

In physical mapping, the DNA is cut by restriction enzyme.

Restriction Site Mapping (RSM)

1. A particular restriction enzyme ^(RSM) is added to a DNA solution, the DNA is cut at particular sites.

For example: EcoRI cuts at GAATTC.

2. The goal of ^{an} RSM is to determine the locations of every site for a given enzyme.

Review: the DNA string cannot be explicitly observed.
(太長, 無法直接看)

3. We use various biochemical techniques to obtain indirect infos about where the restriction sites could be.

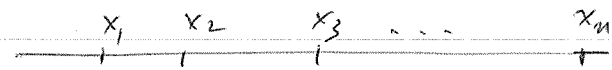
4. Gel electrophoresis is a technique by which the DNA fragments are moved through an electric field, the smaller a fragment is, the faster it will move, its length can be inferred from how far it traveled in the field.

5. The fragment length tells us the distance between some pair of restriction sites.

(2)

6. Given the distances obtained, it is then an algorithmic task to find the locations of all restriction sites on a "long DNA" sequence.
(strand)

Partial Digest Approach

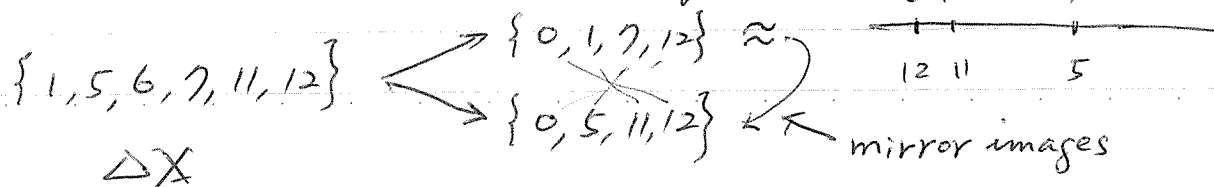


Let $X = \{x_1, x_2, \dots, x_n\}$ be the set of restriction sites on a DNA strand. Then $\Delta X = \{x_i - x_j : 1 \leq j < i \leq n\}$
(on a real line)
is the (multi)set of all $\binom{n}{2}$ distances between sites. Clearly, it is easy to find ΔX by giving X . But, it's not that easy on the reverse direction.

Partial Digest Problem

Given the (multi)set ΔX , find a set X which could have produced ΔX .

(Fact 1) The solutions are not unique.



(3)

Definition (Homometric sets)

If X and Y are the different sets such that $\Delta X = \Delta Y$, then X and Y are homometric sets or simply X and Y are homometric.

(Fact 2) 当 n 很大, 一旦 ΔX 可以决定 $\{k\}$ homometric 的集合。

Definition Let $H(n)$ denote the maximum number of mutually homometric sets of size n for a given ΔX .

(Fact 3) ① $H(n) > \frac{1}{2} n^{0.810}$ for some ΔX ;

② $H(n) < \frac{1}{2} n^{1.233}$ for all n ; and

(*) ③ $H(n) = 2^k$ for some k or $H(n) = 0$.

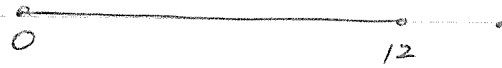
(Fact 4) 正常状况下, "唯一-解", 即 $k=0$.

Backtracking Algorithm

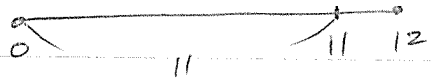
For example $\{1, 5, 6, 7, 11, 12\}$

(4)

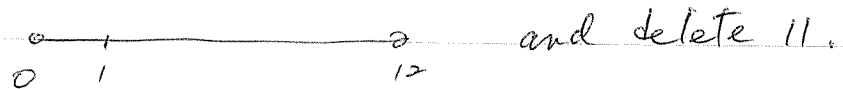
By looking at 12 (maximum), we know



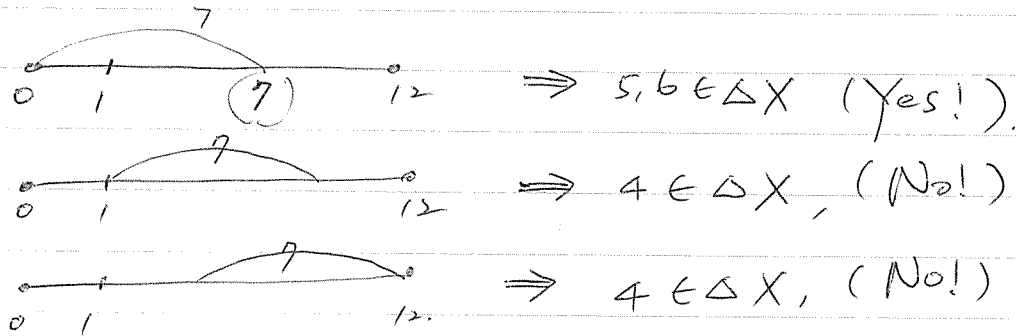
Now, delete 12 and look at the maximum again. Then,



Select one, let it be



Now, the maximum is 7, so we have



We have a solution!

(Note) If we can find a solution from this selection, go back to select another one and keep going!

④

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

$$X \supseteq \{0, 10\}$$



$$x_1 = 0$$

$$x_5 = 10$$

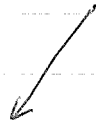


$$\Delta X \supseteq \{2, 2, 3, 3, 4, 5, 6, 7, 8\}$$

Either $x_2 = 2$ or $x_4 = 8$



$$X \supseteq \{0, 2, 10\}, \Delta X \supseteq \{2, 3, 3, 4, 5, 6, 7\}$$



$$x_3 = 3 \Rightarrow 1 \in \Delta X \rightarrow \leftarrow$$

$$x_3 = 5 \text{ or } x_4 = 7$$

$$X \supseteq \{0, 2, 7, 10\}, \Delta X \supseteq \{2, 3, 4, 6\}$$

$$x_3 = 6 \Rightarrow 1 \in \Delta X \rightarrow \leftarrow$$

$$x_3 = 4$$

$$X = \{0, 2, 4, 7, 10\}, \Delta X \supseteq \emptyset$$

Partial Digest (ΔX)1. $width \leftarrow \max \Delta X$ 2. Delete ($width, \Delta X$)3. $X \leftarrow \{0, \underline{width}\}$ 4. Place ($\Delta X, X$)

$$\Delta(X_1, X_2)$$

$$=_{\text{def}} \{ |x_1 - x_2| : x_1 \in X_1, x_2 \in X_2 \}$$

Place ($\Delta X, X$)1. if ΔX is empty2. output X

3. return

4. $y \leftarrow \max \Delta X$ 5. if $\Delta(y, X) \leq \Delta X$ 6. Add y to X and remove lengths $\Delta(y, X)$ from ΔX 7. Place ($\Delta X, X$)8. Remove y from X and add lengths $\Delta(y, X)$ to ΔX 9. if $\Delta(\underline{\text{width} - y}, X) \leq \Delta X$
 $\text{max } \Delta X$ 10. Add $\underline{\text{width} - y}$ to X and remove lengths $\Delta(y, X)$ from ΔX
 y' 11. Place ($\Delta X, X$)12. Remove y' from X
and add lengths
 $\Delta(y', X)$ to ΔX

13. return

(5)

從最大的距離扣起, "Where do we place the largest element $l_{\max}$ of L ?"

Pseudo-code

```

set  $X = \emptyset$ ;

set PartialDigest( list  $L$  )
    width = DeleteMax( $L$ );
     $X = \{0, \text{width}\}$ ;
    if Place( $L$ ) return  $X$ ;
    else return "invalid input";

bool Place(list  $L$ )
    if  $L = \emptyset$  return TRUE;

     $l_{\max} = \text{DeleteMax}(L)$ ;

    if  $\Delta(l_{\max}, X) \subseteq L$                                 //if placing on the left is consistent
         $X = \{l_{\max}\} \cup X$ ;                                //then try this choice
        if Place( $L - \Delta(l_{\max}, X)$ )                    //if subprob is feasible, so is this
            return TRUE;
        else  $X = X - \{l_{\max}\}$ ;                            //otherwise backtrack

    if  $\Delta(\text{width} - l_{\max}, X) \subseteq L$                     //if placing on the right is consistent
         $X = \{\text{width} - l_{\max}\} \cup X$ ;                    //then try that choice
        if Place( $L - \Delta(\text{width} - l_{\max}, X)$ )          //check if feasible
            return TRUE;
        else  $X = X - \{\text{width} - l_{\max}\}$ ;                //otherwise backtrack

    return FALSE;                                          //if neither choice worked, we're infeasible
  
```

Width: 集合中最大與"0"的距離, 即最大數。
數

設 $x_1 = 0$ 。

⑥

一定可以跑出答案! (证明?)

Proposition (Worst-case run time)

The algorithm is $O(2^n n \log n)$.

Proposition (Expected run time)

The algorithm runs in $O(n^2 \log n)$ on "average".

~~~~~

Discussion

Let  $T(n)$  be the maximum time Partial Digest takes to find the solution for an  $n$ -point instance of the PDP.

One viable alternative at every step

$$\Rightarrow T(n) = T(n-1) + \underbrace{O(n)}$$

$\Rightarrow$  quadratic algorithm

the work spent  
adjusting the sets  
 $X$  and  $\Delta X$

Two alternatives

$$\Rightarrow T(n) = 2T(n-1) + O(n)$$

$\Rightarrow$  exponential!

① The solution  $X$  is unique.

②  $\Delta X$  is a set not a multiset.

③ Never backtrack more than a constant <sup>times</sup> <sub># of</sub> in a row.

## Algebraic Set Up!

②

Continued from ③: More about Homometric Sets

Lemma 1 Given two multisets  $U$  and  $V$ . Then  $U+V$  and  $U-V$  are homometric.

Proof. It suffices to prove  $\Delta(U+V) = \Delta(U-V)$ .

( $\subseteq$ ) Let  $d \in \Delta(U+V)$ . Then  $d = (u_2 + v_2) - (u_1 + v_1)$ .

This implies that  $(u_2 - v_1) - (u_1 - v_2) = d \in \Delta(U-V)$ .

( $\supseteq$ ) Let  $d' \in \Delta(U-V)$ . Then  $d' = (u_2 - v_2) - (u_1 - v_1)$

$$= (u_2 + v_1) - (u_1 + v_2) \in \Delta(U+V). \quad \blacksquare$$

Given a multiset  $A = \{a_i\}_{i=1}^n$ . Let  $A(x) = \sum_{i=1}^n x^{a_i}$ . Then

$A(x)$  is a generating function for  $A$ .

Lemma 2 If  $A(x)$  is a generating function for  $A$ , then the generating function  $\Delta A(x)$  for  $\Delta A$  is equal to  $A(x)A(x^{-1})$ .

Proof.  $A(x)A(x^{-1}) = \left( \sum_{i=1}^n x^{a_i} \right) \left( \sum_{j=1}^n (x^{-1})^{a_j} \right)$

$$= \left( \sum_{i=1}^n x^{a_i} \right) \left( \sum_{j=1}^n x^{-a_j} \right)$$

$$= \sum_{i=1}^n \left( x^{a_i} \sum_{j=1}^n x^{-a_j} \right) = \sum_{i=1}^n \sum_{j=1}^n x^{a_i - a_j}$$

$$= \Delta A(x). \quad (\text{這裏的 } \Delta A \text{ 有負的元素!})$$

亦即不一定大的減去小的

②

$$U = \{7, 10, 12\}, \quad V = \{1, 3, 5\}.$$

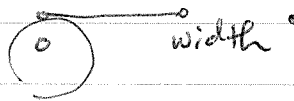
$$U + V = \{8, 10, 12, 11, 13, 15, 13, 15, 17\}$$

$$= \{8, 10, 11, 12, 13, 15, 17\}$$

$$U - V = \{6, 4, 2, 9, 7, 5, 11, 9, 7\}$$

$$= \{2, 4, 5, 6, 7, 9, 11\} = (U + V)^v - b.$$

We may choose "negative numbers", i.e., the sets are not in standard positions



⑧

Lemma 3 Let  $A(x) = U(x)V(x)$  and  $B(x) = U(x)V(x^{-1})$ .

Then  $A$  and  $B$  are homometric sets.

Proof.  $\Delta A(x) = A(x)A(x^{-1})$ ,  $\Delta B(x) = B(x)B(x^{-1})$ . Since

$$A(x)A(x^{-1}) = U(x)V(x)U(x^{-1})V(x^{-1}) = [U(x)V(x^{-1})][U(x^{-1})V(x)]$$

$$= B(x)B(x^{-1}), \Delta A(x) = \Delta B(x) \text{ and thus } A \text{ and } B \text{ are homometric.}$$

是否所有的 homometric sets 可以由此种建構方式获得? No!  $\{0, 1, 2, 5, 7, 9, 12\}$

Theorem 4. (Rosenblatt and Seymour, 1982)

SIAM J. Alg. Discrete Methods 3  
→ 43-350.

Two sets  $A$  and  $B$  are homometric if and only if there exist generating functions  $U(x)$  and  $V(x)$ , and an integer  $\beta$  such that  $A(x) = U(x)V(x)$  and  $B(x) = \pm x^\beta U(x)V(x^{-1})$ .

Proof.

$$(\Leftarrow) \quad \Delta A(x) = U(x)V(x)U(x^{-1})V(x^{-1})$$

$$\Delta B(x) = (x^\beta U(x)V(x^{-1}))(x^{-\beta} U(x^{-1})V(x))$$

$$= U(x)U(x^{-1})V(x^{-1})V(x) = \Delta A(x). \quad \text{done.}$$

$(\Rightarrow)$  Let  $A$  and  $B$  be homometric sets, and  $A(x)$  and  $B(x)$  be their corresponding generating functions.

(9)

Let  $\text{g.c.d.}\{A(x), B(x)\} = p(x)$ . Then,  $A(x) = p(x) Q_A(x)$  and  $B(x) = p(x) Q_B(x)$ . Clearly,  $\text{g.c.d.}\{Q_A(x), Q_B(x)\} = 1$ . Now, consider  $Q_A(x^{-1})$  and  $Q_B(x^{-1})$ . Let  $V(x) = \text{g.c.d.}\{Q_A(x), Q_B(x^{-1})\}$  and thus  $Q_A(x) = V(x) S_A(x)$  and  $Q_B(x^{-1}) = V(x) S_B(x)$ , furthermore  $S_A(x)$  and  $S_B(x)$  are relatively prime. We also conclude that  $S_A(x)$  and  $S_A(x^{-1})$  are relatively prime to both  $S_B(x)$  and  $S_B(x^{-1})$ .

Since  $A$  and  $B$  are homometric,  $A(x)A(x^{-1}) = B(x)B(x^{-1})$ , i.e.,  $p(x)V(x)S_A(x)p(x^{-1})V(x^{-1})S_A(x^{-1}) = p(x)V(x^{-1})S_B(x^{-1})p(x^{-1})V(x)S_B(x)$ . This implies that  $S_A(x)S_A(x^{-1}) = S_B(x)S_B(x^{-1})$ . Since  $S_A(x)$  and  $S_A(x^{-1})$  are relatively prime to both  $S_B(x)$  and  $S_B(x^{-1})$ ,  $S_A(x) = \pm x^a$  and  $S_B(x) = \pm x^b$ . Therefore,

$$A(x) = \pm x^a p(x) V(x) \text{ and } B(x) = \pm x^b p(x) V(x^{-1}).$$

Now, the proof follows by letting  $U(x) = \pm x^a p(x)$  and  $\beta = b - a$  which is an integer.

(註) 這裏的互質只是 polynomial 的部分。

### Example

$$A = \{0, 1, 2, 5, 7, 9, 12\}, \quad B = \{0, 1, 5, 7, 8, 10, 12\}.$$

$$U(x) = 1 + x + x^2 + x^3 + x^4 + x^5 + x^7$$

$$V(x) = x^{-5}(1 - x^3 + x^5)$$

$$U(x) V(x) = x^{-5} (1 - x^3 + x^5) (1 + x + x^2 + x^3 + x^4 + x^5 + x^7)$$

$$= x^{-5} (1 + x + x^2 + \cancel{x^3} + \cancel{x^4} + \cancel{x^5} + \cancel{x^6} - \cancel{x^7} - \cancel{x^8} - \cancel{x^9} - \cancel{x^{10}} + \cancel{x^{11}} + \cancel{x^{12}})$$

$\beta = 5$

$$V(x^{-1}) = x^5(1 - x^{-3} + x^{-5}) = x^5 - x^2 + 1$$

$$x^5 U(x) \underline{V(x^{-7})} =$$
$$\begin{aligned} & 1 + x + x^2 + x^3 + x^4 + x^5 + x^6 \\ & - x^2 - x^3 - x^4 - x^5 - x^6 - x^7 - x^8 \\ & + x^5 + x^6 + x^7 + x^8 + x^9 + x^{10} + x^{12} \\ & \quad \downarrow +x^5 \\ & = 1 + x + x^7 + x^8 + x^{10} + x^{12} = B(x) \end{aligned}$$

## Pseudo-polynomial algorithm for PDP (Rosenblatt and Seymour, 1982)

- Given  $\Delta A = \{d_i\} \quad i=1, 2, \dots, \binom{n}{2}$ .

- $\Delta A(x) = n + \sum_i (x^{d_i} + x^{-d_i})$ .  $\Leftarrow$

$$\begin{cases} x_i - x_j = d_{ij} \\ \Rightarrow x_j - x_i = -d_{ij} \\ \boxed{x_i - x_i = 0} \end{cases}$$

- Factor this "polynomial" into irreducibles over the ring of polynomials with integer coefficients using a factoring algorithm with runtime polynomial in  $\max_i d_i$ .
- The solution  $A(x)$  of the PDP problem must have a form  $\Delta A(x) = A(x) A(x)^{-1}$ .
- Try all  $2^F$  possible subsets  $S$  of  $F$  irreducible nonreciprocal factors of  $\Delta A(x)$  as putative factors of  $A(x)$ .
- Find a set of all possible sets  $A(x)$ .
- Eliminate the sets  $A(x)$  with negative coefficients and sort to remove possible redundant copies.

## Remark

(\*) Noisy Versions (PDP) : NP-hard

(i) Measurement error : the elements in  $\Delta X$  can be observed with 0.1 % error at best.

(ii) Additions : restriction enzymes are not perfect, stray DNA can contaminate the specimen, extraneous elements in  $\Delta X$  may be present.

(iii) Deletions : some elements of  $\Delta X$  may be missing.

(\*) M. Cieliebak et al. proved that variations of PDP reflecting (ii) and (iii) scenarios are NP-hard.



## Min. Partial Digest Superset Problem (MPDSP)

Given a multiset  $L$  with  $|L|=m$ , find the minimum  $n^*$  such that  $|X|=n^*$  and  $L \subseteq \Delta X$ .

For example,  $m=8$ , " $n^*=5$ " is the best choice, but  $n^*$  may be larger!

Trivial bound for  $n^* \leq m+1$ , see it?

Theorem MPDSP is NP-hard.

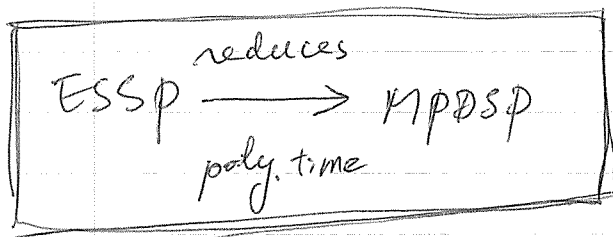
Equal Sum Subsets (ESS) problem (NP-complete problem)

Given a set (not a multiset) of  $m$  elements (numbers)

$L$ , are there two disjoint, non-empty subsets  $A, B \subseteq L$  s.t.

$$\sum_{x \in A} x = \sum_{y \in B} y, \text{ i.e., } \text{sum}(A) = \text{sum}(B).$$

Theorem There is a solution  $(A, B)$  to ESSP if and only if the minimum  $n^*$  to MPDSP is at most  $m$ , i.e.,  $n^* \leq m$ .



In a word, there is a solution!