

Partial Digest Problem

In DNA sequencing, the partial digest problem (PDP) plays an important role on reconstructing the locations 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 multiset of distances between every two distinct restriction sites. Then, PDP is to find the solution X by giving ΔX .

Example of PDP

- Example

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

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

$$|\Delta X| = C(6, 2) = 15.$$

- $X \rightarrow \Delta X$ (Easy!)
- $\Delta X \rightarrow X$ (Much harder!)
- Let $X' = \{x_n - x_i \mid i = 1, 2, \dots, n\}$. Then $\Delta X = \Delta X'$. (This is known as the set of symmetric restriction sites.)

More Solutions of PDP

(*) For a given ΔX , there are possibly more than one solutions except for the symmetric one. For example, $\{0, 1, 2, 6, 8, 11\}$ is also a solution of the example $\Delta X = \{1, 1, 2, 2, 3, 4, 5, 5, 6, 6, 7, 8, 9, 10, 11\}$. (The one in previous page is $\{0, 1, 6, 7, 9, 11\}$).

- A good reference for this problem is:

Skiena et al., A partial digest approach to restriction site mapping, Bulletin of Math. Biology, Vol. 56, no. 2, 275 – 294, 1994.

** If X has n restriction sites, then the number of possible solutions is between $0.5n^{0.81}$ and $0.5n^{1.23}$.*

Algorithm for PDP

- Back-tracking algorithm

We use an example to explain the idea: $\Delta X = \{1, 1, 2, 2, 3, 4, 5, 5, 6, 6, 7, 8, 9, 10, 11\}$.

Step 1. From 11, the maximum integer, we have



Step 2. From 10, we have 1 or 10 to choose, choose “10”. Now, we have



Continued

Step 3. From 9, we can choose 2 or 9, choose “2”. Now, we have



Step 4. From 7 (we already have 8), we can choose 4 or 7, choose “7”.

Now, we have



Here, we have left {1, 2, 5, 6, 6}. Not able to find X!

Continued

Step 5. Go back to find 4 instead of 7. Now we have



The list left is {1, 3, 5, 5, 6} which can be achieved by choosing 5 to get the job done.



Note that this answer is symmetric to {0, 1, 6, 7, 9, 11}.

Algorithm

1. Set $L = \Delta X$ and $x_1 = 0$.
2. Define $\text{Max}(L)$ 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

PD(List L) ... (*Backtracking algorithm*) ...

Combinatorial Problems

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

Problem

1. How large $|A + A|$ may be?
2. How about $|A \cdot A|$ and $|A + A \cdot A|$?

Answer

1. $|A + A|$ may be small if A is a set of arithmetic progression integers.
 2. $|A \cdot A|$ may be small if A is a set of geometric progression integers.
 3. $|A + A \cdot A|$ is always quite large comparing to $|A|$.
- (**) If A is the set of all primes, then $A + A \cdot A$ contains all even integers.

Homo-metric Sets

- Definition

Two sets A and B are homo-metric if $\Delta A = \Delta B$.

- Proposition

Given two sets U and V. Then, $U + V$ and $U - V$ are homo-metric.

Proof. It suffices to claim that $\Delta(U + V) = \Delta(U - V)$. Let $u_1 + v_1$ and $u_2 + v_2$ be two elements in $U + V$. Then, $(u_2 + v_2) - (u_1 + v_1) = (u_2 - v_1) - (u_1 - v_2)$, that is an element of $\Delta(U + V)$ is also an element of $\Delta(U - V)$. It is true for the other way around. The proof follows.

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

$$U + V = \{0, 1, 3, 8, 9, 11, 12, 13, 15\}$$

$$U - V = \{0, 1, 3, 4, 5, 7, 12, 13, 15\}$$

Generating Functions

- Note that not all homo-metric sets can be constructed via finding U and V mentioned in above proposition.
- *We can use generating functions to provide a characterization of two homo-metric sets.*

Definition

Given a set $A = \{a_1, a_2, \dots, a_n\}$. Let $A(x)$ be a generating function of A , that is, $A(x) = \sum x^{a_i}$ for $i = 1, 2, \dots, n$.

Example $A = \{0, 1, 6, 7, 9, 11\}$, $A(x) = 1 + x^1 + x^6 + x^7 + x^9 + x^{11}$, and

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

Important Results

Fact. $A(x) \cdot A(x^{-1})$ is a generating function of $\Delta A(x)$.

Proposition

Let $A(x)$ and $B(x)$ be generating functions of A and B respectively such that $A(x) = U(x) \cdot V(x)$ and $B(x) = U(x) \cdot V(x^{-1})$ where $U(x)$ and $V(x)$ are generating functions of U and V respectively. Then, $\Delta A(x) = \Delta B(x)$.

Proof. ($U + V$ and $U - V$ are homo-metric.)

$$\begin{aligned}\Delta A(x) &= A(x) \cdot A(x^{-1}) = U(x) \cdot V(x) \cdot U(x^{-1}) \cdot V(x^{-1}) = U(x) \cdot V(x^{-1}) \cdot U(x^{-1}) \cdot V(x) = \\ &B(x) \cdot B(x^{-1}) = \Delta B(x).\end{aligned}$$

Stronger Result

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

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

Proof. It is easier to see the sufficiency part. (?)

Now, we prove the other direction. That is to claim: if A and B are homo-metric sets, then there exist $U(x)$ and $V(x)$ as mentioned in the statement. Let $A(x)$ and $B(x)$ be the generating functions of A and B respectively.

Continued

Let $P(x) = \text{g.c.d.}(A(x), B(x))$. Then, $A(x) = P(x) \cdot Q_A(x)$ and $B(x) = P(x) \cdot Q_B(x)$.

Clearly, $\text{g.c.d.}(Q_A(x), Q_B(x))$ is a constant or $Q_A(x)$ and $Q_B(x)$ are relatively prime in the sense of polynomials.

Now, let $V(x) = \text{g.c.d.}(Q_A(x), Q_B(x^{-1}))$. Hence, $Q_A(x) = S_A(x) \cdot V(x)$ and $Q_B(x^{-1}) = S_B(x) \cdot V(x)$. Again, $S_A(x)$ and $S_B(x)$ are relatively prime, so are $S_A(x^{-1})$ and $S_B(x^{-1})$.

Since A and B are homo-metric, $\Delta A(x) = P(x) \cdot Q_A(x) \cdot P(x^{-1}) \cdot Q_A(x^{-1}) = P(x) \cdot Q_B(x) \cdot P(x^{-1}) \cdot Q_B(x^{-1}) = \Delta B(x)$. Thus, $P(x) \cdot S_A(x) \cdot V(x) \cdot P(x^{-1}) \cdot S_A(x^{-1}) \cdot V(x^{-1}) = P(x) \cdot S_B(x^{-1}) \cdot V(x^{-1}) \cdot P(x^{-1}) \cdot S_B(x) \cdot V(x)$, implying $S_A(x) \cdot S_A(x^{-1}) = S_B(x) \cdot S_B(x^{-1})$.

By the fact, $S_A(x)$ and $S_B(x)$ are relatively prime, so are $S_A(x^{-1})$ and $S_B(x^{-1})$, $S_A(x) = \pm x^a$, and $S_B(x) = \pm x^b$ for some integers a and b. Hence, $A(x) = P(x) \cdot Q_A(x) = A(x) = P(x) \cdot S_A(x) \cdot V(x) = \pm x^a P(x) \cdot V(x)$. Now, let $U(x) = \pm x^a \cdot P(x)$, we have $A(x) = U(x) \cdot V(x)$ and $B(x) = \pm x^\beta \cdot U(x) \cdot V(x^{-1})$. ($\beta = b - a$)

An Example

- $A = \{0, 1, 2, 5, 7, 9, 12\}$
- $B = \{0, 1, 5, 7, 8, 10, 12\}$
- $\Delta A = \Delta B = \{1^2, 2^3, 3^2, 4^2, 5^3, 6, 7^3, 8, 9, 10, 11, 12\}$
- Choose $U(x) = 1 + x + x^2 + x^3 + x^4 + x^5 + x^7$ and $V(x) = x^{-5}(1 - x^3 + x^5)$.
- $U(x)V(x) = x^{-5}(1 + x + x^2 + x^3 + x^4 + x^5 + x^7)(1 - x^3 + x^5) = x^{-5}(1 + x + x^2 + x^5 + x^7 + x^9 + x^{12})$ and $U(x)V(x^{-1}) = (1 + x + x^5 + x^7 + x^8 + x^{10} + x^{12})$.
- **Problem**
Given $A(x)$ and $B(x)$, find $U(x)$ and $V(x)$ satisfying the above construction.

Exercise 6

Find an idea to solve the Partial Digest Problem.

Definition

A set A is re-constructible if whenever B is homo-metric to A , we have $B = A + \{v\}$ or $B = -A + \{v\}$ for some integer v .

Definition

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

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

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

(*) By using the idea of generating functions, Rosenblatt and Seymour gave a pseudo-polynomial algorithm for PDP. See the reference in page 12.