

Longest Common Subsequence Problem (LCS)

Definition (Common subsequence)

Let $V = \langle v_1, v_2, \dots, v_n \rangle$ and $W = \langle w_1, w_2, \dots, w_m \rangle$. A common subsequence of V and W is a sequence of indices

$1 \leq i_1 \leq i_2 \leq \dots \leq i_k \leq n$ chosen from V and a sequence

of indices $1 \leq j_1 \leq j_2 \leq \dots \leq j_k \leq m$ chosen from W such that

$$v_{i_t} = w_{j_t} \text{ for } t = 1, 2, \dots, k.$$

e.g. $V = \text{A} \text{ (T)} \text{ (C)} \text{ (T)} \text{ G} \text{ (A)} \text{ T} = \langle v_1, v_2, \dots, v_7 \rangle$

$$W = \text{(T)} \text{ G} \text{ (C)} \text{ (A)} \text{ (T)} \text{ (A)} = \langle w_1, w_2, \dots, w_6 \rangle$$

$$i_1 = 2, i_2 = 3, i_3 = 4, i_4 = 6 \Rightarrow \langle v_2, v_3, v_4, v_6 \rangle$$

$$j_1 = 1, j_2 = 3, j_3 = 5, j_4 = 6 \Rightarrow \langle w_1, w_3, w_5, w_6 \rangle$$

$$v_{i_1} = v_2 = w_1 = w_{j_1}, v_{i_2} = v_3 = w_3 = w_{j_2}, \dots$$

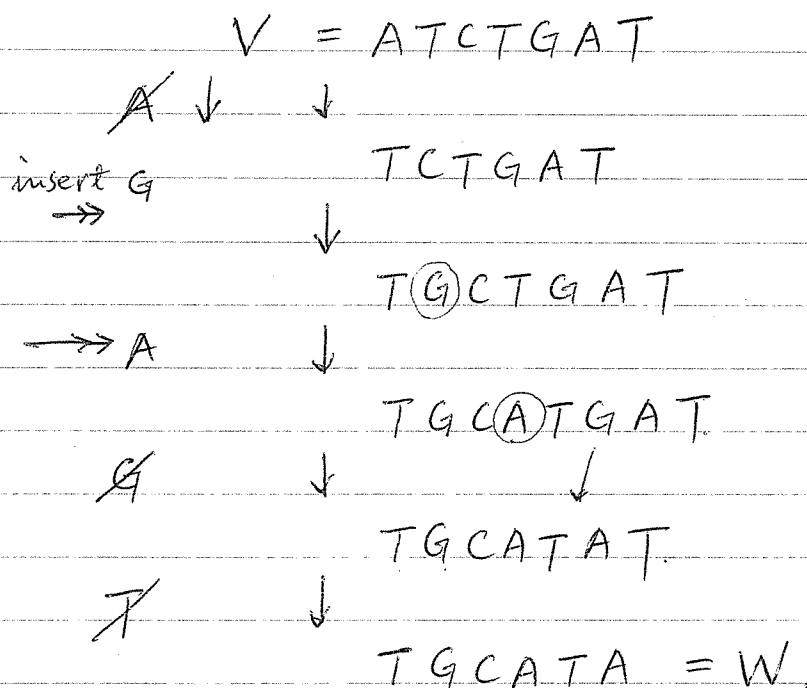
Definition ($s(V, W)$)

The length of a longest common subsequence (LCS) of V and W is denoted by $s(V, W)$.

Definition ($d(V, W)$)

The minimum number of insertions and deletions needed to transform from V into W is called the edit distance from V into W , denoted by $d(V, W)$.

e.g.

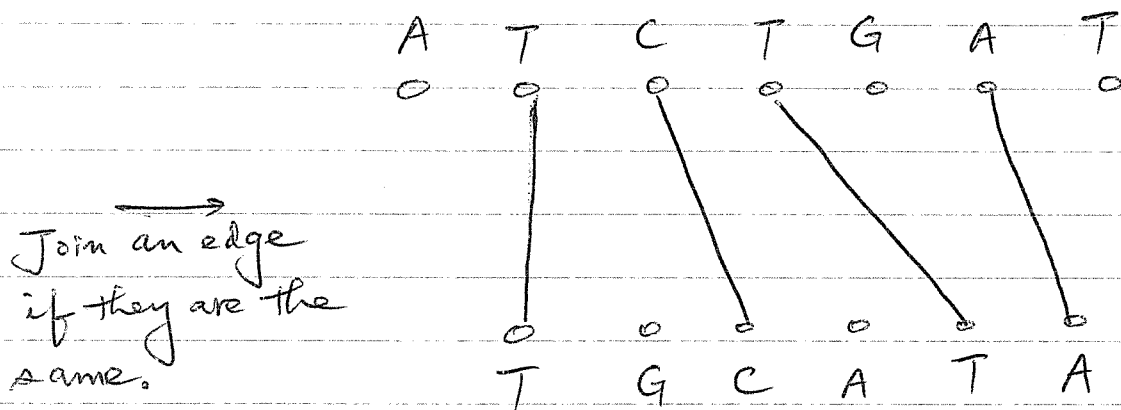


It takes two insertions and three deletions to transform V into W . Hence, $d(V, W) \leq 5$. In fact, we can verify the following property.

Proposition (Fact 1)

$d(V, W) = |V| + |W| - 2s(V, W)$, $|V|$ and $|W|$ are the lengths of V and W respectively.

Proof. We have to delete $|V| - s(V, W)$ terms from V and insert $|W| - s(V, W)$ terms to transform V into W as the best strategy. Moreover, we can edit V into W by the following idea.



So, we have to delete A, G, T in $V = ATCTGAT$ and then insert G, A to obtain W . Now, the number of vertices not incident to an edge of the matching is $d(V, W)$. ■

(*) So, it is sufficient to compute $s(V, W)$ in order to find the edit distance. (This is important in applications!)

Compute $s(V, W)$, $V = \langle v_1 v_2 \dots v_n \rangle$ and $W = \langle w_1 w_2 \dots w_m \rangle$

Algorithm (Dynamic Algorithm)

Define $s_{i,j} = s(V_i, W_j)$ where $V_i = \langle v_1 v_2 \dots v_i \rangle$ and $W_j = \langle w_1 w_2 \dots w_j \rangle$, $s_{i,0} = s_{0,j} = 0$, $\forall 1 \leq i \leq n$ and $1 \leq j \leq m$.

Step 1

$$s_{i,j} = \max \begin{cases} s_{i-1,j} \\ s_{i,j-1} \\ s_{i-1,j-1} + 1 \end{cases} \text{ if } v_i = w_j.$$

One more common term.

	*	T	G	C	A	T	A	
*	0 ₀	0 ₁	0 ₂	0 ₃	0 ₄	0 ₅	0 ₆	
A	0 ₁	0 ₂	0 ₃	0 ₄	1 ₃	1 ₄	1 ₅	
T	0 ₂	1 ₁	1 ₂	1 ₃	1 ₄	2 ₃	2 ₄	
C	0 ₃	1 ₂	1 ₃	2 ₂	2 ₃	2 ₄	2 ₅	
T	0 ₄	1 ₃	1 ₄	2 ₃	2 ₄	3 ₃	3 ₄	
G	0 ₅	1 ₄	2 ₃	2 ₄	2 ₅	3 ₄	3 ₅	
A	0 ₆	1 ₅	2 ₄	2 ₅	3 ₄	3 ₅	4 ₄	
T	0 ₇	1 ₆	2 ₅	2 ₆	3 ₅	4 ₄	4 ₅	

↑
V

$$s(V, W) = 4$$

$$d(V, W) = 5 \text{ (edit distance)}$$

$$\checkmark d(V, W) = 7 + 6 - 2(4) = 5. \quad \text{By (Fact 1).}$$

Combinatorial Relations

Consider a permutation $\pi \in S_n$ (Symmetric group of order n).

(.) Find a longest increasing ^{sub-}sequence of π .

(..) Find a longest increasing subsequence of π is equivalent to find the LCS between π and the permutation

$$(1 \ 2 \ 3 \ \dots \ n).$$

$$\begin{pmatrix} 1 & 2 & 3 & \dots & n \\ 1 & 2 & 3 & \dots & n \end{pmatrix}$$

identity permutation with cycle representation
 $(1)(2)(3) \dots (n)$

(*) Every permutation $\pi = \begin{pmatrix} 1 & 2 & 3 & \dots & n \\ x_1 & x_2 & x_3 & \dots & x_n \end{pmatrix}$ denoted by

$\langle x_1 \ x_2 \ \dots \ x_n \rangle$ has either an increasing or a decreasing subsequence of length at least $\underline{\sim \sqrt{n}}$.

Review (Discrete Mathematics)

Every sequence $\langle a_1, a_2, \dots, a_{n+1} \rangle$ of real numbers contains either an increasing subsequence or a decreasing subsequence of length $n+1$.

Proof. Suppose that there is no increasing subsequence of length $n+1$. For $1 \leq k \leq n+1$, let the longest increasing

subsequence starting at a_k be m_k . By assumption, $m_k \leq n$ for each $k \in \{1, 2, \dots, n+1\}$. By Pigeon-hole principle, there exists an $1 \leq n' \leq n$, such that $m_{k_1} = m_{k_2} = \dots = m_{k_{n+1}} = n'$. Now, consider the subsequence $W = \langle a_{k_1}, a_{k_2}, \dots, a_{k_{n+1}} \rangle$. If $a_{k_i} \leq a_{k_j}$, then m_{k_j} can be larger, a contradiction. Hence, W is a decreasing subsequence of length $n+1$. 

Remark

Pigeon-hole principle plays an important role in showing the existence of certain configurations. Make sure you understand how to apply the principle.

Sequence Comparison

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

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

Definition (Increasing subsequence of a permutation).

An increasing subsequence of a permutation

$\pi = \langle x_1, x_2, \dots, x_n \rangle$ is a sequence of indices $1 \leq i_1 < i_2 < \dots < i_k \leq n$

s.t. $x_{i_1} < x_{i_2} < \dots < x_{i_k}$.

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

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

(As finding $A(V, W)$.)

Young Tableaux

Definition (partition of integer)

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

Notation:

$$\lambda = (\lambda_1, \lambda_2, \dots, \lambda_l) \mapsto \lambda \vdash n \text{ (reading } \lambda \text{ partitions } \underline{n} \text{)}$$

$$\text{or } (\lambda_1, \lambda_2, \dots, \lambda_l) \vdash n.$$

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 $i = 1, 2, \dots, l$.

Examples

1	2	5	8
4	7		
6			
9			

$(4, 2, 1, 1)$

Not a
Young tableau

1	2	3	8
4	5		
6	7		
9			

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

Standard
Young Tableau

left-justified
(*)
↓ 從上而下
right-justified
(*)
有的資料用由左而右

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 $1, 2, \dots, n$ 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, 1977

→ RSK algorithm

$$\pi \xrightarrow[\text{bijection}]{\text{RSK}} (P, Q)$$

{ Conjugacy classes of S_n }

$\longleftrightarrow$ { partitions of n }

$\lambda \vdash n$

$\hookrightarrow$ { 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 $\neq$ Tableau 的个数。

We start with an example.

11

Example

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

P: $\boxed{7_1} \rightarrow \begin{bmatrix} 2_1 \\ 7_2 \end{bmatrix} \leftarrow \text{insert } 2_1 \quad \leftarrow \text{(上升序列长度 1)}$

(Q)

在下角,
依填入顺序

填入 1, 2, ..., 10

在相同 shape

的 Tableau 中。

$$\begin{bmatrix} 2_1 & 8_3 \\ 7_2 \end{bmatrix} \quad \leftarrow \text{接在第一列後面} \quad \leftarrow \text{(上升 2)}$$

$$\begin{bmatrix} 1_1 & 8_3 \\ 2_2 \\ 7_4 \end{bmatrix} \quad \leftarrow \text{insert } 1_1 \quad \leftarrow \text{insert } 2_2$$

$$\begin{bmatrix} 1_1 & 3_3 \\ 2_2 & 8_5 \\ 7_4 \end{bmatrix} \quad \leftarrow \text{insert } 3_3 \quad \leftarrow 8 \text{ 比较大放在右边}$$

$$\begin{bmatrix} 1_1 & 3_3 & 4_6 \\ 2_2 & 8_5 \\ 7_4 \end{bmatrix} \quad \leftarrow 4 \text{ 放右边} \quad \leftarrow \text{(上升 3)}$$

$$\begin{bmatrix} 1_1 & 3_3 & 4_6 & 10_7 \\ 2_2 & 8_5 \\ 7_4 \end{bmatrix}$$

$$\begin{bmatrix} 1_1 & 3_3 & 4_6 & 5_7 & 9_9 \\ 2_2 & 6_5 & 10_8 \\ 7_4 & 8_{10} \end{bmatrix}$$

(註) 觀察第一列
的數字個數
即為 LIS.

(上升 4)

(上升 5)

Next page

Step by Step

$$\pi = \langle 7, 2, 8, 1, 3, 4, 10, 6, 9, 5 \rangle$$

$$\begin{array}{c} \downarrow \\ \begin{array}{|c|} \hline 7_1 \\ \hline \end{array} \\ \downarrow \\ \begin{array}{|c|} \hline 2_1 \\ \hline \end{array} \\ \begin{array}{|c|} \hline 7_2 \\ \hline \end{array} \\ \downarrow \\ \begin{array}{|c|c|} \hline 2_1 & 8_3 \\ \hline \end{array} \\ \begin{array}{|c|} \hline 7_2 \\ \hline \end{array} \end{array}$$

$$\begin{array}{c} \downarrow \\ \begin{array}{|c|c|} \hline 1_1 & 8_3 \\ \hline \end{array} \\ \begin{array}{|c|} \hline 2_2 \\ \hline \end{array} \\ \begin{array}{|c|} \hline 7_4 \\ \hline \end{array} \end{array}$$

$$\begin{array}{c} \downarrow \\ \begin{array}{|c|c|} \hline 1_1 & 3_3 \\ \hline \end{array} \\ \begin{array}{|c|c|} \hline 2_2 & 8_5 \\ \hline \end{array} \\ \begin{array}{|c|} \hline 7_4 \\ \hline \end{array} \end{array}$$

$$\begin{array}{c} \downarrow \\ \begin{array}{|c|c|c|} \hline 1_1 & 3_3 & 4_6 \\ \hline \end{array} \\ \begin{array}{|c|c|} \hline 2_2 & 8_5 \\ \hline \end{array} \\ \begin{array}{|c|} \hline 7_4 \\ \hline \end{array} \end{array}$$

$$\begin{array}{c} \downarrow \\ \begin{array}{|c|c|c|c|} \hline 1_1 & 3_3 & 4_6 & 10_7 \\ \hline \end{array} \\ \begin{array}{|c|c|} \hline 2_2 & 8_5 \\ \hline \end{array} \\ \begin{array}{|c|} \hline 7_4 \\ \hline \end{array} \end{array}$$

① 右下角数字代表新增的格子

② 向右填格子代表上升的量

(P, Q)

$$\begin{array}{|c|c|c|c|c|} \hline 1_1 & 3_3 & 4_6 & 5_7 & 9_9 \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|} \hline 2_2 & 6_5 & 10_8 \\ \hline \end{array}$$

$$\begin{array}{|c|c|} \hline 7_4 & 8_{10} \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|c|c|} \hline 1_1 & 3_3 & 4_6 & 6_7 & 9_9 \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|} \hline 2_2 & 8_5 & 10_8 \\ \hline \end{array}$$

$$\begin{array}{|c|} \hline 7_4 \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|c|} \hline 1_1 & 3_3 & 4_6 & 6_7 \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|} \hline 2_2 & 8_5 & 10_8 \\ \hline \end{array}$$

$$\begin{array}{|c|} \hline 7_4 \\ \hline \end{array}$$

✓ 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

$x \leftarrow x_R^+$

$R \leftarrow$ 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. $(P_0, Q_0) = (\emptyset, \emptyset), (P_1, Q_1), (P_2, Q_2), \dots, (P_n, Q_n) = (P, Q)$.

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

(Note 2) (P_k, Q_k) is obtained from inserting x_k into P_{k-1} and put k in corresponding cell to make sure P_k 's and Q_k '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 $k = 2, \dots, n$.

Permutation $\langle 4\ 2\ 3\ 6\ 5\ 1\ 7 \rangle$, $n=7$

$$x_R^+ = 4$$

↓

$$2 \leftarrow 4$$

P's

4

2
4

2	3
4	

2	3	6
4		

2	3	5
4	6	

1	3	5
2	6	
4		

1	3	5	7
2	6		
4			

Q's

1

1
2

1	3
2	

1	3	4
2		

1	3	4
2	5	

1	3	4
2	5	
6		

1	3	4	7
2	5		
6			

(P_1, Q_1)

(P_2, Q_2)

(P_3, Q_3)

(P_4, Q_4)

(P_5, Q_5)

(P_6, Q_6)

$(P_7, Q_7) = (P, Q)$

$3 > 2$
($x=3$ is the 3rd row
the largest)

$6 > 2$ and $6 > 3$

5 取代 6,
6 到下-列.

$$x_R^+ = 2$$

$$1 \leftarrow 2 \downarrow \text{下-列}$$

$$x_R^+ = 4$$

$$2 \leftarrow 4 \downarrow \text{下-列}$$

$7 > 1, 3, 5$

How to return to $\langle 4\ 2\ 3\ 6\ 5\ 1\ 7 \rangle$ from knowing (P, Q) ?

① Use (P_7, Q_7) find $x_7 = 7$ since 7 is in the 7th position in Q_7 and it is in the 1st row.

② Delete $\boxed{7}$, we have
(in P_7 and Q_7)

1	3	5
2	6	
4		

1	3	4
2	5	
6		

$x = 4$, 往上一列看 "2" 小于 4 且为最大,

用 4 取代 2; $x = 2$ 再往上一列看用 2 取代 1.

③ Delete "1",

2	3	5
4	6	

1	3	4
2	5	

, $x_6 = 1$

④ $x = 6$, 往上一列 5 是 x_5 , 用 5 取代 6, $x_5 = 5$

⑤

2	3	6
4		

1	3	4
2		

 $x_4 = 6$

⑥

2	3
4	

1	3
2	

 $x_3 = 3$

⑦

2
4

1
2

 $x = 4$, 往上一列 2 < 4, 用 2 取代 4, $x_2 = 2$
 $x_1 = 4$

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 $x_{n-1}, x_{n-2}, \dots, x_1$.

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

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

$P_k(i, j)$ must have been the last element to be displaced in the construction of P_k , and thus " x_k " can be found. Now, How?

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 $P_k(i, j)$

(*) $\begin{array}{l} \text{第 } i \text{ 列的上} \\ \text{一行第 } 0 \text{ 列} \end{array}$ $R \leftarrow \text{the } (i-1)\text{-st row of } P_k$

(**) x_R 为 $x-1$, While R is not the zeroth row of P_k
 的最大数 (R 中)

Replace x_R by x in R

(***) 可以往上移动就移 $R \leftarrow \text{next row up}$
 $x_k \leftarrow x$

Lemma If $\pi = \langle x_1 x_2 x_3 \dots x_n \rangle$ and x_k enters P_{k-1}

(to become P_k) in column j , then the longest increasing subseq. of π ending in x_k has length j .

Proof. By induction on k . Clearly, it is true for $k=1$.

So, suppose it holds for all values up to $k-1$. First, we

claim, indeed, there exists an increasing subsequence of length

j ending in x_k . Remind that x_k enters P_{k-1} in column j .

Let y be the element of P_{k-1} in cell $(1, j-1)$. Then, $y < x_k$ since

x_k enters in column j . By induction, there is an increasing

subseq. ending in y with length $j-1$, thus we have an increasing

subseq. of length j ending in x_k .

Now, we prove that there cannot be a longer increasing

subseq. ending in x_k . If there exists one, let x_i be an
(length $> j$)

element ~~preceding~~ x_k . By induction, x_i enters (earlier) to
(in this increasing subsequence)

the right of column j . Now, the element in the cell $(1, j)$, z ,

of P_i satisfies $z \leq x_i < x_k$. So, in order to enter column j ,

$x_k < z$, a contradiction. (By RSK algorithm.)



Theorem Let $P(\pi)$ be the standard tableau of π . Then, the length of the longest increasing subseq. of π is the length of the first row of $P(\pi)$.

Proof. $P(\pi)$ is a standard tableau, (The first row has largest length.)
(maximum)

Average Length of Longest Common Subsequence

(1) Let $\mathcal{U}$ and $\mathcal{W}$ be two sets of sequences of length n .

(2) Let $p : \mathcal{U} \times \mathcal{W} \rightarrow \mathbb{R}$ be a measure. For example,

let $p : \mathcal{U} \times \mathcal{W} \rightarrow \left\{ \frac{1}{|\mathcal{U}| \cdot |\mathcal{W}|} \right\}$ a constant mapping.

(*) The average length of LIS between $\mathcal{U}$ and $\mathcal{W}$ is

$$s(n) = \frac{1}{|\mathcal{U}| \cdot |\mathcal{W}|} \cdot \sum_{\substack{V \in \mathcal{U}, \\ W \in \mathcal{W}}} s(V, W),$$

(**) Average longest ^{increasing} subseq. in random permutation in S_n ,

$s_{\text{per}}(n) = \frac{1}{1 \cdot n!} \sum_{\pi \in S_n} \text{lis}(\pi)$ where $\text{lis}(\pi)$ is the ^{length of the} longest increasing subsequence of π .

(*) The problem of finding $s_{\text{per}}(n)$ was raised by Ulam, 1961; and Hammersley in 1972 proved that

$$\lim_{n \rightarrow \infty} \frac{s_{\text{per}}(n)}{\sqrt{n}} = s_{\text{per}} \text{ (a constant).}$$

He also proved that $\frac{\pi}{2} \leq s_{\text{per}} \leq e$. (Surprisingly!)

Kingman, 1973, $1.59 < s_{\text{per}} < 2.49$.

Logan and Shepp, 1977, and Vershik and Kerov, 1977,

✓ $s_{\text{per}} = 2$.

How about Longest Common Subsequence $s_k(n)$ in k -letter alphabet. (DNA sequences: $k=4$).

So, both $\mathcal{V}$ and $\mathcal{W}$ contain all k^n n -letter words in k -letter alphabet, $p(V, W) = \frac{1}{k^n \cdot k^n}$ for each $V \in \mathcal{V}$ and $W \in \mathcal{W}$.

Chvátal and Sankoff, 1975: $\lim_{n \rightarrow \infty} \frac{s_k(n)}{n} = s_k$ (a constant).

Conjectures (1) $\lim_{k \rightarrow \infty} (s_k \cdot \sqrt{k}) = 2$.

(2) $s_k = \frac{2}{1 + \sqrt{k}}$. (??)