Lecture 2

10,2 ~ 10,3

Sequence Comparison (Continued)

$$\pi = \begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ x_1 & x_2 & x_3 & \cdots & x_n \end{pmatrix} \in S_n$$ (Symmetric group defined on $\{1, 2, 3, \cdots, n\}$)

π is denoted by $\langle x_1 x_2 \cdots x_n \rangle$.

Definition (Increasing subsequence of a permutation).

An increasing subsequence of a permutation $\pi = \langle x_1 x_2 \cdots x_n \rangle$ is a sequence of indices $1 \le i_1 \le i_2 \le \cdots \le i_k \le n$ s.t. $x_{i_1} < x_{i_2} < \cdots < x_{i_k}$.

(*) Finding the longest increasing subsequence (LIS) of a permutation $\pi = \langle x_1 x_2 \cdots x_n \rangle$ is equivalent to finding the longest common subsequence (LCS) of π and $\langle 1 2 3 \cdots n \rangle$.

(**) For sure, we can also use dynamic algorithm to find the answer. But, we shall use a non-dynamic programming approach to find an ILS in what follows.

Young Tableaux

Definition (Partition of integer)

A partition of an integer n is a sequence of positive integers $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_\ell$ s.t. $\sum_{i=1}^{\ell} \lambda_i = n$.

Notation:

$$\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_\ell) \longmapsto \lambda \vdash n \quad (\text{reading } \underline{\lambda \text{ partitions } n})$$

or $(\lambda_1, \lambda_2, \cdots, \lambda_\ell) \vdash n$.

Definition (Young diagram of shape λ)

The Young diagram of shape λ is an array of n cells into ℓ left-justified rows with row i containing λ_i cells for $i = 1, 2, \cdots, \ell$.

left-justified (*) ↓ 從上而下

(*) right- 有的資料用由左而右

Examples

1	2	5	8
4	7		
6			
9			

$(4,2,1,1) \vdash 8$

1	2	3	8
4	5		
6	7		
9			

$(4,2,2,1) \vdash 9$

↑ Young Tableau

Definition (Young Tableaux)

A Young tableau (of shape λ) is an array obtained by replacing the cells of the Young diagram λ with the members $1, 2, \cdots, n$ bijectively. A tableau is standard if its rows and columns are increasing sequences.

Permutation and (P,Q)-bitableau

Every permutation π of S_n can be represented by a pair of tableaux $(\lambda \vdash n)$, and there is a bijection between permutations and pairs of tableaux.

Note ① proposed first; Robinson 1938,

② In connection with representation theory, Schensted 1961,

③ Knuth generalized the algorithm for the case of LCS, 19?

→ RSK algorithm

$$\pi \xrightarrow{RSK} (P,Q)$$

$$\{\text{Conjugacy classes of } S_n\}$$

$$\xleftrightarrow{1-1} \{\text{partitions of } n\}$$

$$\lambda \vdash n$$

$$\hookrightarrow \{\text{Irreducible Representations of } S_n\}$$

More about Representation Theory

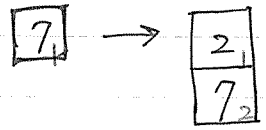
Theorem Let d_λ be the number of distinct standard tableaux corresponding to $\lambda \vdash n$. Then

$$|S_n| = n! = \sum_{\lambda \vdash n} d_\lambda^2.$$

Note Tableaux 不是 Tableau 的複數。

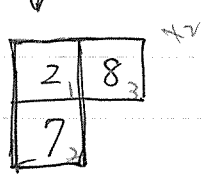
Example $\pi = \langle 7\ 2\ 8\ 1\ 3\ 4\ 10\ 6\ 9\ 5 \rangle \longleftrightarrow (P, Q)$

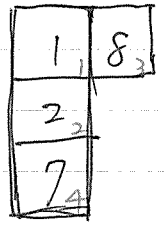
P:

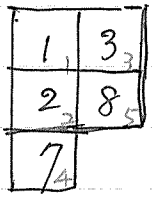


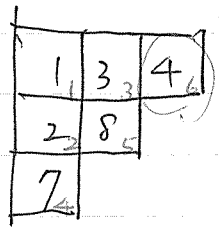
Q

右下角，依嵌入順序填入 1, 2, …, 10 在相同 shape 的 Tableau 中。

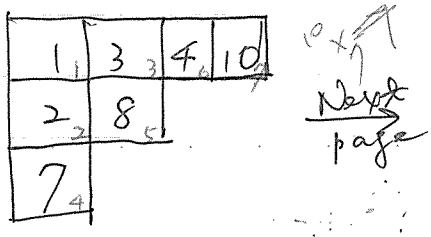


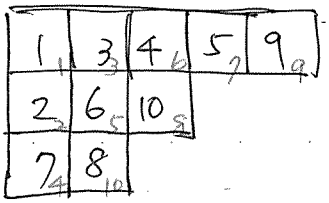






(註) 觀察第一列的數字個數即為 LIS.





1 3 4 10 ← 6 = x_8

2 8

7

1 3 4 6

2 8 ← 10

7

1 3 4 6 ← 9 = x_9

2 8 10

7

1 3 4 6 9 ← 5 = x_{10}

2 8 10

7

1 3 4 5 9

2 8 10 ← 6

7

1_1 3_3 4_6 5_7 9_9

2_2 6_5 10_8

$\wedge$ 8

Definition (Partial tableau)

A partial tableau P is a Young diagram with distinct entries whose rows and columns increase.

For a row R and an element x, define x_R^+ as the smallest element of R greater than x and x_R^- as the largest element of R smaller than x. For x not in P, let the "row insertion of x into P" be obtained by the following algorithm:

$R \leftarrow$ the first row of P

While x is less than some element of row R

Replace x_R^+ by x in R

$x \longleftarrow x_R^+$

$R \leftarrow$ next row (down)

Place x at the end of row R.

(If x is greater than every element of R)

Algorithm for finding (P, Q) from π.

(P, Q) will be obtained by a sequence of tableaux,

i.e. $(P_0, Q_0) = (\emptyset, \emptyset), (P_1, Q_1), (P_2, Q_2), \cdots, (P_n, Q_n) = (P, Q)$.

(Note 1) Compare (Refer) the following statements with example in next page.

(Note 2) (P_k, Q_k) is obtained from inserting x_k into P_k and put k in corresponding cell to make sure P_i's and Q_i's are of the same shape.

Starting from $(\emptyset, \emptyset)$, insert x_1 to obtain $(P_1, Q_1) =$ $(\boxed{x_1}, \boxed{1})$ respectively. Then, insert x_k to P_{k-1} to obtain P_k, and thus corresponding Q_k for $k = 2, \cdots, n$.

Theorem The map $\pi \xrightarrow{RSK} (P,Q)$ is a bijection between elements of S_n and pairs of standard tableaux of the same shape $\lambda \vdash n$

Proof. It suffices to prove that a pair of tableaux (P,Q) for $\lambda \vdash n$ can determine a unique permutation. We shall obtain this permutation starting from finding x_n and then $x_{n-1}, x_{n-2}, \cdots, x$

Assume that (P_k, Q_k) has been constructed, $k = n, n-1, \cdots, 2$.

Then, we will find (P_{k-1}, Q_{k-1}) and x_k. First, we find the cell (i,j) containing k (the largest number) in Q_k.

$P_k(i,j)$ must have been the last element to be displaced in the construction of P_k, and thus "x_k" can be found. Now, we shall use the following procedure to delete $P_k(i,j) = x_k$ from P_k. This gives P_{k-1} and Q_{k-1} can be obtained accordingly

Set $x \leftarrow P_k(i,j)$ and erase $P_k(i,j)$

(*) 第一列的上一列第"0"列。 $R \leftarrow$ the $(i-1)$-st row of P_k

(**) $\bar{x}_R$ 为比 x 小的最大数(R中) While R is not the zeroth row of P_k

Replace $\bar{x}_R$ by x in R

(***) 可以往上移就移 $R \leftarrow$ next row up

$x_k \leftarrow x$

∎

Theorem The length of the LIS of π is the length of the first row of P (respectively Q) where $\pi \longleftrightarrow (P, Q)$ (for tableaux P and Q).

Proof. Let $\pi = \langle x_1 x_2 \cdots x_n \rangle$. It suffices to show that if x_k ($2 \le k \le n$) enters P_{k-1} in column j, then the longest increasing subsequence of π ending in x_k has length j. (Therefore, the one "x_k" ending at the column with largest index gives the LIS.) Since P is a standard tableau, the length of the first row of P gives the answer for LIS.

Since x_k enters P_{k-1} in column j, there exists a y in the cell $(1, j-1)$ of P, $y < x_k$. By induction, there exists a subsequence (increasing) of length $j-1$ ending at y and thus we have an increasing subsequence ending at x_k of length j. In fact, the longest one ending at x_k is of length j. Suppose not. Assume that there exists a longer one ending at x_i, $i \le k-1$. This implies that x_i enters into P_{i-1} at column index larger than j.

Now, consider the cell $(1,j)$ in P_i, let the entry be y.

Clearly, $y \leq x_i < x_k$. This is impossible since by RSK, x_k enters y in order to get x_k ~~end at top row~~ P_{k-1} in column j. ▨

Average Length of LCS

Let $\mathcal{V}$ and $\mathcal{W}$ be two sets of n-letter strings over the same alphabet. Let $p: \mathcal{V} \times \mathcal{W} \to \mathbb{R}$ be a measure, i.e., $\forall\ V \in \mathcal{V}, W \in \mathcal{W}$, $p(V,W)$ is a real number. For simplicity, we may let $p(V,W) = \frac{1}{|\mathcal{V}||\mathcal{W}|}$.

(1) Longest increasing subsequence in random permutation

$s_{per}(n)$.

$\mathcal{V} = \{\langle 1\ 2\ 3\ \cdots\ n\rangle\}$ and $\mathcal{W} = S_n$. Let $p(V,W) = \frac{1}{n!}$.

Definition (Average length of LCS)

$$s(n) = \sum_{V \in \mathcal{V},\, W \in \mathcal{W}} s(V,W) \cdot p(V,W).$$

Definition (Average length of LIS)

$$s_{per}(n) = \sum_{W \in S_n} s(V,W) \cdot \frac{1}{n!}. \quad (V = \langle 1\ 2\ 3\ \cdots\ n\rangle)$$

Ulam (1961), proposed to find $s_{per}(n)$.

Hammersley (1972), proved that $\lim_{n\to\infty} \frac{s_{per}(n)}{\sqrt{n}} = s_{per}$, i.e., the limit exists.

The best result

Logan and Shepp, Vershik and Kerov (1977) proved that $s_{per} = 2$. This implies that in (average) the longest increasing subsequence of an n-letter permutation is about $2\sqrt{n}$.

In fact, in 1999, Baik et al. proved that

$$s_{per}(n) = 2\sqrt{n} - \mu n^{1/6} + o(n^{1/6}) \text{ where } \mu = 1.711\ldots.$$

NOTE By the correspondence between permutations and Young tableaux, this implies that we are interested in finding the average length of the first rows.

NOTE More analysis technique is needed when "average" LCS's are considered.

(**) We let p_n be the probability that $P(\pi)$ (tableau $\leftrightarrow \pi$) contains n in the first row. Then

$$p_n \leq \frac{1}{\sqrt{n}}. \quad \text{(See p. 108 for proof.)}$$

Peuzner's book

Shortest Superstring Problem

Example set of strings: {001, 100, 101}

001100101 (Concatenation)

0010100 (Better)

Problem Given a set of strings $\vec{s}_1, \vec{s}_2, \cdots, \vec{s}_n$, find the shortest string $\vec{s}$ such that $\forall\ i = 1, 2, \cdots, n$, $\vec{s}_i$ is a a substring of $\vec{s}$.

Example set of strings {CATGC, CTAAGT, GCTA, TTCA, ATGCATC}.
① ② ③ ④ ⑤

$\vec{s}$: GCTAAGTTCATGCATC

Greedy Algorithm (把可以接在一起的依次接好，重叠部份愈多愈好。)

Step 1. ① + ⑤ CATGCATC ⑥

Step 2. ② + ③ GCTAAGT ⑦

Step 3. ④ + ⑥ TTCATGCATC ⑧

Step 4 ⑦ + ⑧ GCTAAGTTCATGCATC

(Answer)

Idea (Prefix graph)

Definition (prefix and overlap)

prefix $(\vec{s}_i, \vec{s}_j)$: First "letters" of $\vec{s}_i$ where overlap $(\vec{s}_i, \vec{s}_j)$ is removed.

overlap $(\vec{s}_i, \vec{s}_j)$: The maximum overlap between $\vec{s}_i$ and $\vec{s}_j$.

$\vec{s}_i$: ACGGCTAT

$\vec{s}_j$: CTATTAGC

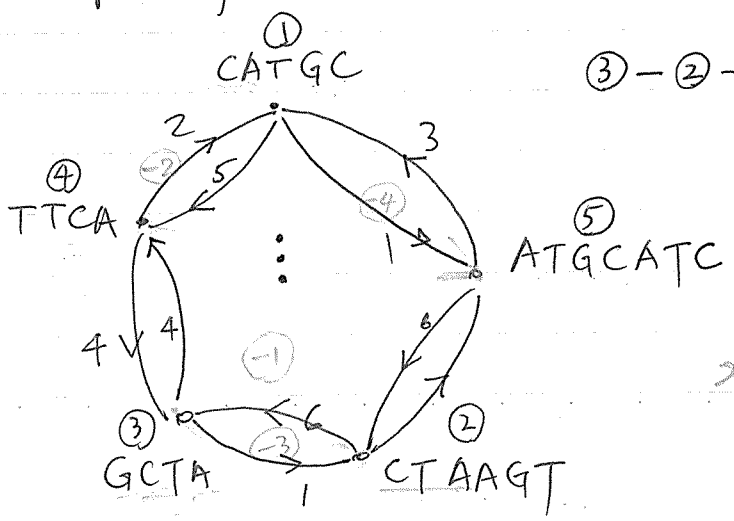
overlap $(\vec{s}_i, \vec{s}_j)$ = CTAT

prefix $(\vec{s}_i, \vec{s}_j)$ = ACGG

Definition (prefix graph G)

Use the idea of Hamiltonian path!

The prefix graph G of a set of strings $\{\vec{s}_1, \vec{s}_2, \cdots, \vec{s}_n\}$ is a double-weighted complete digraph such that $V(G) = \{\vec{s}_1, \vec{s}_2, \cdots, \vec{s}_n\}$ and the weight of $(\vec{s}_i, \vec{s}_j)$ = $|\vec{s}_i| -$ overlap $(\vec{s}_i, \vec{s}_j)$.



③ − ② − ④ − ① − ⑤

26 − 10 = 16