

Genome Rearrangements

Lecture 11

12/24-25

①

Genome Comparison $\leftrightarrow$ Gene Comparison

Late 1980s, Jeffrey Palmer et al.

Compare the mitochondrial genomes of

基因组

Brassica oleracea and *Brassica campestris*.
(cabbage) 甘蓝 (包心菜) (turnip) 萝卜

They are very closely related (many genes are 99% identical).

遗传分子

(*) These molecules are almost identical in gene sequences,

differ dramatically in gene order.

1 -5 4 -3 2 (cabbage)

1 -5 4 -3 -2

1 -5 -4 -3 -2

↓

1 2 3 4 5 (turnip)

↓ prove

(•) Genome rearrangements represent a common mode of molecular evolution.

(2)

(Fact 1) Every study of genome rearrangements involves solving a combinatorial "puzzle" to find a series of rearrangements that transform one genome into another.

(Fact 2) Some are simple, but most of them are complicated.

1 2 31 34 28 26 17 29 4 9 36 18 35 19 1 16 14 32 33 22 15 11 27 5 20 13 30 23 10 6 3
24 21 8 25 27

} Uses 26 reversals (Can you do it?)

1 2 3 4 . . .

36

Definition Let $\pi = \langle a_1 a_2 \dots a_i a_{i+1} \dots a_j a_{j+1} \dots a_n \rangle$ be a sequence.

An (i, j) -reversal, denoted by $\rho(i, j)$ of π is the permutation

$$\begin{pmatrix} 1 & 2 & \dots & i-1 & i & i+1 & \dots & j & j+1 & \dots & n \\ 1 & 2 & \dots & i-1 & j & j-1 & \dots & i & j+1 & \dots & n \end{pmatrix}.$$

$$\pi \cdot \rho(i, j) = \langle a_1 a_2 \dots a_{i-1} \underbrace{a_j \dots a_{i+1} a_i}_{\text{reversed}} a_{j+1} \dots a_n \rangle.$$

(*) For convenience, we apply this idea on $\pi \in S_n$, the set of permutations of an n -elements set $\{1, 2, \dots, n\}$.

The example in Fact 2 shows the mouse X chromosome is transformed into human X chromosome by using 26 reversals.

优点

(不是优点) ③

缺点

(*) Genome comparison has certain merits and demerits as compared to classical gene comparison: genome comparison ignores actual DNA sequences of genes, while gene comparison ignores gene order. The ultimate goal would be to combine the merits of both genome and gene comparison in a single algorithm.

(Note) The analysis of genome rearrangements in molecular biology was pioneered in the late 1930s by Dobzhansky and Sturtevant:

17 inversions for the species of *Drosophila* fruit fly. (1938, *Genetics* 23, 28-64.)

(来源)

With the advent of large-scale mapping and sequencing, the number of genome comparison problems is rapidly growing in different areas, including viral, bacterial, yeast, plant and animal evolution.

Sorting by Reversals (Permutations)

Definition (Reversal distance) ^{(of permutations (or between))}

Given permutations π and α , the reversal distance is the minimum number t such that $\pi \cdot p_1 p_2 \cdots p_t = \alpha$ where p_i 's are reversals. We call t the reversal distance between π and α , denoted by $d(\pi, \alpha) = t$.

→ $d(\pi)$ in short

(4)

Definition (Prefix-reversals)

A reversal of the form $p(1, i)$ is called a prefix-reversal.

Sorting by prefix reversals problem (Pancake flipping problem)

Given an arbitrary permutation π , find the minimum number of prefix reversals sorting π (to e).

$$d_{\text{pref}}(\pi)$$

為什麼又稱爲是 pancake flipping 問題?

(Harry Dweigter)

The chef in our place is sloppy, and when he prepares a stack of pancakes they come out all different sizes. Therefore, when I deliver them to a customer, on the way to a table I arrange them (so that the smallest winds up on top, and so on, down to the largest at the bottom) by grabbing several from the top and flipping them over, repeating this (varying the number I flip) as many times as necessary. If there are n pancakes, what is the maximum number of flips that I will ever have to use to rearrange them?

Example $\pi = (1, 3, 2, 5, 4)$

~~(1, 3, 2, 5, 4)~~
(2, 3, 1, 5, 4)

~~(2, 3, 1, 5, 4)~~
(3, 2, 1, 5, 4)

~~(3, 2, 1, 5, 4)~~
(1, 2, 3, 5, 4)

~~(1, 2, 3, 5, 4)~~
(4, 5, 3, 2, 1)

↓
↓
↓
↓
↓

$$k \leq 6$$

$$\overline{d_{\text{pref}}(\pi)} \leq 6$$

$$\text{But, } \underline{d(\pi) = 2.}$$

最多 2n 次?

(直觀)

William H. Gates

(5)

Late 1970s, (Bill Gates and Cristos Papadimitriou),
[1979, D.M.] 27, 47-57.

(*) They proved that the prefix reversal diameter of the symmetry group, $d_{\text{pref}}(n) = \max_{\pi \in S_n} d_{\text{pref}}(\pi)$, is at most $\frac{5}{3}(n+1)$ and there are infinite many n such that $d_{\text{pref}}(n) \geq \frac{17}{16}n$.

($\therefore$) The pancake flipping problem still remains unsolved.

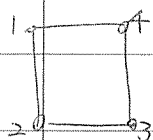
Note We may consider subgroups of S_n .

Example

$(3\ 2\ 1\ 4) \rightarrow (1\ 2\ 3\ 4)$

pref. reversal

$(3\ 4\ 1\ 2) \rightarrow (4\ 3\ 1\ 2) \rightarrow (2\ 1\ 3\ 4) \rightarrow (1\ 2\ 3\ 4)$



$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{pmatrix}$

Problem What is $d'_{\text{pref}}(n) = \max_{\pi \in D_n} d_{\text{pref}}(\pi)$?

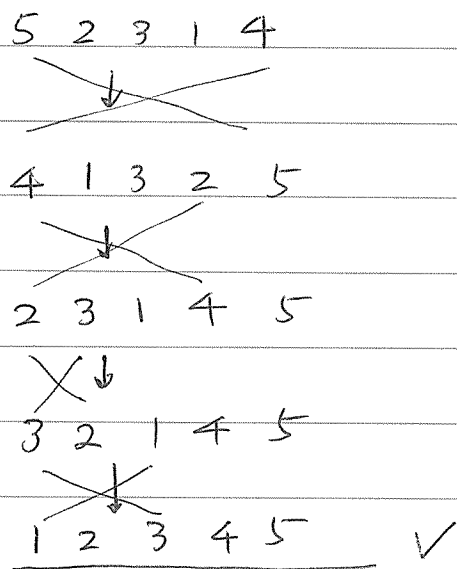
L. Bulteau, Guillaume and Irena Rusu (France)
↓
Laurent

prove that the pancake flipping problemⁱⁿ its original (unburnt) variant, is NP-hard. (Nov. 2011)

Burnt version : 有一边註明是烤焦一些。

6

Example



Let $f(n)$ (and $g(n)$) denote the maximum number of flips required to sort (flip) the unburnt pancakes (burnt).
"n" pancakes

(•) $f(n)$ is known for $n \leq 19$.

(•) $g(n)$ is known for $n \leq 17$.

(•) $\frac{15n}{14} \leq f(n) \leq \frac{18n}{11} + O(1)$.

(•) $\lfloor \frac{(3n+3)}{2} \rfloor \leq g(n) \leq 2n-6$. (Gates et al. considered burnt version first)
↑
for $n \geq 16$

Theorem [G. and P.]

$$17n < f(n) < (5n + 5/2) \dots$$

Defn. (The Breakpoint Graph)

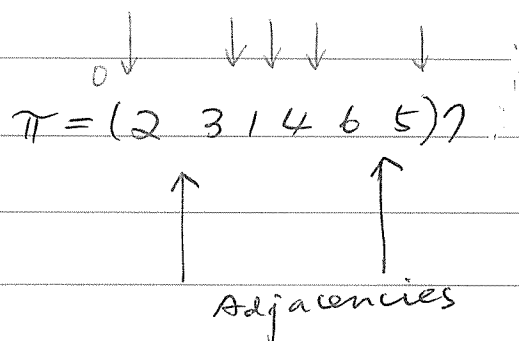
②

Let $\tilde{\pi} = \pi_0 \pi_1 \pi_2 \dots \pi_n \pi_{n+1}$ where π is a permutation $\pi_1 \pi_2 \dots \pi_n$.
For $0 \leq i \leq n$, a pair (π_i, π_{i+1}) of π is called an adjacency if $|\pi_i - \pi_{i+1}| = 1$ and a breakpoint otherwise.

(Fact) An identity permutation has no breakpoints.

(Fact) Sorting by reversals corresponds to eliminating breakpoints.

Example



So, there are two adjacencies and five breakpoints.

(Fact 2) Every reversal can eliminate at most "2" breakpoints.

(Fact 3) Let $b(\pi)$ be the number of breakpoints in π .
Then the reversal distance of π , $\boxed{d(\pi) \geq \frac{b(\pi)}{2}}$.

Note Sometimes, a reversal may eliminate one breakpoint and create another breakpoint.

$2 \ 3 \ 1 \ 4 \ 6 \ 5 \ 7$
 $\quad \quad \quad \times$

Reversal Distance Problem

Given two permutations in S_n , find a shortest series of reversals that transforms one permutation to the other.

k-reversal Distance Problem

→ $p(i, j)$ with $j - i = k$
(Fixed-length reversals)

Sorting by Reversals Problem

Transforms a permutation into identity permutation.

(*) Simple Reversal Sort

```

1  for  $i \leftarrow 1$  to  $n-1$ 
2       $j \leftarrow$  position of element  $i$  in  $\pi$  (i.e.,  $\pi_j = i$ )
3      if  $j \neq i$ 
4           $\pi \leftarrow \pi \circ p(i, j)$ 
5          output  $\pi$ 
6      if  $\pi$  is the identity permutation
7          return
  
```

Example

$\underline{6} \ 1 \ 2 \ 3 \ 4 \ 5 \rightarrow 1 \ \underline{6} \ 2 \ 3 \ 4 \ 5 \rightarrow 1 \ 2 \ \underline{6} \ 3 \ 4 \ 5 \rightarrow \dots \rightarrow 1 \ 2 \ 3 \ 4 \ 5 \ 6$
 $\pi_2 = 1 \qquad \pi_3 = 2$

Clever

$6 \ 1 \ 2 \ 3 \ 4 \ 5 \rightarrow 6 \ 5 \ 4 \ 3 \ 2 \ 1 \rightarrow 1 \ 2 \ 3 \ 4 \ 5 \ 6$

Breakpoint Reversal Sort (π)

- 1 while $b(\pi) > 0$
- 2 Among all reversals, choose reversal ρ minimizing $b(\pi \circ \rho)$
- 3 $\pi \leftarrow \pi \circ \rho$
- 4 output π
- 5 return

- Can this algorithm stop?
- Minimizing $b(\pi \circ \rho)$ is not easy.

Definition (Strip)

A strip in a permutation π is an interval between two consecutive breakpoints, i.e., a maximal segment without break points. A strip is increasing (decreasing) if the consecutive elements are in increasing (decreasing) order.

e.g.

0 • 2 1 • 3 4 5 • 8 7 6 • 9

increasing strip

decreasing strip

Theorem (The above algorithm stops)

If a permutation π contains a decreasing strip, then there is a reversal ρ that decreases the number of breakpoints in π , i.e., $b(\pi \circ \rho) < b(\pi)$.

The # of breakpoints in π

Proof. Among all decreasing strips in π choose the strip containing the

$k-1$. Therefore, $k-1$ belongs to an increasing strip; moreover, $k-1$ terminates this strip. (不然它后面就是 k) Therefore, k and $k-1$ correspond to "two" breakpoints. (0 1 2 7 6 5 8 4 3 9) Now, reverse the segment between k and $k-1$ (containing k but not $k-1$). This reversal will decrease the ^{number of} breakpoints by at least "one".

(0 1 2 3 4 8 5 6 7 9)

↓
(0 1 2 3 4 8 5 6 7 9)

↓
0 1 2 3 4 8 7 6 5 9

↓
0 1 2 3 4 5 6 7 8 9

Another example.

0 8 2 7 6 5 1 4 3 9

↓

0 2 8 7 6 5 1 4 3 9

↓

0 2 3 4 1 5 6 7 8 9

↓

0 4 3 2 1 5 6 7 8 9

↓

0 1 2 3 4 5 6 7 8 9

How about "no" decreasing strips.

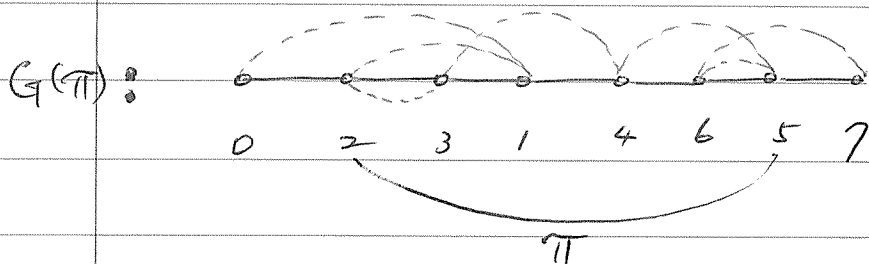
(0 1 5 6 7 2 3 4 8 9)

8

More Accurate Estimation

Definition (Breakpoint graph of π)

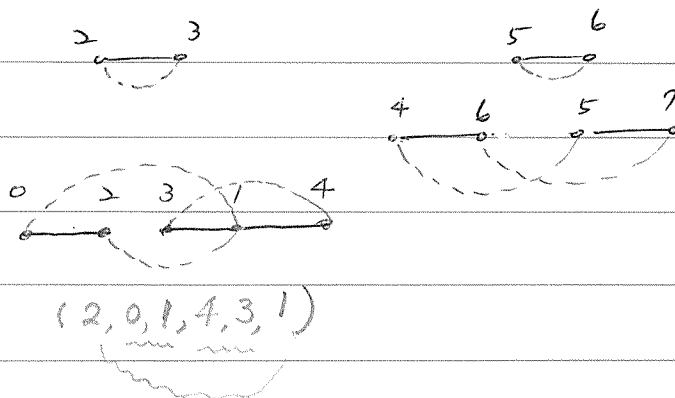
A breakpoint graph of a permutation π is a 2-edge-colored graph $G(\pi)$ defined on $n+2$ vertices $\pi_0, \pi_1, \dots, \pi_n, \pi_{n+1}$.
 For each $0 \leq i \leq n$, (π_i, π_{i+1}) is colored by "b",
 and (π_i, π_j) is also an edge if $|\pi_i - \pi_j| \leq 1$ colored by "g".



The breakpoint graph of π .

(Fact) $G(\pi)$ is an eulerian graph, in fact a balanced 2-colored eulerian graph.

(Fact) $G(\pi)$ can be decomposed into 2-colored circuits.



Note $G(\pi)$ can be decomposed into "5" circuits but two

Bafna and
[Perzner, 1996, SIAM Computing 25]

⑨

Theorem Let $c(\pi)$ denote the maximum number of edge-disjoint alternating cycles obtained from the decomposition of $G(\pi)$. Then $d(\pi) \geq n+1 - c(\pi)$.

Proof. Omitted. (观察: 对于距离而言, 每一个 alternating cycle 至少都可以节省一次 reversal.)

Point Reducing the reversal distance problem to the maximal cycle decomposition problem.