Lecture 11, Map Assembly

Date / / No. 10

(•) Physical mapping can be understood in terms of the following analogy:

1. Several copies (clones) of a book cut by scissors into thousand of pieces. (Break the DNA molecule into small pieces.)

2. Each copy is cut in a individual way such that a piece from one copy may overlap a piece from another copy.

↓ Cloning incorporates a fragment of DNA into some self-replicating host.

3. For each piece and each word from a list of key words (fingerprinting) we are told whether the piece contains the key word.

4. Given this data, we wish to determine the patterns of overlaps of the pieces.

↑ clone library

(•) After a clone library is constructed, we are aimming at reconstructing the relative placements of the clones along the DNA molecule. Some infos may have lost, so, the reconstruction starts with the fingerprinting of the clones.

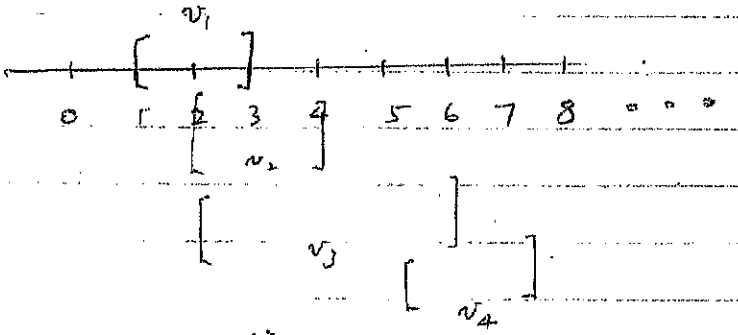
Idea from Graph Theory

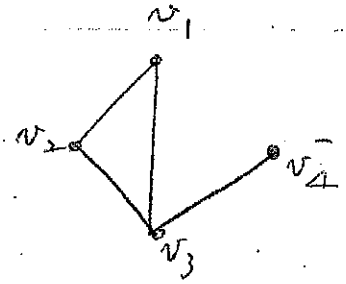
Interval Graphs

An interval graph $G=(V,E)$ is the graph where $V(G)$ is a set of closed intervals in real line and two vertices are adjacent iff their corresponding intervals have non-empty intersection.

For example:

$$V = \{[1,3],[2,4],[2,6],[5,7]\}$$





G

Lemma 1 Every interval graph is triangulated, i.e., every cycle of length longer than 3 has a "chord". (Chordal Graph)

Proof. Let C be a cycle of G with $|V(G)| \geq 4$ and

C has no chords. Let $C = (u_1, u_2, \cdots, u_k)$, $k \geq 4$. W.L.O.G assume that we may $a_1 \leq a_2 \leq \cdots \leq a_k$ where $u_i = [a_i, b_i]$, $i = 1, 2, \cdots, k$.

(If $a_3 < a_2$, then $u_1 \sim u_3$ since $b_1 \geq a_2 > a_3$.)

Since $u_1 \sim_G u_k$, $a_k \leq b_1$. Now, if $b_1 \leq b_j$ for some $j = 2, \cdots, k-2$, then $a_k \leq b_1 \leq b_j$ and therefore $u_j \sim_G u_k$, $\rightarrow\leftarrow$. Hence, $b_1 > b_j$ $\forall$ $j = 2, \cdots, k-2$. ($b_1 \geq a_j$ $\wedge$) This implies $u_1 \sim u_j$, $\rightarrow\leftarrow$. ∎

Definition (Transitive orientation property)

$G = (V, E)$

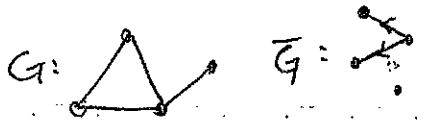
An undirected graph has a transitive orientation property if G has an orientation (on E) such that the resulting digraph $D_G(V, A)$ satisfies $\forall\, a, b, c \in V$, $(a, b) \in A$ and $(b, c) \in A$ (whenever exist) imply $(a, c) \in A$. (G is said to be transitively orientable)

(If G contains no K_3's, then G is transitively orientable.)

Definition (Comparability graph)

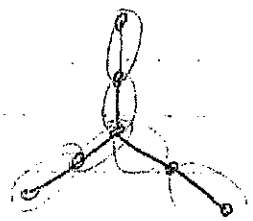
An undirected graph that is transitively orientable is called a comparability graph.

Lemma 2 The complement of an interval graph is a comparability graph.

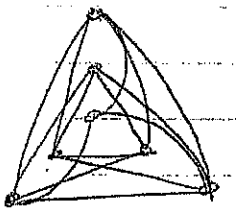


Proof. Let $\bar{G}$ be the complement of an interval graph G. Then, $[i,j]$ and $[i',j']$ of $V(G)$ are adjacent in $\bar{G}$ if and only if $[i,j] \cap [i',j'] = \phi$. Now, define $[i,j] \to [i',j']$ (orientation) provided $j < i'$. Now, it is easy to check if $[i_1,j_1] \to [i_2,j_2]$ and $[i_2,j_2] \to [i_3,j_3]$, then $[i_1,j_1] \to [i_3,j_3]$ since $i_3 > j_2 \geq i_2 > j_1$. ∎

Example Not all graphs possess this property!



G



$\tilde{G}$

Not a comparability graph

(?)

Theorem 3 (Gilmore and Hoffman, 1964) Let G be an undirected graph. The following statements are equivalent:

(1) G is an interval graph.

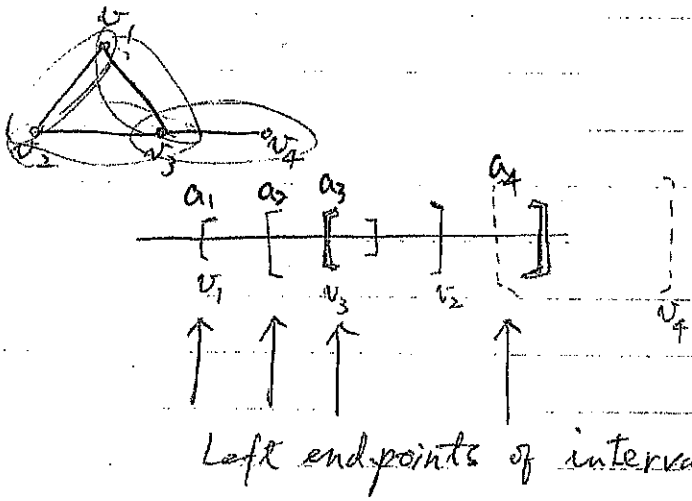
(2) G is a triangulated graph and its complement is a comparability graph.

(3) The maximal cliques of G can be linearly ordered s.t. $\forall\, x \in V(G)$, the maximal cliques containing x occur consecutively.

Proof. (1) $\Rightarrow$ (2) (By Lemma 1, 2.)

(2) $\Rightarrow$ (3) complicate.

(3) $\Rightarrow$ (1) $\forall x \in V(G)$, let $I(x)$ denote the set of all maximal cliques of G that contains x. The sets $I(x)$ form the intervals of the interval graph.



(*) Theorem 3 reduces the problem of recognition of an interval graph to the problems of recognition of Triangulated and comparability graphs (Fulkerson and Gross, 1965; Pnueli et al., 1971).

Mapping with Restriction Fragment Fingerprints

The simplest case of mapping with restriction fragment fingerprints is Single Complete Digest mapping (SCD mapping). (Olson et al. 1986)

In this case the fingerprint of a clone is a multiset of the sizes of its restriction fragments in a digest by a restriction enzyme. An SCD map is a placement of clones and restriction fragments consistent with the given SCD data (Gillett et al. 1996)

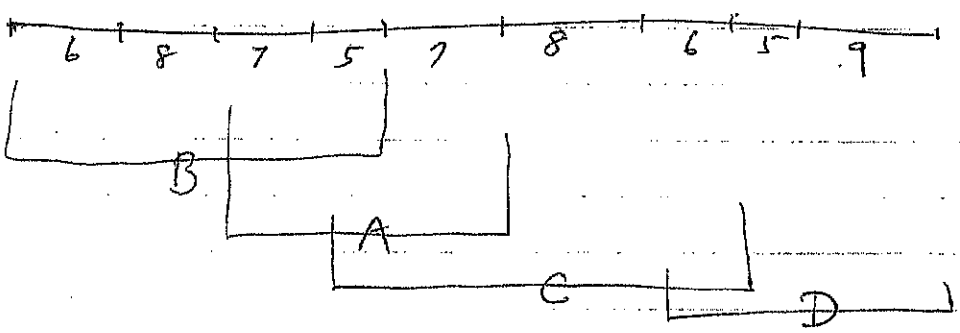
SCD Mapping Problem

Find a most compact map (i.e., a map with the minimum number of restriction fragments) that is consistent with SCD data.

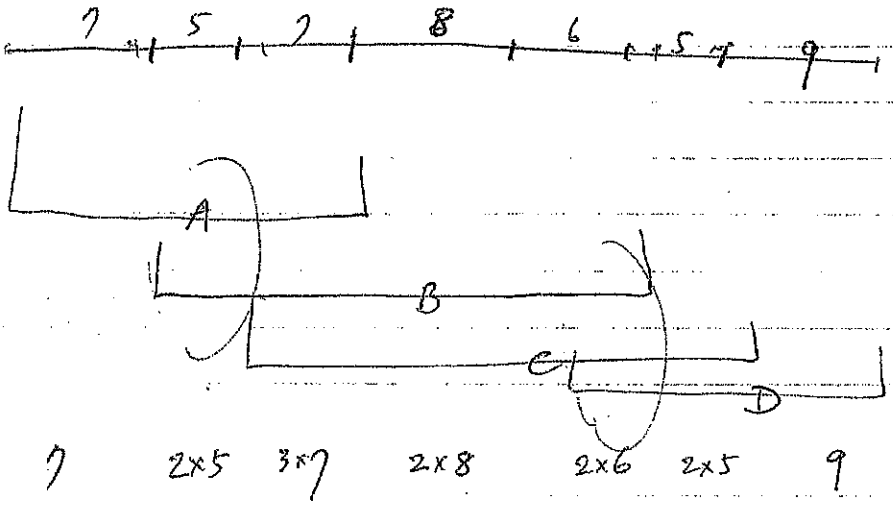
$A = \{5, 7, 7\}$, $B = \{5, 6, 7, 8\}$, $C = \{5, 6, 7, 8\}$, $D = \{5, 6, 9\}$

Solutions

①



②



7 2×5 3×7 2×8 2×6 2×5 9

Solution ② is better which has 7 restriction fragments.

(*) The problem of finding the most compact map is NP-hard.

(**) 在实际的应用上，"Fingerprints of clones" 通常可以用统计的结果来估计 clones 未识的顺序 (Ordering)。

(**)

Jiang and Karp (1998) formulated SCD mapping with known clone ordering as a constrained path cover problem on a special multistage graph.

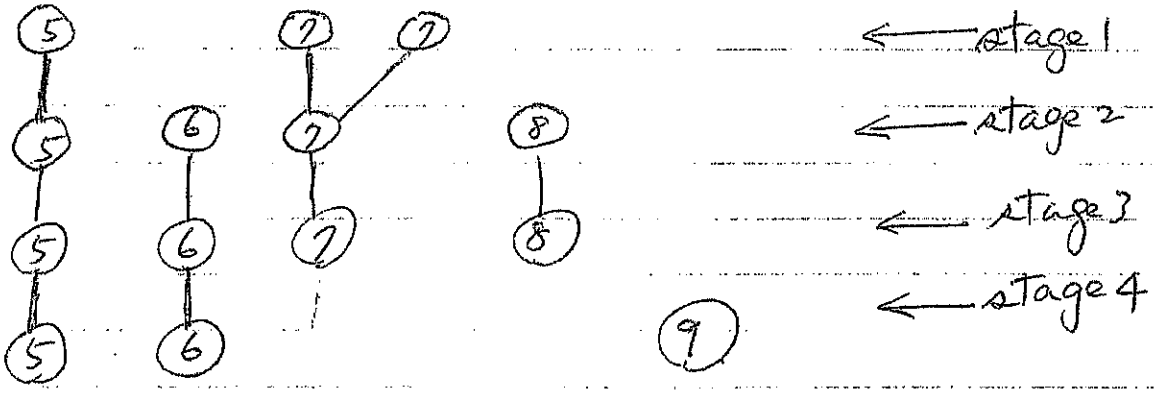
Let $S = \{S_1, S_2, \ldots, S_n\}$ be an instance of SCD mapping, where S_i is a multiset representing the fingerprint of the i-th clone in the clone ordering by the left endpoints.

Definition (labeled multistage graph)

A labeled multistage graph G (called a clone-fragment graph) consists of n stages, with the i-th stage containing $|S_i|$ vertices. At stage i, G has a vertex for each element x of S_i (including duplicates), with label x. Two vertices are adjacent if they are at adjacent stages and have identical label.

Example.

$S_1 = \{5, 7, 7\}$, $S_2 = \{5, 6, 7, 8\}$, $S_3 = \{5, 6, 7, 8\}$, $S_4 = \{5, 6, 9\}$



$$\begin{bmatrix} 5 & 0 & 7 & 7 & 0 & 0 \\ 5 & 6 & 7 & 0 & 8 & 0 \\ 5 & 6 & 7 & 0 & 8 & 0 \\ 5 & 6 & 0 & 0 & 0 & 9 \end{bmatrix}$$

Definition (Path cover)

A path cover is a collection of paths such that every vertex is contained in exactly one path.

(*) Any map for S corresponds to a path cover of G of the same cardinality.

From ②, we see that the map gives a path cover with "7" paths and they are "7", "2×5", "3×7", "2×8", "2×6", "2×5" and 9.

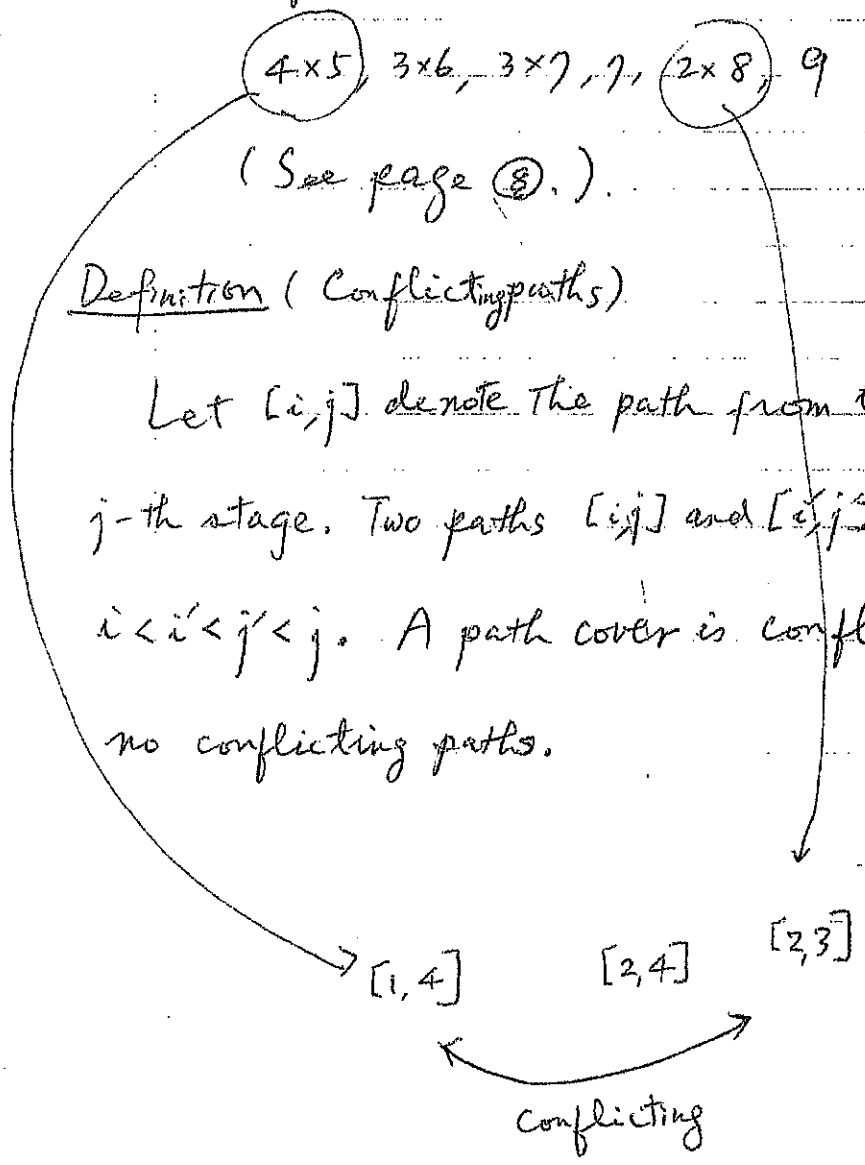
(**) We may have a better path cover, but it corresponds to no map! For example

4×5, 3×6, 3×7, 7, 2×8, 9

(See page ⑧.)

Definition (Conflicting paths)

Let [i,j] denote the path from the i-th stage to the j-th stage. Two paths [i,j] and [i',j'] are conflicting if $i < i' < j' < j$. A path cover is conflict-free if it has no conflicting paths.

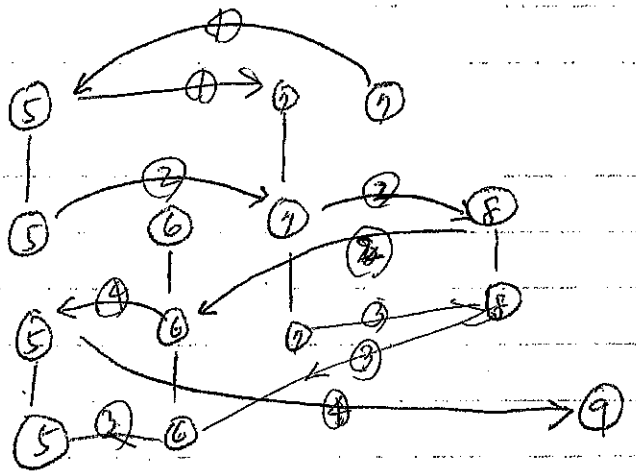


• The cardinality of a map of S is the # of fragments.

Definition (consistent path cover)

A path cover of G is consistent if it corresponds to a map of S with the same cardinality.

Example, Solution ② corresponds to the following path cover:



"7 paths" ⟷ 7 fragments

Proposition A path cover is consistent if and only if it is conflict-free. (A set of intervals)

Proof. Derived from the fact of interval graphs. (?)

(*) Jiang and Karp (1998) obtained a 2-approximation algorithm for SCD mapping with a given ordering of clones.

(*) In order to get the correct map, we need signals.

Finding Signals in DNA

(11)

- The first signal in DNA was found in 1970 by Hamilton Smith after the discovery of the Hind II restriction enzyme.
- The palindromic site of the restriction enzyme is a signal that initiates DNA cutting.

(Fact) 1. Hamilton Smith was lucky: restriction sites are the simplest signals in DNA. (Reliably find in DNA.)

2. Most other signals (promoters, splicing sites, etc.) are so complicated that we don't yet have good models or reliable algorithms for their recognition.

(**) Understanding gene regulation is a major challenge in computational biology. For example, regulation of gene expression may involve a protein binding to a region of DNA to affect transcription of an adjacent gene.

(Fact) Since protein–DNA binding mechanisms are still insufficiently understood to allow "in silico" prediction of binding sites, the common experimental approach is to locate the approximate position of the binding site.

(***) These experiments usually lead to identification of a DNA fragment of length n that contains a binding site (an unknown magic word) of length $l \ll n$.

•• Of course, one experiment is insufficient for finding the binding site, but a sample of experimentally found DNA fragments gives one hope of recovering the magic word.

GATTCTTAGGC
TATACGTTCGA
TGATTGACTTC

Problem

Given a sample of K sequences where an (unknown) magic word appears at different (unknown) positions in these K sequences. Find the magic word!

Common Sense Approach (Staden, 1989; Wolfertstetter et.al, 1996; Tompa, 1999)

Since $l \ll n$, we may test all words of length l and find those that appear in all (or almost all) K sequences.

Note l 可能也不知道有多長；所以可從較小的長度測試再增長；一直到有"唯一"未現的 Magic word 為止。

① The described approach usually works fine for short continuous words such as GAATTC the restriction site of EcoRI.

② In the idea of randomness, this is also quite difficult, since we have 4 (6) words to compete for the magic word.

③ The problem gets even more difficult (complicated) whe the magic word has gaps.

For example,

$CCAN_qTGG$ (Xcm I restriction enzyme)

↑ a gap of length q

$(Pu)^m C N_{40-2000} (Pu)^m C$ (McrBC Endonuclease)

↑ A or G

$TTGACAN_{17}TATAAT$ (E. Coli promoters)

↑ a gap of length 17

Enumeration and check of all patterns of the above types are hardly possibly due to computational complexity.

Remark

寻找 Magic Words 基本上就像密码学中寻找 Key 的概念，除了利用各种可能提供的讯息之外，还要发挥一些想像力来协助判断；把所有可能的答案都试一次绝非良策。

• DNA linguistics is at the heart of the pattern-driven approach to signal finding, which is based on enumerating all possible patterns and choosing the most frequent or the fittest among them.

(*) The fitness measures vary from estimates of the statistical significance of discovered signals to the information content of the fragments that approximately match the signal.

Steps for pattern-driven approach

Step 1. Define the frequency or fitness measure. (f.f.m.)

Step 2. Calculate the f.f.m. of each word w.r.t. a sample DNA fragments.

Step 3. Report the fittest words as potential signals.

(Fact)

Note that if A denotes the set of alphabets, then the search space for patterns of length l is $|A|^l$.

⊙ DNA texts are not easy to decipher, and there is little doubt that Nature can construct an "enigma" (謎) of the kind which human ingenuity (創造力) may not resolve.

• A popular approach in DNA linguistics is based on the assumption that frequent or rare words may correspond to signals in DNA.

• A word occurs considerably more (or less) frequently than expected has the potential to become a "signal".

↓

What is its biological meaning?

For example, Gelfand and Koonin (1997) showed that the most avoided 6-palindrome in the archaeon M. jannaschii is likely to be the recognition site of a restriction-modification system.

(W)

◎ To find frequent and rare words in a text, one has to compute expected value $E(W)$ and the variance $\sigma^2(W)$ for the number of occurrences (frequency) of each word W.

◎ Afterwards, the frequent and rare words are identified as the words with significant deviations from expected frequencies.

Remark

In many DNA linguistics papers, the variance $\sigma^2(W)$ of the number of occurrences of a word in a text was erroneously assumed to be $E(W)$.

要探討 word's occurrence 的和率就不是一件直截(或簡單)的事。參考 "Correlation Polynomial" (Autocorelation)