

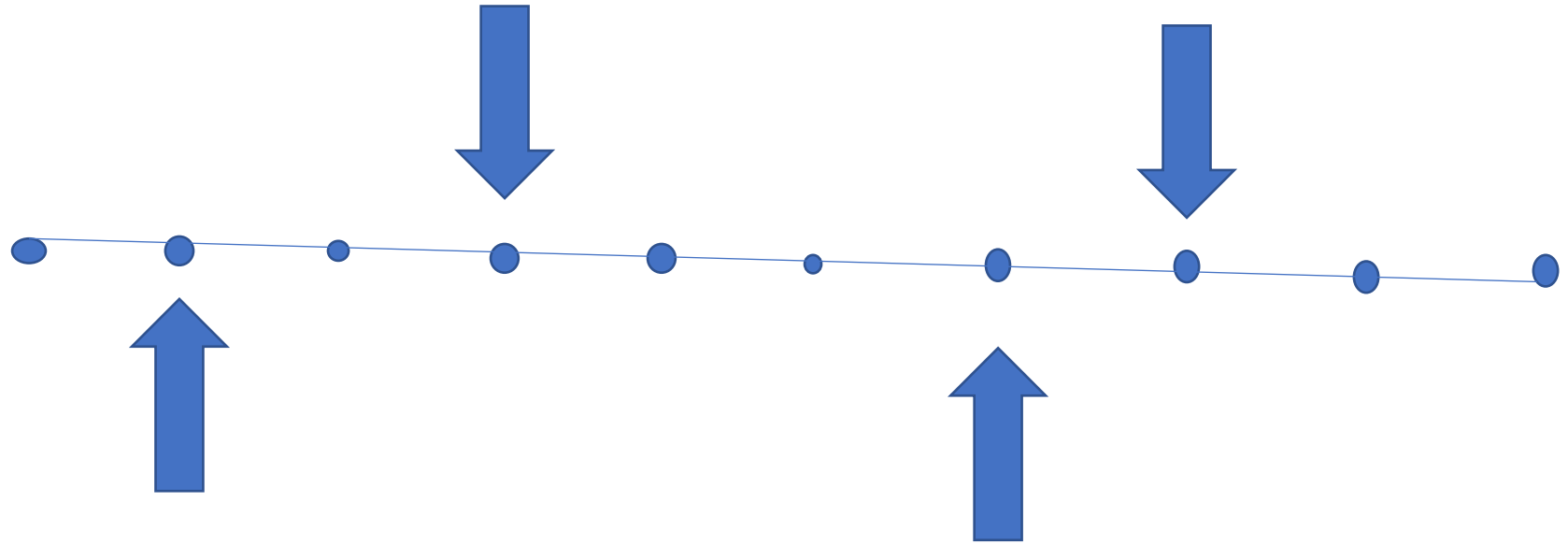
## *Double Digest Problem (Physical Maps)*

- Hamilton Smith discovered in 1970 that the restriction enzyme *hindII* cleaves DNA molecules at every occurrence of a sequence **GTGCAC** or **GTTAAC**. Later in 1973, Danna et al. constructed the first **restriction map for Simian Virus 40 DNA**.
- **Restriction maps** also known as **physical maps** representing DNA molecules with points of cleavage (sites) by restriction enzymes have become fundamental data structures in molecular biology.
- To build a restriction map, biologists use different *biochemical* techniques to derive indirect information about the map and use **combinatorial methods** to reconstruct the map from the data.

# Methods

- Partial Digest problem is one of the well-known method to reconstruct the restriction sites.
- *Double Digest* is a much simpler experimental mapping technique than *Partial Digest*.
- In this approach, a biologist maps the positions of sites of two restriction enzymes by complete digestion of DNA in such a way that only fragments between *consecutive sites* are formed.
- The problem of determining the positions of the cuts from fragment length data is known as the *Double Digest Problem* (DDP).

## *Examples of DDP*



## DDP

- As shown in previous page, there are two enzymes A and B (from above and bottom respectively).
- From A, we have distances between two adjacent sites: 3, 4 and 2. For convenience, we let  $A = \{2, 3, 4\}$ , the set of distances. Similarly, we have  $B = \{1, 3, 5\}$ . Combining A and B,  $A \wedge B = \{1, 2, 3, 1, 2\} = \{1^2, 2^2, 3\}$ .

### Problem

*Given A, B and  $A \wedge B$ , find a physical map.*

*Note: In DDP, we only collect those distances from adjacent sites.*

## Another Example

### Example

$A = \{1, 2, 3, 3, 4, 4, 5, 5\}$ ,  $B = \{1, 2, 3, 3, 3, 7, 8\}$  and  $A \wedge B = \{1^6, 2^5, 3, 4^2\}$ .

### Solutions

(1)  $\mathbf{A} = (\mathbf{1}, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, \mathbf{1})$ ,

(2)  $\mathbf{B} = (\mathbf{1}, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, \mathbf{1})$  and

(3)  $\mathbf{A} \wedge \mathbf{B} = (\mathbf{1}, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, \mathbf{1})$ .

Note that the “1” at the left and right end means the end points respectively.

(\*) The answer uses the support idea for  $(0, 1)$  – vectors in  $Z_{28}$ .

## *Support of (0, 1) - vectors*

- We can use a set form to represent a (0, 1)-vector, the set is called the support of the vector.
- For example, the support of  $(1, 0, 1, 1, 0, 0, 0, 1)$  is  $\{0, 2, 3, 7\}$ .
- The answer mentioned in previous page is:
  - (1)  $\{\mathbf{A}\} = \{0, 1, 5, 8, 13, 15, 19, 24, 27\}$ ,
  - (2)  $\{\mathbf{B}\} = \{0, 3, 4, 7, 9, 17, 20, 27\}$  and
  - (3)  $\{\mathbf{A} \wedge \mathbf{B}\} = \{0, 1, 3, 4, 5, 7, 8, 9, 13, 15, 17, 19, 20, 24, 27\}$ .

## Answer 2

- (1)  $\{A\} = \{0, 1, 5, 10, 12, 16, 19, 24, 27\}$ ,
- (2)  $\{B\} = \{0, 3, 6, 14, 15, 18, 20, 27\}$  and
- (3)  $\{A \wedge B\} = \{0, 1, 3, 5, 6, 10, 12, 14, 15, 16, 18, 19, 20, 24, 27\}$ .

- Clearly, the above two solutions are not symmetric.

### Definition (Similar solutions)

Two solutions of DDP are **similar** if one solution can be obtained from transforming another solution to this solution via suitably defined *transformations*.

# The Idea of Cassette

## Definition (Cassette)

Let  $A = \{a_1, a_2, \dots, a_m\}$ ,  $B = \{b_1, b_2, \dots, b_n\}$  and  $C = \{c_1, c_2, \dots, c_l\}$  where  $C = A \wedge B$ . For the interval  $I = [i, j]$  with  $1 \leq i \leq j \leq l$ , define  $I_C = \{c_k \mid i \leq k \leq j\}$  is the set of fragments obtained from  $c_i$  to  $c_j$ . The cassette defined by  $I_C$  is the pair of fragments  $(I_A, I_B)$ , where  $I_A$  and  $I_B$  are the set of fragments of  $A$  and  $B$  respectively that contain a fragment from  $I_C$ .

## Example

$A = \{1, 4, 3, 5, 2, 4, 5, 3\}$  ( $m = 8$ ),  $B = \{3, 1, 3, 2, 8, 3, 7\}$  ( $n = 7$ ) and

$A \wedge B = \{1, 2, 1, 1, 2, 1, 1, 4, 2, 2, 2, 1, 4, 3\}$  ( $l = 14$ ). Let  $I = [3, 7]$ . Then  $I_A = \{4, 3, 5\}$  and  $I_B = \{1, 3, 2\}$ .

## Boundary of Cassettes

- For the **cassette** defined by  $I_C : (I_A, I_B)$ , let  $m_a$  and  $m_B$  be the starting positions of the leftmost fragments of  $I_A$  and  $I_B$  respectively,  $n_A$  and  $n_B$  be ending positions of the rightmost fragments of  $I_A$  and  $I_B$ .
  - For example, in the previous page  $m_a - m_B = 1 - 3 = -2$  and  $n_A - n_B = 13 - 9 = 4$ . (See it?)
  - *For the solutions mentioned in page 6 and 7, we are able to find that the cassette defined on  $I_C$   $I = [3, 7]$  (in Sol. 1) is exactly the same as the cassette defined on  $I_C$   $I = [8, 12]$  (in Sol. 2) .*
- (\*) In fact, the Sol. 2 can be obtained from Sol. 1 via cassettes exchange.

## *Comments on Cassette Exchange*

- There are solutions for DDP which can not be transformed each other.
- DDP problem is one of the most important one in DNA sequencing.

### Reference

Schmitt and Waterman, Multiple solutions of DNA restriction mapping problem, *Advances in Applied Math.* 12(1991), 412 – 427.

### Problem

How to transform one map into another by cassette transformation.

# Alternating Eulerian Circuits

## Theorem (Kotzig, 1968)

Let  $G$  be an edge-colored connected even graph. Then,  $G$  has an alternating Eulerian circuit if and only if  $G$  is balanced.

## Corollary

If  $G$  is a bicolored connected balanced even graph, then  $G$  contains an alternating Eulerian circuit.

## Definition

Order exchange:  $F_1 F_2 F_3 F_4 F_5 \dots \longrightarrow F_1 F_4 F_3 F_2 F_5 \dots$

Order reflection:  $F_1 F_2 F_3 F_4 F_5 \dots \longrightarrow F_1 F_2^{-1} F_3 F_4 F_5 \dots$  (Reverse  $F_2$ )

# Important Result

## Theorem

*Every two alternating Eulerian circuits in a bicolored graph  $G$  can be transformed into each other by a series of order transformations.*

## Definition (Fork)

*Let  $A = \{a_1, a_2, \dots, a_m\}$ ,  $B = \{b_1, b_2, \dots, b_n\}$  and  $C = \{c_1, c_2, \dots, c_l\}$  where  $C = A \wedge B$ . Let the corresponding fragments of  $A$ ,  $B$  and  $C$  be  $\{A_1, A_2, \dots, A_m\}$ ,  $B = \{B_1, B_2, \dots, B_n\}$  and  $\{C_1, C_2, \dots, C_l\}$ . A fork of fragment  $A_i$  (resp.  $B_j$ ),  $F(A_i)$  (resp.  $F(B_j)$ ), is the set of double digested fragments  $C_k$  contained in  $A_i$  (resp.  $B_j$ ):  $F(A_i) = \{C_k \mid C_k \text{ is contained in } A_i\}$ ,  $F(B_j) = \{C_k \mid C_k \text{ is contained in } B_j\}$ .*

*Notice that any two forks contain at most one common fragment of  $C_k$ 's.*

# Fork Graph

By using the length of fragments in  $C$ , we are able to define a fork graph accordingly. That is if there are  $h$  different length of the fragments in  $C$ , then the vertex set of the fork graph is  $\{l_1, l_2, \dots, l_h\}$ . The arc set has  $l$  arcs, starting from the leftmost fragment of either  $A$  or  $B$ . Moreover, we use **red** lines if  $A_i$ 's are used, **blue** lines for  $B_j$ 's.

**Fact 1. Every physical map  $(A, B)$  defines an alternating Eulerian path in its fork graph.**

**Fact 2. Cassette transformations of physical maps correspond to order transformations in the fork graph.**

# Conclusion of Cassette Transformations

## Theorem

*Every alternating Eulerian path in the fork graph of  $(A, B)$  corresponds to a map obtained from  $(A, B)$  by cassette transformations.*

## 3-partition problem

*Given  $3n$  integers  $a_1, a_2, \dots, a_{3n}$  and an integer  $h$  such that the sum of the above  $3n$  integers is  $h \cdot n$  and  $h/4 < a_i < h/2$ , can we partition  $\{a_1, a_2, \dots, a_{3n}\}$  into  $n$  triples such that each triple has a sum  $h$ ?*

## *Examples of 3-partition*

Possible case:  $h = 16$ ,  $n = 6$  and  $\{5^{12}, 6^6\}$ .

Impossible case:  $h = 16$ ,  $n = 6$  and  $\{5^{14}, 6^2, 7^2\}$ .

### Theorem (Known result)

*3-partition problem is of strong NP-completeness.*

### Theorem

*DDP is strongly NP-complete.*

Proof. Let  $A = \{a_1, a_2, \dots, a_{3n}\}$  and  $B = \{b_j = h \mid j = 1, 2, \dots, n\}$  where  $\sum A = h \cdot n$ .  
Now, the solution for DDP in the case  $C = A$  is equivalent to find a 3-partition of  $A$ . This completes the proof. ■

## *Errors occurred in DDP*

There are essentially four types of errors.

1. Partial cleavage: an enzyme can fail to cut at some restriction site.
2. Fragment length: determining the exact length of a fragment from “gel electrophoresis” is not easy.
3. Missing fragments especially the small length fragments.
4. Doublets: two different fragments with almost the same length may generate two spots in the “gel electrophoresis” that overlap. Thus, only one of the fragments is recognized.

# *Research Problem*

## Maximum 4-partition problem

Given a multiset  $Q = \{q_1, q_2, \dots, q_{4n}\}$  of  $4n$  integers such that their sum is  $n \cdot h$  and  $h/5 < q_i < h/3$ ,  $i = 1, 2, 3, \dots, 4n$ . Find a maximum number of disjoint quadruples (4-subsets) in  $Q$  such that each quadruple has a sum  $h$ .

(\*) This approach can handle the cases with errors. (?)