

In DNA sequencing, the partial digest problem (PDP) plays an important role on reconstructing the location sites in DNA.

Model

Let $X = \{x_1, x_2, \dots, x_n\}$ be the set of restriction sites with $0 = x_1 < x_2 < \dots < x_n$ where x_i 's are positive integers.

Let $\Delta X = \{x_j - x_i \mid 1 \leq i < j \leq n\}$ be the multi-set of distances between every two distinct restriction sites.

Then, the PDP is to find the solutions X by knowing ΔX .

e.g. $X = \{0, 1, 6, 7, 9, 11\}$.

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

$$|\Delta X| = \binom{|X|}{2}.$$

$$X \Rightarrow \Delta X \quad (\text{Easy})$$

$$\Delta X \Rightarrow X \quad (\text{Much harder!})$$

If we take $X' = \{11-11, 11-9, 11-7, 11-6, 11-1, 11-0\}$
 $= \{0, 2, 4, 5, 10, 11\}$, then $\Delta \bar{X} = \Delta \bar{X}'$.

(This fact is obtained from the symmetric "distances".)
 Symmetry

(*) Not only that, there are other solutions X for the same $\Delta \bar{X}$. For example, $X = \{0, 1, 2, 6, 8, 11\}$.

(*) Skiena et al., A partial digest approach to restriction site mapping, Bulletin of Math. Biology, vol. 56, no. 2, 275-294, 1994.

(*) The number of possible solutions is between
 $\frac{1}{2} n^{0.81}$ and $\frac{1}{2} n^{1.23}$ if $\bar{X} = \{x_1, x_2, \dots, x_n\}$.

(*) We can use back-tracking algorithms for finding the solutions. (Exponential time!)

Observation

Starting from larger distance

1. $\max \Delta \bar{X} = x_n$

2. x_2 or x_{n-1} come from 2nd largest distance.
 e.g. $10 \Rightarrow x_2 = 1$ or $x_{n-1} = 10$.

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

From 11,



From 10, choose 10, new list $\{1, 2, 2, 3, 4, 5, 5, 6, 6, 7, 8, 9\}$

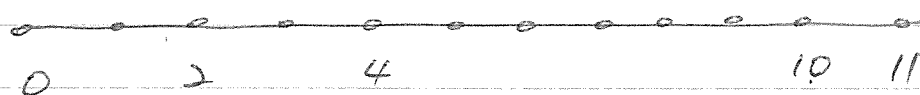


From 9, choose 2, new list $\{1, 2, 3, 4, 5, 5, 6, 7\}$.

From 7, choose 7, new list $\{1, 2, 5, 6, 6\}$

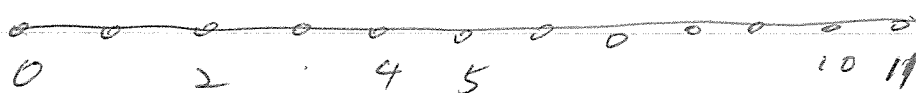
↓
No other choice can achieve the goal!

So, go back to choose 4 for "7".



new list $\{1, 3, 5, 5, 6\}$

From 6, choose 5



We have $X = \{0, 2, 4, 5, 10, 11\}$, another answer, but it is symmetric to $X = \{0, 1, 6, 7, 9, 11\}$.

1. Set $L = \Delta X$ and $x_1 = 0$.
2. Define the function $\text{Delete_Max}(L)$ as "returns the maximum value of L and removes it from L ".
3. Let $\Delta(X, Y)$ denote the multiset of all distances between a point of X and a point of Y .
4. Let X and width be two variables.

Set X

int width

$\text{PD}(\text{List } L)$

$\text{width} = \text{Delete_Max}(L)$

$X = \{0, \text{width}\}$

$\text{place}(L)$

$\text{place}(L)$

if $L = \emptyset$ then

output X (solution)

exit

$y = \text{Delete_Max}(L)$

if $\Delta(\{y\}, X) \leq L$ then

$X = X \cup \{y\}$

$\text{place}(L \setminus \Delta(\{y\}, X))$

$X = X \setminus \{y\}$

if $\Delta(\{\text{width} - y\}, X) \leq L$ then

$X \cup \{\text{width} - y\}$

Place a point at
right and $\Leftarrow$
"backtracking"

$\Rightarrow$ to $\{y\}$ to output of place

place a point $\Leftarrow$
at left and
"backtracking"

Place $(L \setminus \Delta(\{\text{width}-y\}, X)$
 $X \setminus \{\text{width}-y\}$

Homometric Sets

- (*) In general, "reconstruct X from ΔX " provides more than one solution.

Definition (Homometric Sets)

Two sets A and B are homometric if $\Delta A = \Delta B$.

Proposition Given two multisets U and V . Then,

$U+V$ and $U-V$ are homometric. ($U \pm V = \{u \pm v \mid u \in U, v \in V\}$)

Proof. We claim that $\Delta(U+V) = \Delta(U-V)$.

Say something
about this!
(See '5')

Let $\overset{eU}{u_1} + \overset{eV}{v_1}$ and $\overset{eU}{u_2} + \overset{eV}{v_2}$ be any two elements in $U+V$. Then,

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

$\Delta(U+V) \subseteq \Delta(U-V)$. On the other hand, let $u_2 - v_2$ and

$u_1 - v_1$ be two elements in $U-V$. Then, $(u_2 - v_2) - (u_1 - v_1) =$

$(u_2 + v_1) - (u_1 + v_2) \in \Delta(U+V)$. The proof follows. 

Given a set A with cardinality $|A|$.

Questions

1. How large $|A+A|$ may be?
2. How large $|A \cdot A| = |\{a \cdot a' \mid a \in A, a' \in A\}|$ may be?
3. $|A \cdot A + A| = ?$

Answers (Partial)

1. $|A+A|$ may be very small, which is about $|A|$, if A is an arithmetic progression sequence (等差数列). But, if $A = \{2^0, 2^1, \dots, 2^n\}$, then $|A+A|$ is very large.
 2. $|A \cdot A|$ may be small if A is a geometric progression sequence (等比数列) and $|A \cdot A|$ is large if A is an arithmetic progression sequence.
 3. Combining $A \cdot A$ and A , the size of $|A \cdot A + A|$ is always quite large comparing to $|A|$.
- (*) If A is the set of primes, then $A \cdot A + A$ contains all even (positive) integers. (陈景润)

e.g. $U = \{6, 7, 9\}$ and $V = \{-6, 2, 6\}$

$$\begin{cases} U+V = \{0, 1, 3, 8, 9, 11, 12, 13, 15\} \\ U-V = \{0, 1, 3, 4, 5, 7, 12, 13, 15\} \end{cases}$$

Not symmetric!

(*) Not all homometric sets can be constructed via finding U and V , and use the above proposition.

(*) As a matter of fact, we can provide a characterization of two homometric sets.

(*) Given a multiset $A = \{a_i\}_n = \{a_1, a_2, \dots, a_n\}$. Then, let

$$A(x) = \sum_{i=1}^n x^{a_i} \text{ be a generating function of } A.$$

Since

(*) ΔA is a multiset, let $\Delta A(x)$ be a generating function

of ΔA . Then $\Delta A(x) = A(x) \cdot A(\bar{x})$. That is,

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

e.g. $A = \{0, 1, 6, 7, 9, 11\}$

$$A(x) = (1 + x + x^6 + x^7 + x^9 + x^{11})$$

$$A(\bar{x}) = (1 + \bar{x}^{-1} + \bar{x}^{-6} + \bar{x}^{-7} + \bar{x}^{-9} + \bar{x}^{-11})$$

$$A(x) A(\bar{x}) = \dots$$

$$= 6 + 2x^1 + 2x^2 + x^3 + x^4 + \dots$$

$$2\bar{x}^{-1} + 2\bar{x}^{-2} + \bar{x}^{-3} + \bar{x}^{-4} + \dots$$

(*) We can use generating functions to prove the $U+V$ and $U-V$ construction.

Proposition Let $A(x)$ and $B(x)$ be generating functions of $\mathcal{A}$ and $\mathcal{B}$ respectively such that $A(x) = U(x)V(x)$ and $B(x) = U(x)V(\bar{x}')$ where $U(x)$ and $V(x)$ are generating functions of $\mathcal{U}$ and $\mathcal{V}$ respectively.

Then, $\Delta A(x) = A(x)A(\bar{x}') = U(x)V(x) \cdot U(\bar{x}')V(\bar{x}')$
 (Proof.)
 $= U(x)V(\bar{x}') \cdot U(\bar{x}')V(x) = B(x)B(\bar{x}') = \Delta B(x).$ $\square$

Theorem (Rosenblatt and Seymour, The structure of homometric sets, SIAM J. Alg. Discrete Methods, ³(1982), 343-350.)

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

Proof. ($\Leftarrow$) Sufficiency

Since $\Delta A(x) = U(x)V(x)U(\bar{x}')V(\bar{x}')$, and

$$\Delta B(x) = (\pm x^\beta U(x)V(\bar{x}'))(\pm x^{-\beta} U(\bar{x}')V(x)) = U(x)V(x)U(\bar{x}')V(\bar{x}'),$$

the proof follows.

($\Rightarrow$) Let A and B be homometric sets. Let $p(x) = \gcd(A(x), B(x))$.

Then, $A(x) = p(x) \cdot Q_A(x)$ and $B(x) = p(x) \cdot Q_B(x)$, moreover $\gcd(Q_A(x), Q_B(x))$ is a constant. ($Q_A(x)$ and $Q_B(x)$ are relatively prime.)

Now, let $V(x) = \gcd(Q_A(x), Q_B(x^{-1}))$. Hence,

$Q_A(x) = S_A(x) \cdot V(x)$ and $Q_B(x^{-1}) = S_B(x) \cdot V(x)$, and ($S_A(x)$ and $S_B(x)$ are

relatively prime). This implies that $S_A(x)$ and $S_A(x^{-1})$ are relatively prime to both $S_B(x)$ and $S_B(x^{-1})$.

By assumption that A and B are homometric, we have

$$\Delta A(x) = p(x) Q_A(x) p(x^{-1}) Q_A(x^{-1}) = p(x) Q_A(x) p(x^{-1}) Q_B(x^{-1}) = \Delta B(x). \text{ Thus,}$$

$$\underbrace{p(x)} \underbrace{V(x)} S_A(x) \underbrace{p(x^{-1})} \underbrace{V(x^{-1})} S_A(x^{-1}) = \underbrace{p(x)} \underbrace{V(x^{-1})} S_B(x^{-1}) \underbrace{p(x^{-1})} \underbrace{V(x)} S_B(x),$$

implying that $S_A(x) S_A(x^{-1}) = S_B(x) S_B(x^{-1})$. Therefore,

$S_A(x) = \pm x^a$ and $S_B(x) = \pm x^b$ for some integers a and b . This

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

Now, by letting $U(x) = \pm x^a p(x)$, we conclude the proof. $\square$

Notice that " β " can be negative.

ex. 9.

$$A = \{0, 1, 2, 5, 7, 9, 12\}$$

$$\Delta A = \{1, 2, 5, 7, 9, 12, 1, 4, 6, 8, 11, 3, 5, 7, 10, 2, 4, 7, 2, 5, 3\}$$

$$= \{1, 1, 2, 2, 2, 3, 4, 4, 5, 5, 5, 6, 7, 7, 7, 8, 9, 10, 10, 11, 12\}$$

$$B = \{0, 1, 5, 7, 8, 10, 12\}$$

$$\Delta B = \{1, 5, 7, 8, 10, 12, 4, 6, 7, 9, 11, 2, 3, 5, 7, 1, 3, 5, 2, 4, 2\}$$

$$= \{1, 1, 2, 2, 2, 3, 3, 4, 4, 5, 5, 5, 6, 7, 7, 7, 8, 9, 10, 11, 12\}$$

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

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

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

$$\rightarrow +x^3 + x^4 + x^5 + x^6 + x^7 + x^8 + x^{10}$$

$$+x^5 + x^6 + x^7 + x^8 + x^9 + x^{10} + x^{12}$$

$$= (1 + x + x^2 + x^5 + x^7 + x^9 + x^{12}) \cdot x^{-5}$$

$$A(x) = x^5 \cdot U(x)V(x)$$

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

$$= 1 + x + x^2 + x^3 + x^4 + x^5 + x^7$$

$$- x^2 - x^3 - x^4 - x^5 - x^6 - x^7 - x^9$$

$$+ x^5 + x^6 + x^7 + x^8 + x^9 + x^{10} + x^{12}$$

$$= 1 + x + x^5 + x^7 + x^8 + x^{10} + x^{12}$$

Question

Given $A(x)$ and $B(x)$, find $U(x)$ and $V(x)$ satisfying the above construction.

Definition (Reconstructible)

A set A is reconstructible if whenever B is homometric to A , we have $B = A + \{v\}$ or $B = -A + \{v\}$.

(We don't assume $x_1 = 0$ here.)

A is said to be symmetric if $-A = A + \{v\}$.

e.g. $A = \{0, 1, 4, 5\}$, $-A = \{0, -1, -4, -5\}$, $A + \{-5\} = -A$.

$$A(x) = 1 + x + x^4 + x^5$$

$$A(x^{-1}) = 1 + x^{-1} + x^{-4} + x^{-5}$$

$$A(x^{-1}) = x^{-5}(1 + x + x^4 + x^5) = x^{-5} \cdot A(x)$$

A poly. $A(x)$ is symmetric if the corresponding set is symmetric.

Theorem A set is reconstructible if and only if $A(x)$ has at most one prime factor that is not symmetric.

Proof. It follows from the previous theorem. ($U(x)$)

(*) By using the idea of generating functions, Rosenblatt and Seymour gave a pseudo-polynomial algorithm for PDP.
(See the reference in page 7.)