Lecture 11

DDP - continued

Proof. (DDP is NP-complete)

It suffices to prove that an instance of DDP is equivalent to solve the 3-partition problem, also a special case.

Let $A = \{a_i \mid i \in [1,3n]\}$, $B = \{b_i = h \mid i \in [1,n]\}$, and $A = C$. Assume that $a_i = c_i = q_i$ where $\frac{h}{4} < q_i < \frac{h}{2}$ and $\sum_{i=1}^{3n} q_i = hn$.

Now, if there is a solution for the 3-partition problem, i.e., there exist n disjoint triples of q_i's (and a_i's as well) such that each triple sums up to h, then starting from 0, we arrange the distances a_i on a line satisfying each three a_i's belong to the same triple are adjacent. This order is also used for c_i's. So, clearly, we have a solution for DDP.

On the other hand, if there is a solution for DDP, then by selecting three consecutive coordinates (sum up to h),

Disjoint DDP

(*) Two enzymes cut at disjoint restriction sites. ($A \cap B = \emptyset$.)

(*) DDDP is NP-complete

Proof. We show that if there is a solution for 3-partition problem on a special instance, then a suitably designed DDDP can be solved.

Let $\frac{h}{4} < q_i < \frac{h}{2}$, $i = 1, 2, \cdots, 3n$, $\sum_{i=1}^{3n} q_i = h \cdot n = s$ and $t = (n+1) \cdot s$.

Now, let $A = \{ a_i, a_j' \mid a_i = q_i, i \in [1, 3n] \text{ and } a_j' = 2t, j \in [1, n-1] \}$,

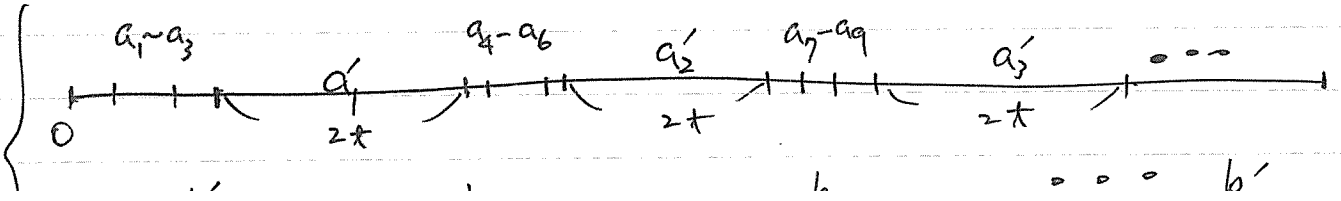
$B = \{ b_j, b_k' \mid b_j = h + 2t, j \in [1, n-2] \text{ and } b_k' = h + t, k \in [1, 2] \}$, and

$C = \{ c_i, c_j' \mid c_i = q_i, i \in [1, 3n] \text{ and } c_j' = t, j \in [1, 2n-2] \}$.

Now, if there exists a solution for the 3-partition instance, let $\sum_{j=1}^{3} a_{3i+j} = h$, $i = 0, 1, 2, \cdots, n-1$

$\Downarrow$

$\sum_{j=1}^{3} c_{3i+j} = h$, $i = 0, 1, 2, \cdots, n-1$, and assign them as following:



We obtain a solution for DDDP. (Check C!)

On the other direction, let $\{A\}$, $\{B\}$ and $\{C\}$ be a solution (coordinates) for DDDP. Observe that for any $n+1$ points of $\{B\}$, the distances obtained consist a multiple of h and a multiple of t. Hence, any two of the points in $\{B\}$, their distance is a multiple of h. Since $\{B\} \subseteq \{C\}$, and C_j's are of distance t, the sum "h" are obtained from C_i's and three consecutive points give a sum "h". This is a solution of (n h's are needed.) the 3-partition problem. (instance)

Errors occurred! (In DDP.)

There are essentially four types of errors.

1. Partial clearage: an enzyme can fail to cut at some restriction site.

2. Fragment length: determining the exact length of a fragment from gel electrophoresis

3. Missing small fragments

4. Doublets : Two different fragments with almost the same length may generate two spots in the gel electrophoresis (凝膠電泳) that overlap. Thus, only one of the fragments is recognized.

(•) The error occurs because of partial clearage is in general "easier" to handle, still very hard.

(••) An approach comes from the so-called Maximum 4-partition problem.

Maximum 4-partition

Given a multiset $Q = \{q_1, q_2, \cdots, q_{4n}\}$ of $4n$ integers such that $\sum_{i=1}^{4n} q_i = nh$ and $\frac{h}{5} < q_i < \frac{h}{3}$, $i = 1, 2, \cdots, 4n$. Find a maximum number of disjoint subsets $S_1, S_2, \cdots, S_m \subseteq Q$, s.t. for $j = 1, 2, \cdots, m$, $\sum_{x \in S_j} x = h$.

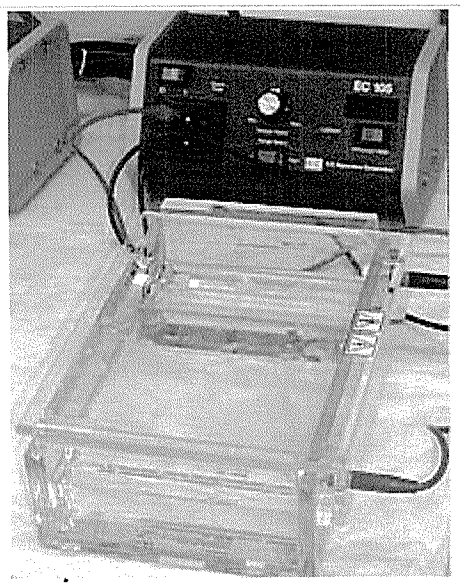
Gel electrophoresis

Gel electrophoresis is a method for separation and analysis of macromolecules (DNA, RNA and proteins) and their fragments, based on their size and charge. It is used in clinical chemistry to separate proteins by charge or size (IEF agarose, essentially size independent) and in biochemistry and molecular biology to separate a mixed population of DNA and RNA fragments by length, to estimate the size of DNA and RNA fragments or to separate proteins by charge.[1]

Nucleic acid molecules are separated by applying an electric field to move the negatively charged molecules through a matrix of agarose or other substances. Shorter molecules move faster and migrate farther than longer ones because shorter molecules migrate more easily through the pores of the gel. This phenomenon is called sieving.[2] Proteins are separated by charge in agarose because the pores of the gel are too large to sieve proteins. Gel electrophoresis can also be used for separation of nanoparticles.

Gel electrophoresis uses a gel as an anticonvective medium or sieving medium during electrophoresis, the movement of a charged particle in an electrical field. Gels suppress the thermal convection caused by application of the electric field, and can also act as a sieving medium, retarding the passage of molecules; gels can also simply serve to maintain the finished separation, so that a post electrophoresis stain can be applied.[3] DNA Gel electrophoresis is usually performed for analytical purposes, often after amplification of DNA via polymerase chain reaction (PCR), but may be used as a preparative technique prior to use of other methods such as mass spectrometry, RFLP, PCR, cloning, DNA sequencing, or Southern blotting for further characterization.

Gel electrophoresis



Gel electrophoresis apparatus – an agarose gel is placed in this buffer-filled box and an electrical field is applied via the power supply to the rear. The negative terminal is at the far end (black wire), so DNA migrates toward the positively charged anode (red wire).

Classification	Electrophoresis
Other techniques	
Related	Capillary electrophoresis SDS-PAGE Two-dimensional gel electrophoresis Temperature gradient gel electrophoresis

Contents

Partial Digest Problem

$\Delta X = \{1,1,2,2,3,4,5,5,6,6,7,8,9,10,11\}$

$\Downarrow$

$X = \{0,1,6,7,9,11\}$ or $\{0,1,2,6,8,11\}$

Skiena et al.

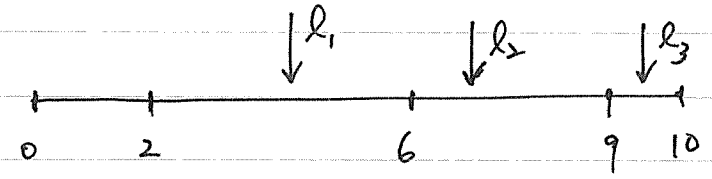
|Possible solutions| is between $\frac{1}{2}n^{0.81}$ and $\frac{1}{2}n^{1.23}$.

So, which one is "correct"?

For the purpose of finding the right answer, labels are added.

1. End-labeled PDP: Add a label at one end.
2. Labeled (both ends) PDP: Add labels to both ends. (Better) $O(n^2 \log n)$
3. Probed PDP: Add a probe (label) somewhere inside target DNA.
4. Mid-labeled PDP (黨禕延):

   k labels are added, one label for a fragment.
   (isolated) radiosotope
   ↑ at most



$X = \{0, 2, 6, 9, 10\}$

$L = \{l_1, l_2, l_3\}$

=

General forms

$X = \{x_1, x_2, \cdots, x_n\}$, $\quad 0 = x_1 < x_2 < \cdots < x_n$;

$L = \{l_1, l_2, \cdots, l_k\}$, $\quad l_1 < l_2 < \cdots < l_k$;

$\Delta X_\phi = \{x_i - x_j \mid l_v \notin (x_i, x_j), \quad 1 \le v \le k\}$

(Distances contain no labels.)

$\Delta X_{l_s, l_t} = \{(x_j - x_i \mid l_u \in (x_i, x_j), \forall s \le u \le t;\ l_v \notin (x_i, x_j), v < s \text{ or } v > t\}$

(Distances contain exactly labels $l_s, l_{s+1}, \cdots, l_t$.)

Example: $k = 3, n = 5$.

$\Delta X_\phi = \{2\}$ $\quad \Delta X_{l_1, l_3} = \{8, 10\}$ $\quad \Delta X_{l_3, l_3} = \{1\}$.

$\Delta X_{l_1, l_1} = \{4, 6\}$ $\quad \Delta X_{l_2, l_2} = \{3\}$

$\Delta X_{l_1, l_2} = \{7, 9\}$ $\quad \Delta X_{l_2, l_3} = \{4\}$

(*) 有 label 的地方，只看位置，不在乎它与两端的距离。

Goal

1. Locate the restriction sites between labels l_{m_1} and l_{m_1+1} where $m_1 = 1, 2, \cdots, k-1$.
2. The rest sites outside of these k labels.

Algorithm

1. For each label, find the nearest site.
2. Use the number of restriction sites between l_{i-1} and l_i for $i = 1, 2, \cdots, k$ to locate restriction sites.

(*) It takes $O(\sqrt{n}\, 2^n \log_2 n)$ and $O(n^{\frac{3}{2}} \cdot 2^{\frac{2n}{k+1}} \log_2 n)$ to find all the sites for $k=1$ and $k>1$ respectively.

Probed DDP

In this method DNA is partially digested with a restriction enzyme, thus generating a collection of DNA fragments between any two cutting sites.

After this action, we attach a labeled probe to the DNA between any two cutting sites which hybridized to the partially digested DNA, and the size of fragments of which the probe hybridized are measured.

The problem is to reconstruct the positions of sites from this multiset of measured lengths.

(利用 probes 來找到 Cutting sites.) → 8'

Optical Mapping

Schwartz et al. (1993), conference paper

(*) In optical mapping, single copies of DNA molecules are stretched and attached to a glass support

由於標記之後，每一片段皆由"+,-"連接而成，因此，數學模型可以表成

$$A=[-s,0],\quad B=[0,t],\quad s,t\in\mathbb{Z}^+,$$

In PPDP, the experiment provides the multiset $E=\{b-a \mid a\in A \text{ and } b\in B\}$.

The problem is "to find $X\subseteq[-s,t]$.

Newberg: The # of solutions is more than $n^{1.72}$.

under a microscope. When restriction enzymes are activated, they "observe" and "cleave" the DNA molecules at their restriction sites. The molecules remain attached to the surface, but the elasticity of the stretched DNA pulls back the molecule ends at the cleaved sites.

(*) These can be identified under the microscope as tiny gaps in the fluorescent line of molecule. Thus, a "photograph" of the DNA molecule with gaps at the position of cleavage sites gives a snapshot of the restriction map.