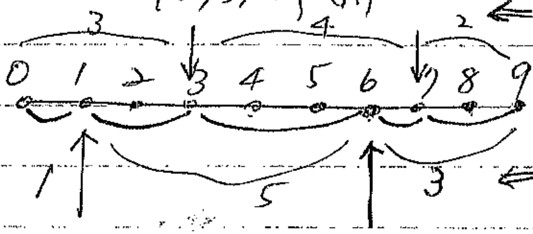
Lecture 9 Double Digest Problem (DDP)

{2, 3, 4} (A)



← Enzyme A

← Enzyme B.

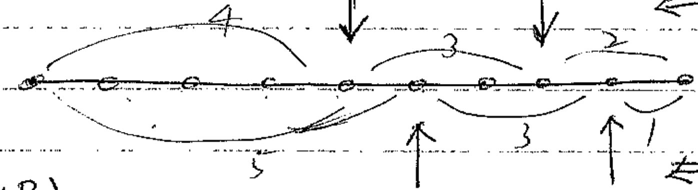
collect the distance {1,1,2,2,3} (A+B).

{1, 3, 5} (A)

(分段) 与 PDP 不同

The above figure shows the physical map of two restriction enzymes. (Map 1)

{2, 3, 4} (A)



← Enzyme A

collect the distance → {1, 1, 1, 2, 4} (A+B)

← Enzyme B.

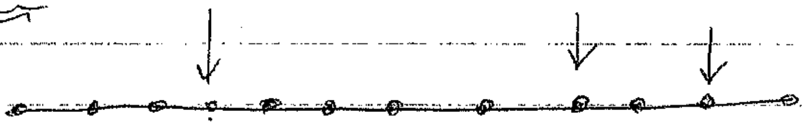
{1, 3, 5} (A)

(Map 2)

(*) Clearly, these two maps have different restriction sites.

Problem Given A, B and A+B, find a physical map.

(Note) In DDP, we only collect those distance from adjacent sites.



{3, 5, 2, 1} (four distances for three sites)

As mentioned in PDP, the solutions for DDP are also not unique in general.

Here is an example with given A, B and A+B.

6 1's, 5 2's

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

(•) $\sum_{x\in A} x = \sum_{y\in B} y = \sum_{z\in A+B} z = 27$

(••) We can use 28-dim. vectors (Binary) to represent the maps for A, B and A+B respectively.

Sol.(1)

End

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

$\vec{B} = (1,0,1,1,0,0,1,0,1,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,1)$

End

$\vec{A+B} = (1,1,0,1,1,1,0,1,1,1,0,0,0,1,0,1,0,1,0,1,1,0,0,0,1,0,0,1)$

Representing by supports (in Z_8)

$\begin{cases} 0+1=1 \\ 1+0=1 \\ 0+0=0 \\ 1+1=1 \end{cases}$

$\{A\} = \{0,1,5,8,13,15,19,24,27\}$

$\{B\} = \{0,3,4,7,9,12,20,27\}$

$\{A+B\} = \{0,1,3,4,5,7,8,9,13,15,17,19,20,24\}$

$\| \quad A \cup B$

Date / /

No. 3

Sol. (2)

$\{A\} = \{0, 1, 5, 10, 12, 16, 19, 24, 27\}$, $A = \{1, 2, 3, 3, 4, 4, 5, 5\}$

$\{B\} = \{0, 3, 6, 14, 15, 18, 20, 27\}$, $B = \{1, 2, 3, 3, 3, 2, 8\}$

$\{A+B\} = \{0, 1, 3, 5, 6, 10, 12, 14, 15, 16, 18, 19, 20, 24, 27\}$

$A+B = \{1^6, 2^5, 3, 4^2\}$.

(•) These two solutions are not symmetric.

Definition (Similar solutions)

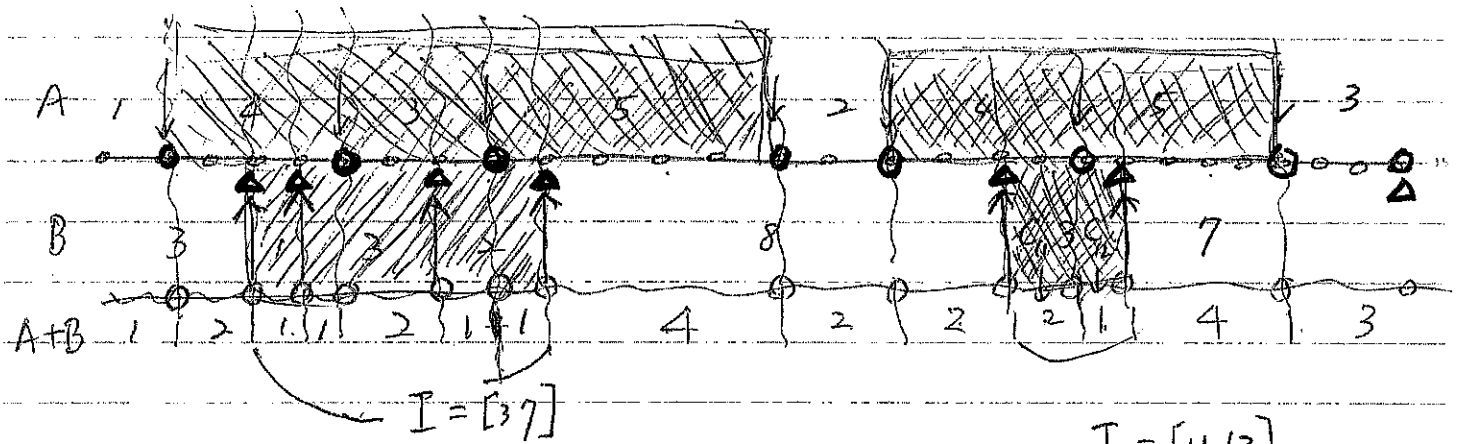
Two solutions of DDP are similar if one solution can be obtained from transforming another solution to this solution via suitably defined "transformations".

Definition (Cassette)

Let $A = \{A_1, A_2, \cdots, A_m\}$, $B = \{B_1, B_2, \cdots, B_m\}$ and $C = \{C_1, C_2, \cdots, C_l\}$ where $C = A + B$. For an interval $I = [i, j]$ with $1 \le i \le j \le l$, define $I_C = \{C_k \mid i \le k \le j\}$ as the set of fragments between C_i and C_j.

The cassette defined by I_C is the pair of sets of fragments

(I_A, I_B), where I_A and I_B are the set of all fragments of A and B respectively that contain a fragment from I_C. Let m_A and m_B be the starting positions of the leftmost fragments (n_A, n_B) (rightmost) of I_A and I_B respectively. The left overlap of (I_A, I_B) is the (right) distance $m_A - m_B$ (resp. $n_A - n_B$).



I = [11, 12]
Cassette defined on $\{C_{11}, C_{12}\}$.

$A = \{1, 4, 3, 5, 2, 4, 5, 3\}$ $m = 8$

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

$A+B = \{1, 2, 1, 1, 2, 1, 1, 4, 2, 2, 2, 1, 4, 3\}$ $l = 14$

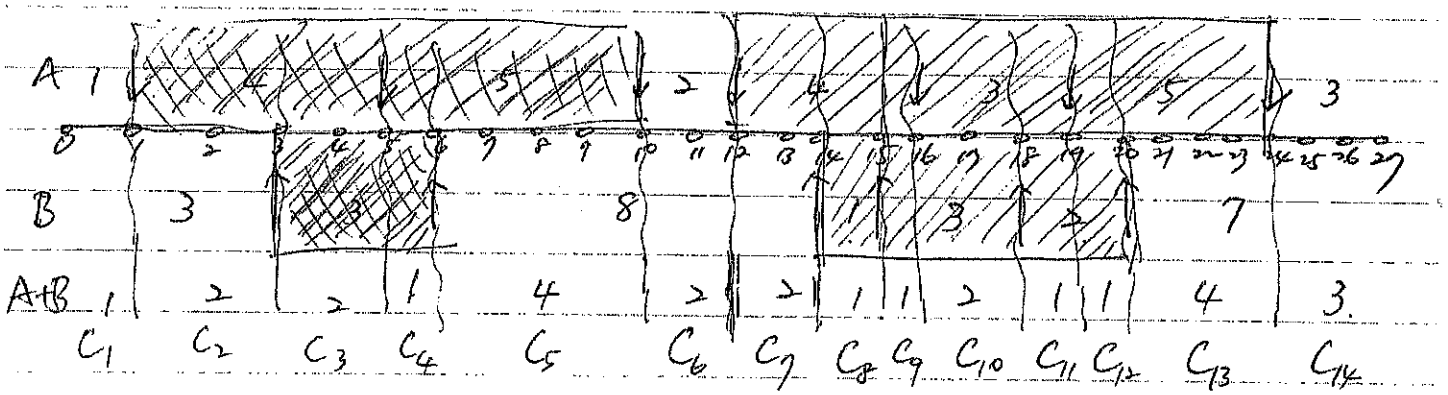
$I = [3, 7]$, 求得部份是 Cassette defined by $I_C = \{C_3, C_4, \cdots, C_7\}$.

$I_A = \{4, 3, 5\}$, $I_B = \{1, 3, 2\}$, $m_A - m_B = 1 - 3 = -2$, $n_A - n_B = 13 - 9 = 4$.

Date / /

No. 5

Sol.(2)



$I = [8, 12]$

The cassette defined on I_C, $I = [8, 12]$ in Sol.(2) is exactly the same as the cassette defined on I_C, $I = [3, 7]$ in Sol.(1).

Similarly, for $I = [3, 4]$ (Sol.(2)), it is the same as the cassette defined on I_C, $I = [11, 12]$ in Sol.(1).

(*) We obtain Sol.(2) by using cassettes exchange from Sol.(1).

(•) Under the condition that (a) they must have the same left overlap and right overlap and (b) they can not have fragments in common.

$\{A\} = \{1, 5, 8, 13, 15, 19, 24\}$

$\{B\} = \{3, 4, 7, 9, 17, 20\}$) Sol. (1)

+14 −11

$\{A\} = \{1, 5, 10, 12, 16, 19, 24\}$

$\{B\} = \{3, 6, 14, 15, 18, 20, 27\}$) Sol. (2)

Using exchanges: Sol. (2) → Sol (1)

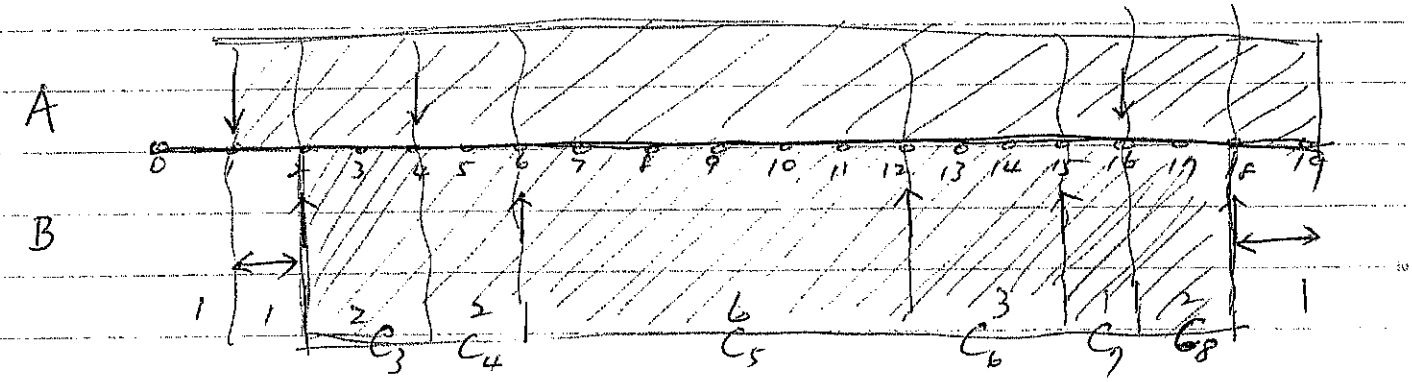
$\{1, 5, 8, 13, 15, 19, 24\}$.

Fact Many solutions can be obtained by using cassettes exchange and also the following cassettes reflection.

$A = \{1, 3, 12, 3\}$ $\quad I = [3, 8]$

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

$A+B = \{1, 1, 2, 2, 6, 3, 1, 2, 1\}$



$m_A - m_B = -1$, $n_A - n_B = 1$ (left overlap = − right overlap)

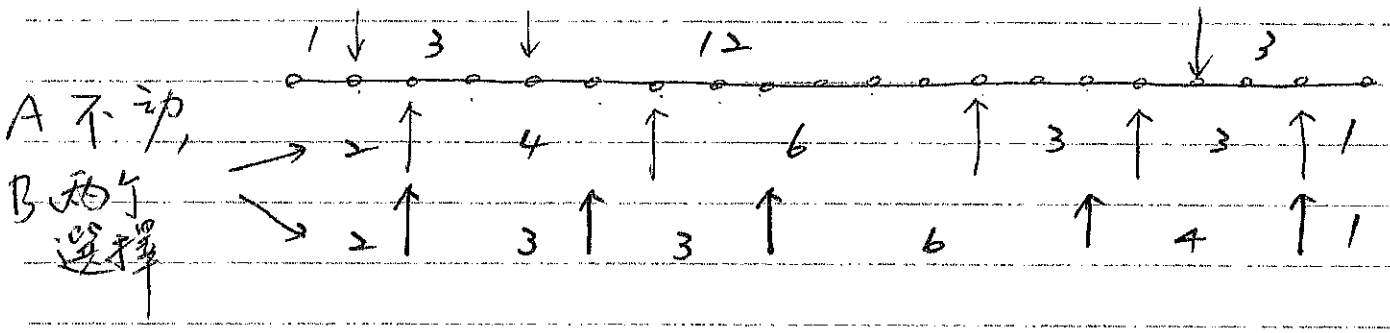
After reflection

A remains the same,

$B = \{2, 3, 3, 6, 4, 1\}$, and $A+B = \{1, 1, 2, 1, 3, 6, 2, 2, 1\}$.

(*) Not all DDP solutions can be transformed into one another by cassette exchange.

e.g. Two solutions for $A = \{1, 3, 3, 12\}$, $B = \{1, 2, 3, 3, 4, 6\}$ and $A+B = \{1, 2, 3, 6\}$ are as follows. (Mentioned above!)



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

How to transform one map into another by cassette transformations?

Review

(•) An k-edge-coloring π of G is a mapping $\pi: E(G) = \{1, 2, \cdots, k\}$.

For convenience, we call G is π-colored, and the number of colors c used around a vertex v is denoted by $d_c(v)$.

(•) If $d_\pi(v) = \max_c d_c(v) \leq \frac{1}{2} d(v)$ for each $v \in V(G)$, (c is a color used in π) then the π-colored graph G is balanced. (用最多次的顏色不超過 degree 的一半!)

Theorem (Kotzig, 1968)

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

Proof. It suffices to pair off the edges (colors) incident to each vertex v. This result is a direct consequence of balanced coloring. (See 8')

Theorem

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

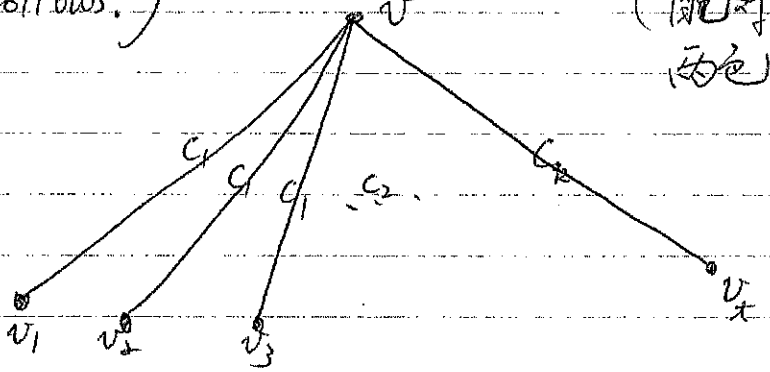
Proof. Since G is bicolored and balanced, $d_1(v) = d_2(v)$ for each vertex $v \in V(G)$ where $d_i(v)$ is the number of edges incident to v which are colored i, $i = 1, 2$.

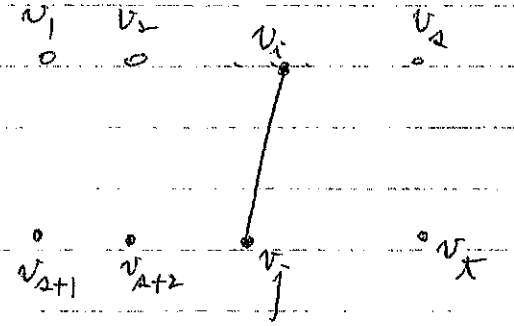
... Let $c_1, c_2, \cdots, c_k$ be the colors occur around the vertex v. For convenience, let $d_{c_i}(v)$ be the number of edges incident to v which are of color c_i and $d_{c_1}(v) \geq d_{c_2}(v) \geq \cdots \geq d_{c_k}(v)$.

Clearly, $t = \sum_{i=1}^{k} d_{c_i}(v) = d(v)$ (degree of v). (See Figure)

Now, let $s = \lceil \frac{t}{2} \rceil = \frac{t}{2}$. ($t$ is even) and $\tilde{G}$ be a bipartite graph with $V(\tilde{G}) = \{v_1, v_2, \cdots, v_{t/2}\} \cup \{v_{s+1}, v_{s+2}, \cdots, v_t\}$ and $v_i v_j$ is an edge if $c_i \neq c_j$. It is not difficult to see that $\tilde{G}$ has a perfect matching (?) since $\max_{i=1}^{k} d_{c_i} \leq s$. (Pair off the vertex v_i with the vertex v_k where vv_k is of color c_i and the process follows.)

(配对出现最多次数的颜色!)



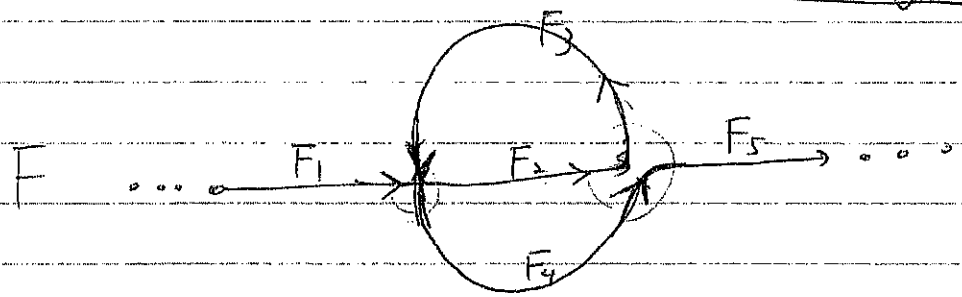


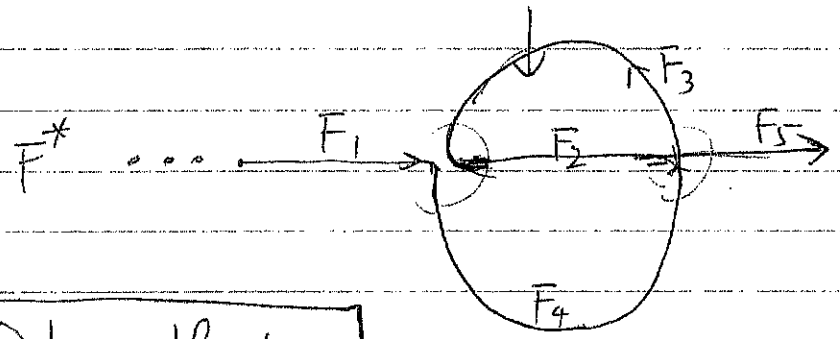
Mathematical Model of Cassette Transformations

(a) Order transformations of alternating paths (2-colored G).

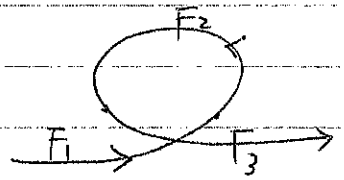
$F = \cdots F_1 F_2 F_3 F_4 F_5 \cdots \longrightarrow F^* = \cdots F_1 F_4 F_3 F_2 F_5 \cdots$

five consecutive paths → **order exchange**



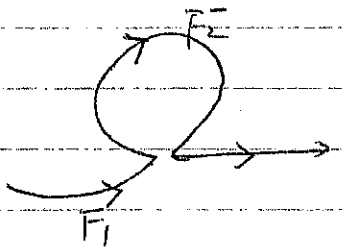


Order reflection



$F = \cdots F_1 F_2 F_3 \cdots$

$F^* = \cdots F_1 F_2 F_3 \cdots$



→ Tedious (but true)

Theorem Every two alternating eulerian circuits in a bicolored graph G can be transformed into each other by a series of order transformations (exchanges and reflections)

The idea of "Fork Graph" (Why fork?)

(•) For simplicity, we assume that digests A and B do not cut DNA at the same position.

(•) $A = \{A_1, A_2, \cdots, A_m\}$, $B = \{B_1, B_2, \cdots, B_n\}$ and $A+B = \{C_1, C_2, \cdots, C_l\}$.

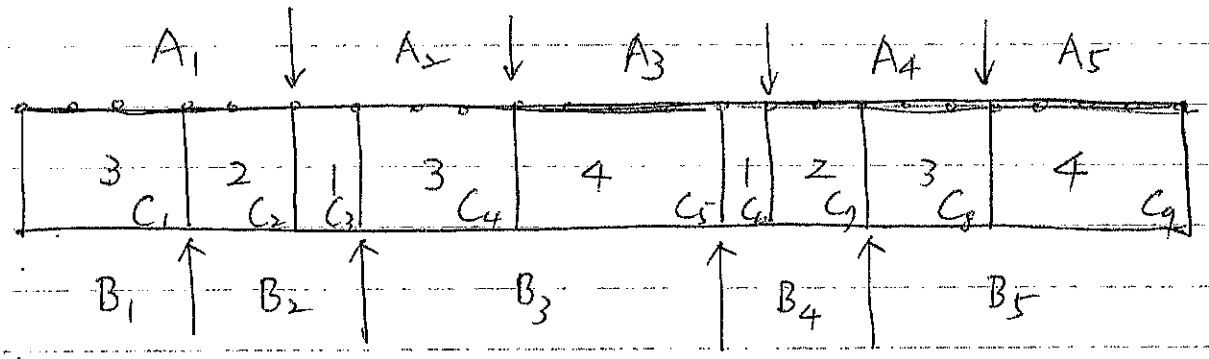
$l = m+n-1$

Definition (Fork)

A fork of fragment A_i (resp. B_j) is the set of double digest fragments C_k contained in A_i (resp. B_j):

$$F(A_i) = \{C_k \mid C_k \subseteq A_i,\ 1 \le k \le l\}.$$

$$(F(B_j) = \{C_h \mid C_h \subseteq B_j,\ 1 \le h \le l\}.)$$



e.g. $F(A_3) = \{C_5, C_6\}$, $F(B_4) = \{C_6, C_7\}$, $F(A_3) \cap F(B_4) = \{C_6\}$.

border fragment

(*) Any two forks contain at most one common fragment!

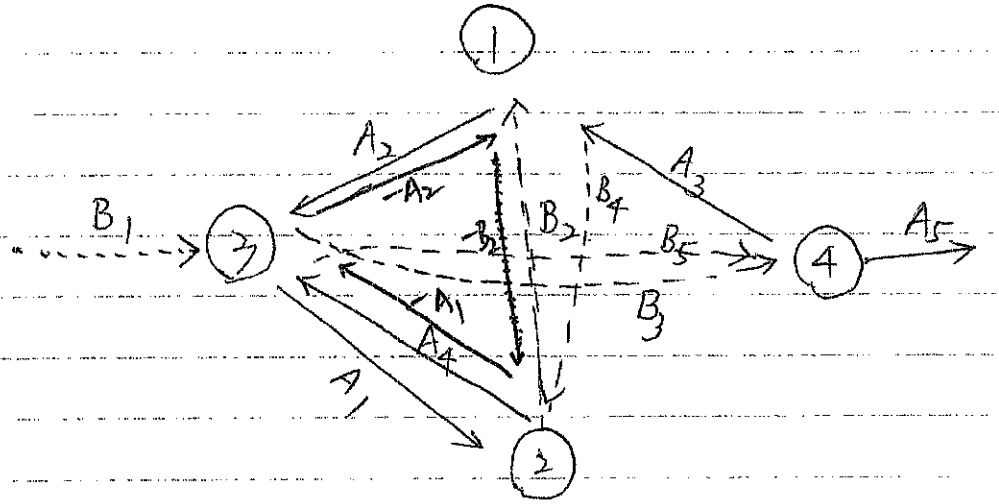
Border fragments : C_1 and C_2 ; and the fragment contained in two forks $F(A_i)$ and $F(B_j)$.

e.g. C_2 is the border fragment of A_1 and B_2,

C_3 is the border fragment of A_2 and B_2.

(°) We can use the lengths of border fragments to define a graph : fork graph.

(°°) Example The DDP in page 10 has four different lengths of fragments. Hence, the fork graph of the DDP has vertex set $\{1, 2, 3, 4\}$.



Black lines for A and dot lines for B.

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

Fact 2 Cassette transformations of a physical map do not change the set of forks of this map.

(*) Cassette transformations of physical maps correspond to order transformations in the fork graph. (?)

Theorem (**)

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

以下我们将会说明 DDP 是一个困难的问题。

3-partition problem

Given $3n$ integers $a_1, a_2, \cdots, a_{3n}$ and an integer h with $\sum_{i=1}^{3n} a_i = nh$ and $\frac{h}{4} < a_i < \frac{h}{2}$, can we partition $\{a_1, a_2, \cdots, a_{3n}\}$ into n triples such that each triple has a sum h?

For example, let $h = 16$, $n = 6$, and $\{5^{12}, 6^6\}$. (Yes!)

But, $\{5^{14}, 6^2, 7^2\}$ is not possible.

(Here, $4 \le a_i < 8 \Rightarrow a_i \in \{5, 6, 7\}$.

(*) 3-Partition problem is of strong NP-completeness.

(NP-complete in the strong sense!)

Theorem Double Digest Problem (DDP) is strongly NP-complete.

Proof. Let $A = \{a_i \mid 1 \le i \le 3n\}$ and $B = \{b_j = h \mid 1 \le h \le n\}$ where $\sum_{i=1}^{3n} a_i = hn$. Now, the solution for DDP in case that $C = A$ is equivalent to find a 3-partition of A. Therefore, we can reduce 3-partition to DDP. By the fact that 3-partition problem is NP-complete, so is DDP. ∎