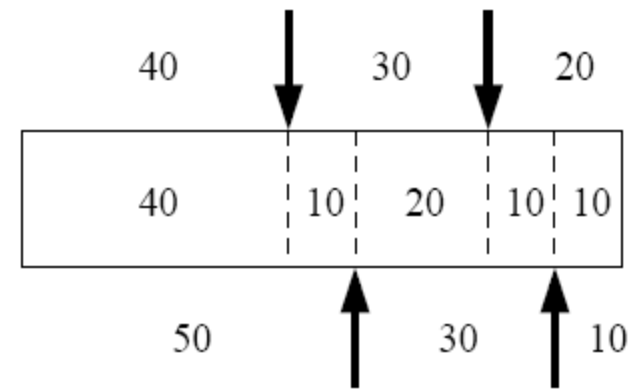
If the lengths of restriction fragments for double digest A + B are 10, 10, 10, 20, 40, then the correct map should be (b) in Figure 5.1.



(a)

Enzyme A

Enzyme B



(b)

## Double Digest Problem (DDP).

Let

$$A = \{a_1 = 0, a_2, \dots, a_m = t \mid a_i < a_{i+1}\}$$

be a set of points in the interval  $[0, t]$ . We define

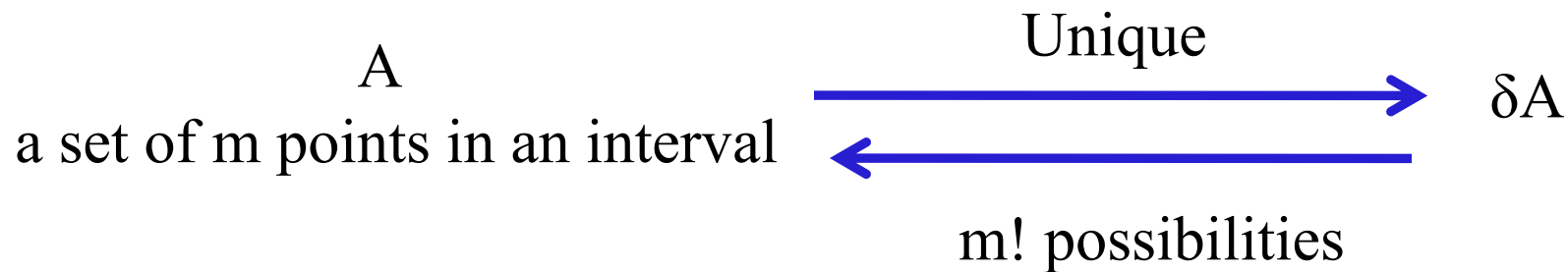
$$\delta A = \{a_2 - a_1, a_3 - a_2, \dots, a_m - a_{m-1}\}.$$



Then, the DDP is formulated as

**Instance:** (Given)  $\delta A, \delta B, \delta(A \cup B)$ .

**Question:** Find  $A \cup B$ .



Given  $\delta A$  and  $\delta B$ ,

there are  $(|\delta A|+1)!$  configurations for the points in A, and

there are  $(|\delta B|+1)!$  configurations for the points in B.

Hence, it is impossible to search for a solution.

**Theorem:** *The DDP is an NP-hard problem.*

This is another problem for which solution is easy to check, but hard to find.

A well-known NP-hard problem is the *set partition problem*: given a set of non-negative integers  $X = \{x_1, x_2, \dots, x_n\}$ , find a disjoint partition  $X = X_1 \cup X_2$  such that the sum of all the integers in  $X_1$  is the same as those in  $X_2$ .

For  $X = \{x_1 = 2, x_2 = 3, x_3 = 4, x_4 = 1, x_5 = 1, x_6 = 1\}$ ,

$X_1 = \{x_1, x_2, x_5\}$  and  $X_2 = \{x_3, x_4, x_6\}$

give a solution to the set partition problem for  $X$ .

For any set of non-negative integers  $X = \{x_1, x_2, \dots, x_n\}$  we define an instance of the double digest problem as

**Instance:** Given  $\delta A = X, \delta B = \{K/2, K/2\}, \delta(A \cup B) = X$ .

**Question:** Find  $A \cup B$ , which is identical to  $A$ .

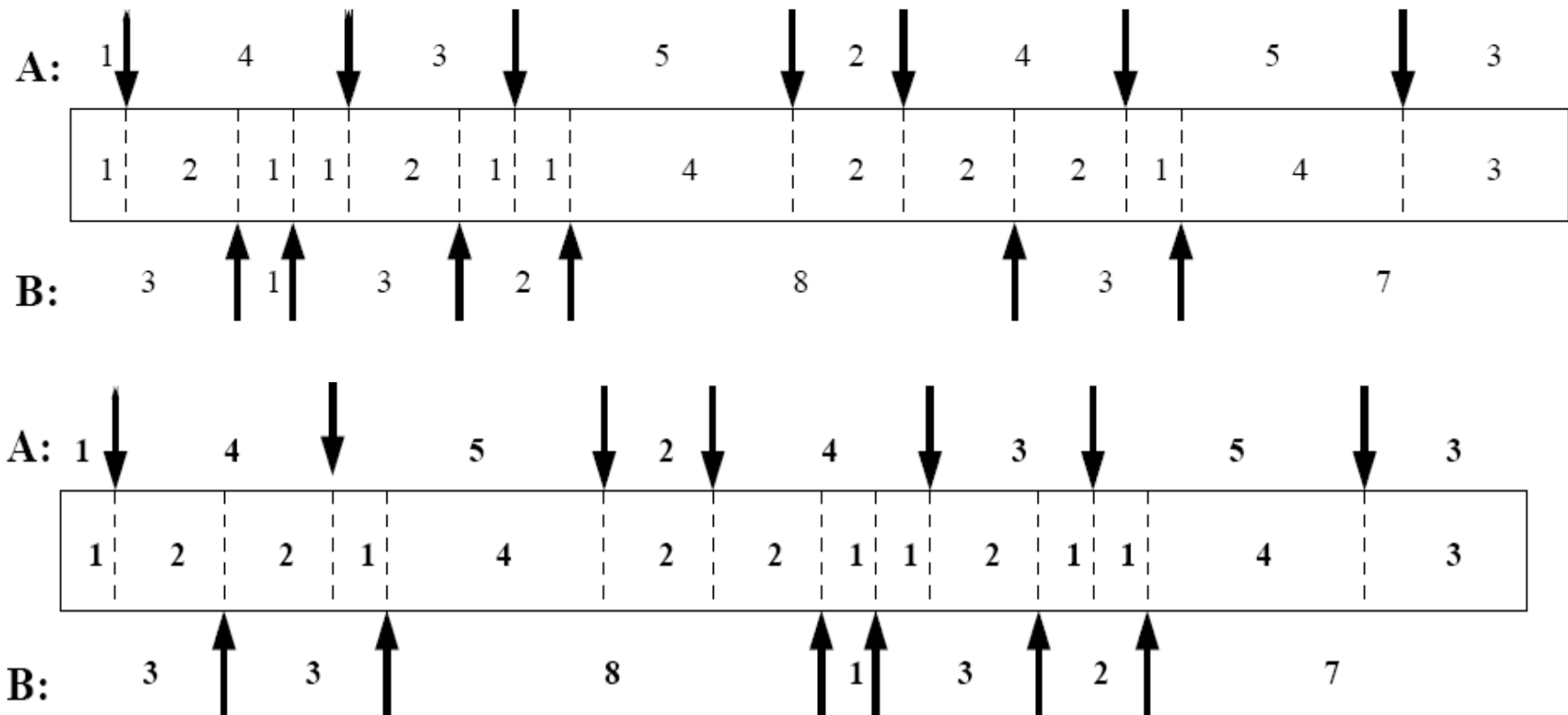
**Example** Let

$$\delta A = \{1, 2, 3, 3, 4, 4, 5, 5\}$$

$$\delta B = \{1, 2, 3, 3, 3, 7, 8\}$$

$$\delta(A \cup B) = \{1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4\}.$$

Then, there are two solutions of the DDP:



**Example** Let

$$\delta A = \{1, 3, 3, 12\}$$

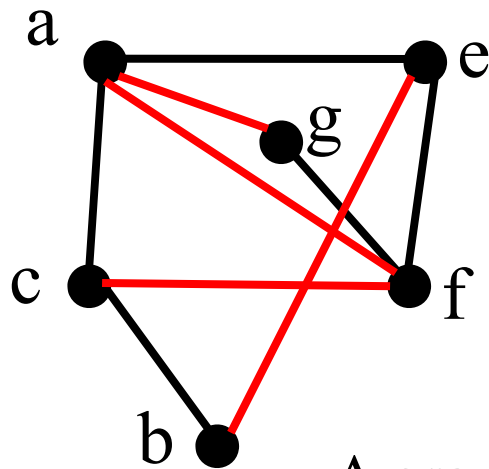
$$\delta B = \{1, 2, 3, 3, 4, 6\}$$

$$\delta(A \cup B) = \{1, 1, 1, 1, 2, 2, 2, 3, 6\}.$$

# Transformation Among Solutions

---

Consider a bicolored graph  $G = G(V, E)$  with vertex set  $V$  and edge set  $E$ , in which edges are colored in two colors.

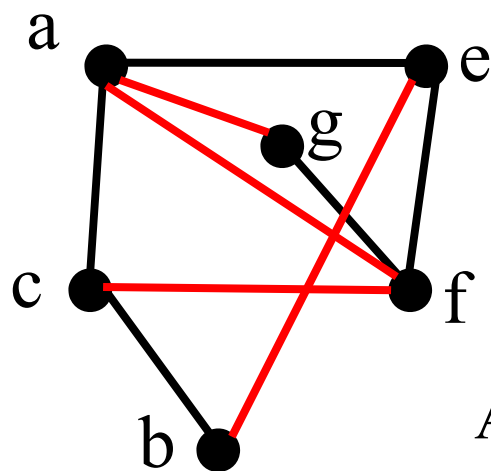


Path:  $a \xrightarrow{\text{red}} g \xrightarrow{\text{black}} f \xrightarrow{\text{black}} e$   
 $a \quad b \quad c \quad g \quad \text{X}$

Cycle:  $a \xrightarrow{\text{red}} g \xrightarrow{\text{black}} f \xrightarrow{\text{black}} e \xrightarrow{\text{black}} a$

A graph is connected if there is a path from each vertex to any other vertex.

Let  $P = x_1x_2 \cdots x_m$  be a path. We use  $P^- = x_mx_{m-1} \cdots x_1$  to denote the reverse path.



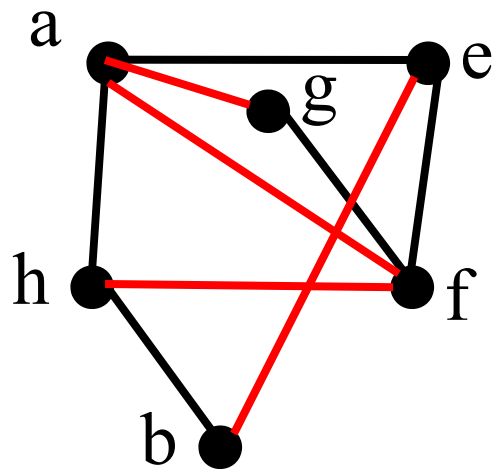
If  $P = a \rightarrow g \rightarrow f \rightarrow c \rightarrow a$

$P^- = e \rightarrow f \rightarrow g \rightarrow a$

Alternate cycle:  $a \rightarrow g \rightarrow f \rightarrow c \rightarrow a$

A path in  $G$  is called *alternating* if the colors of every two consecutive edge  $(x_i, x_{i+1})$  and  $(x_{i+1}, x_{i+2})$  are distinct. Similarly, we can define an alternating cycle in which we consider  $(x_{m-1}, x_m)$  and  $(x_1, x_2)$  as consecutive edges. A path (cycle)  $P$  in  $G$  is called *Eulerian* if every edge  $e \in E$  is traversed by  $P$  exactly once.

Let  $v$  be a vertex in  $G$  and  $c$  a color used in  $G$ . We use  $d_c(v)$  to denote the number of  $c$ -colored edges of  $E$  incident to  $v$ . It is **balanced** if  $d_c(v) = d_{c'}(v)$  for the colors  $c$  and  $c'$  used in  $G$ . The following theorem is a corollary from the Kotzig theorem.



$$d_{\text{black}}(a) = 2; \quad d_{\text{red}}(a) = 2$$

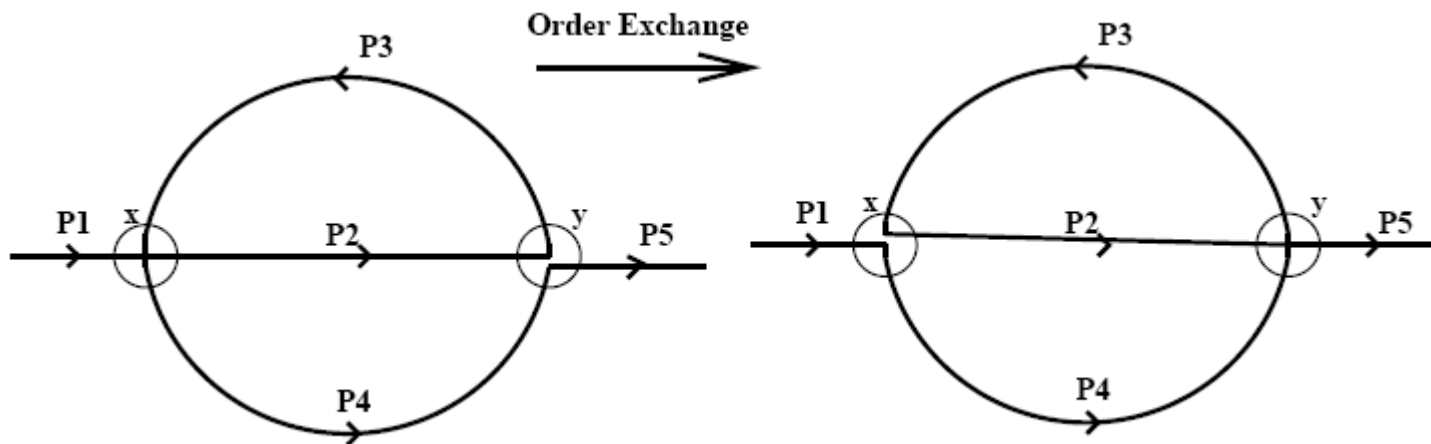
$$d_{\text{black}}(h) = 2; \quad d_{\text{red}}(h) = 1$$

**Theorem 5.3.2** *Let  $G$  be a bicolored connected graph with edges colored by  $r$  and  $b$ . There is an alternating Eulerian cycle in  $G$  if and only if all the vertices in  $G$  are balanced.*

Consider an alternating path

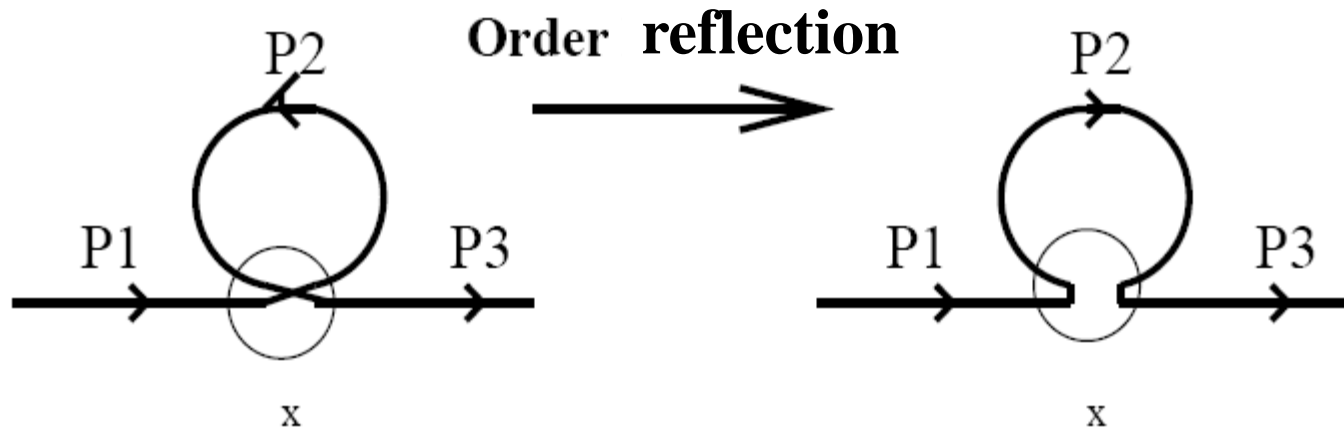
$$P = \underbrace{a_1 a_2 \cdots a_l}_{P_1} \overset{x}{\square} \underbrace{b_1 b_2 \cdots b_m}_{P_2} \overset{y}{\square} \underbrace{c_1 c_2 \cdots c_n}_{P_3} \overset{x}{\square} \underbrace{d_1 d_2 \cdots d_p}_{P_4} \overset{y}{\square} e_1 e_2 \cdots e_q$$

$$P' = \underbrace{a_1 a_2 \cdots a_l}_{P_1} \overset{x}{\square} \underbrace{d_1 d_2 \cdots d_p}_{P_4} \overset{y}{\square} \underbrace{c_1 c_2 \cdots c_n}_{P_3} \overset{x}{\square} \underbrace{b_1 b_2 \cdots b_m}_{P_2} \overset{y}{\square} e_1 e_2 \cdots e_q$$



Consider an alternation path of the form

$$P = \underbrace{a_1 a_2 \cdots a_l}_{P_1} x \underbrace{b_1 b_2 \cdots b_m}_{P_2} x \underbrace{c_1 c_2 \cdots c_n}_{P_3}$$



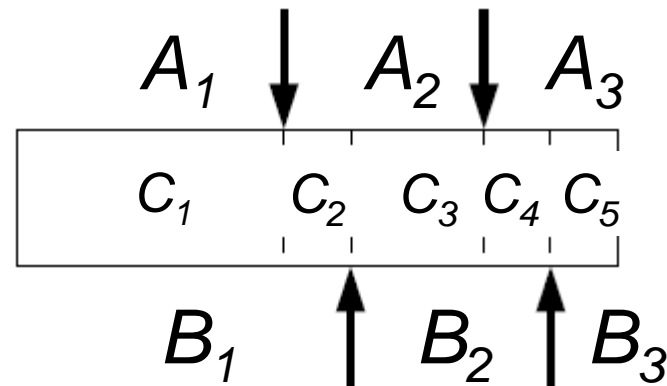
$$P' = \underbrace{a_1 a_2 \cdots a_l}_{P_1} x \underbrace{b_m b_{m-1} \cdots b_1}_{P_2^{-1}} x \underbrace{c_1 c_2 \cdots c_n}_{P_3}$$

**Theorem**

*Every two alternating Eulerian cycles in a bi-colored graph  $G$  can be transformed into each other by a series of order exchanges and reflections.*

Consider a restriction map  $R$  given by (ordered) fragments of single digest  $A$ , and  $B$  and double digest  $A + B$ :

$$\{A_1, A_2, \dots, A_m\}, \{B_1, B_2, \dots, B_n\}, \{C_1, C_2, \dots, C_l\}.$$

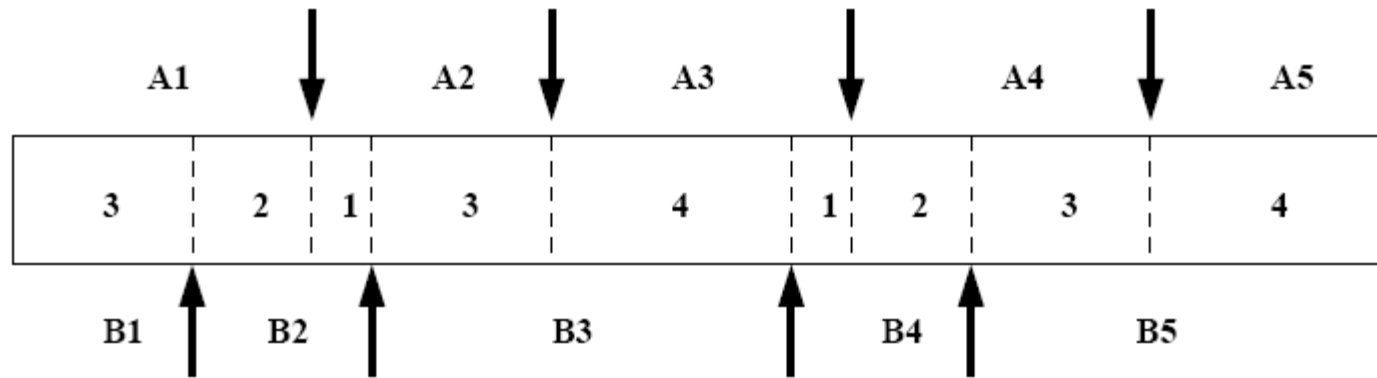


For simplicity, we assume that  $A$  and  $B$  do not cut at the same positions. Then,  $l = n + m - 1$ .

A **fork**  $F(A_i)$  of fragment  $A_i$  is the set of double digest fragments  $C_j$  contained in  $A_i$ :

$$F(A_i) = \{C_j : C_j \subseteq A_i\}.$$

Similarly, a fork of  $B_j$  can be defined. A fork containing at least two fragments is called a **multifork**.



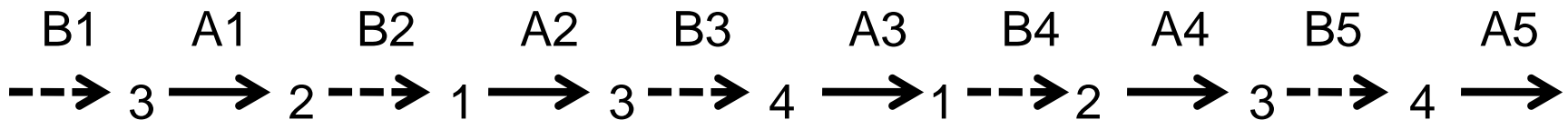
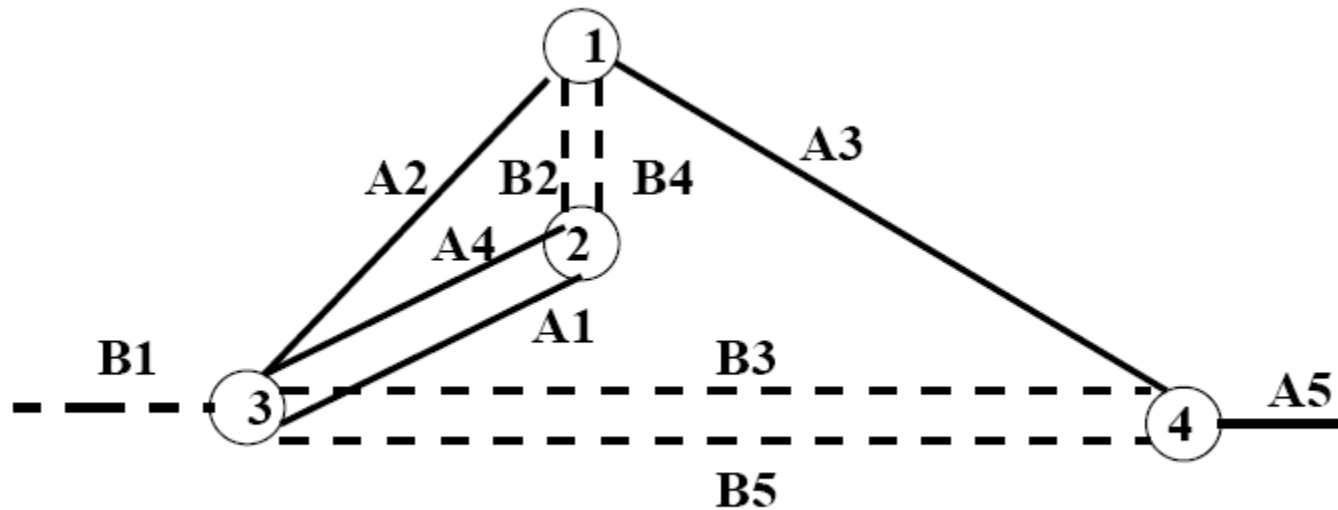
Leftmost and rightmost fragments of a multifork are *border fragments*. Obviously,  $C_1$  and  $C_l$  are border fragments. In addition,

*Every border fragment, excluding  $C_1$  and  $C_l$ , belongs to exactly two multiforks  $F(A_i)$  and  $F(B_j)$  for some  $i$  and  $j$ . Border fragments  $C_1$  and  $C_l$  belong to exactly one multifork.*

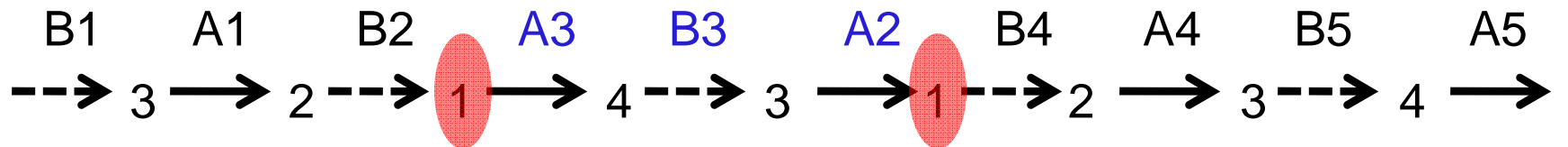
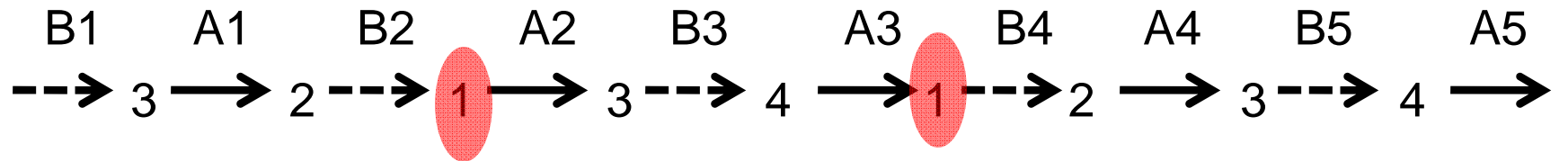
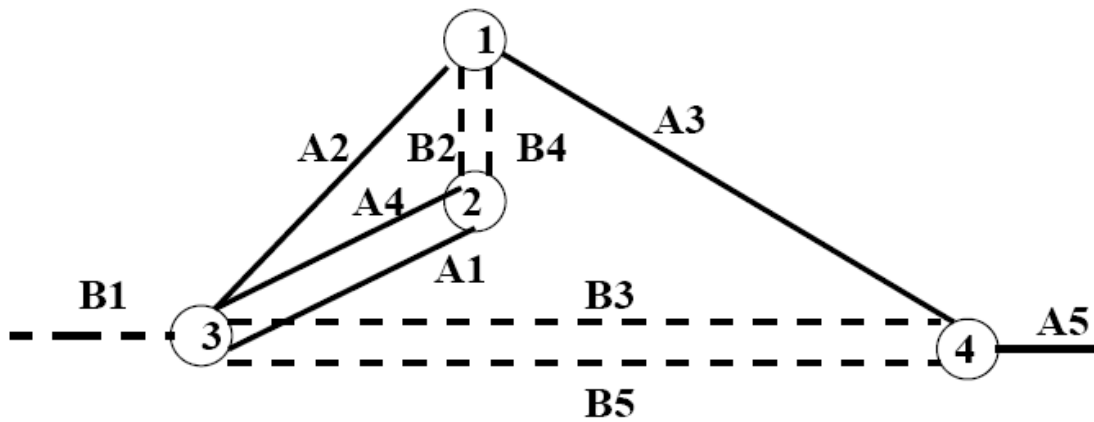
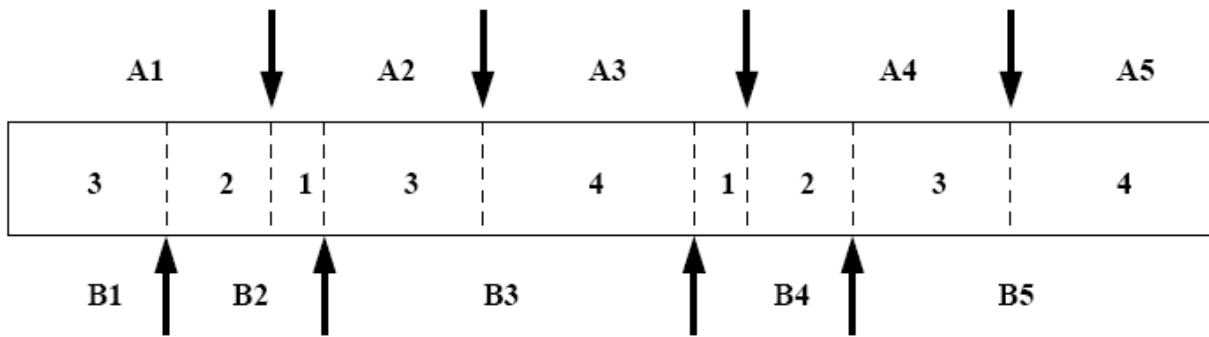


**Theorem 5.3.4** (1) Every restriction map  $R$  defines one-to-one an alternating Eulerian path in its fork graph  $G(R)$ .

(2) Every alternating Eulerian path in  $G(R)$  corresponds to a restriction map



**Remark:** Some fragments in a restriction map may not appear in the corresponding Alternating Eulerian path.



Give an example in detail to show that  
Not all fragments appear in the Fork graph.

With that, it takes about 1.5 hours.