

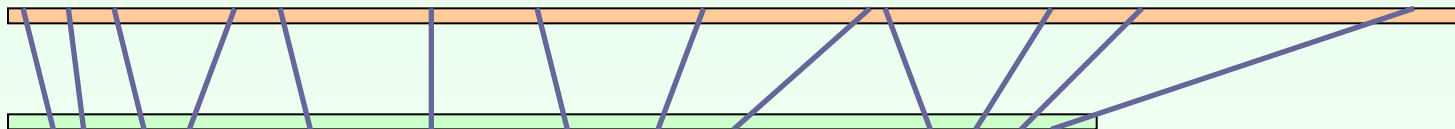
# Максимально длинная общая подпоследовательность (МДП, LCS)

- Последовательность  $U$  является подпоследовательностью  $A$ , если существует монотонно возрастающая последовательность целых чисел  $r_1 \dots r_{|U|}$  такая, что  $U[j] = A[r_j]$ ,  $1 \leq j \leq |U|$ ,  $1 \leq r_j \leq |A|$ .
- Если  $\exists p_1 \dots p_{|U|}$  такая, что  $(p_i < p_k \ \forall i < k) \ \& \ (U[j] = B[p_j])$ ,  $U$  – общая подпоследовательность  $A$  и  $B$
- Если  $\gamma(a_i \rightarrow b_j) \geq \gamma(a_i \rightarrow \varepsilon) + \gamma(\varepsilon \rightarrow b_j)$ , то при переводе  $A$  в  $B$  используются только операции вставки и устранения, а элементы, составляющие LCS, остаются без изменения.

Если  $\gamma(a_i \rightarrow b_j) = 2$  (при  $a_i \neq b_j$ ), а  $\gamma(a_i \rightarrow \varepsilon) = \gamma(\varepsilon \rightarrow b_j) = 1$ ,

$$D(A, B) = m + n - 2L(A, B), \quad m = |A|, n = |B|$$

$$2L(A, B) = m + n - D(A, B)$$



# Вычисление длины МПД:

$$L(i, j) = \begin{cases} L(i-1, j-1) + 1, & \text{если } a_i = b_j \\ \max(L(i, j-1), L(j-1, j)) & \text{в остальных случаях} \end{cases}$$

$$L(0, j) = 0; \quad L(i, 0) = 0;$$

$$0 \leq i \leq m; \quad 0 \leq j \leq n;$$

Пример. A = aacacbb; B = ababc.

$$L(A, B) = 3.$$

Всего 13 МДП:

3 – abb,

6 – aab,

4 – aac

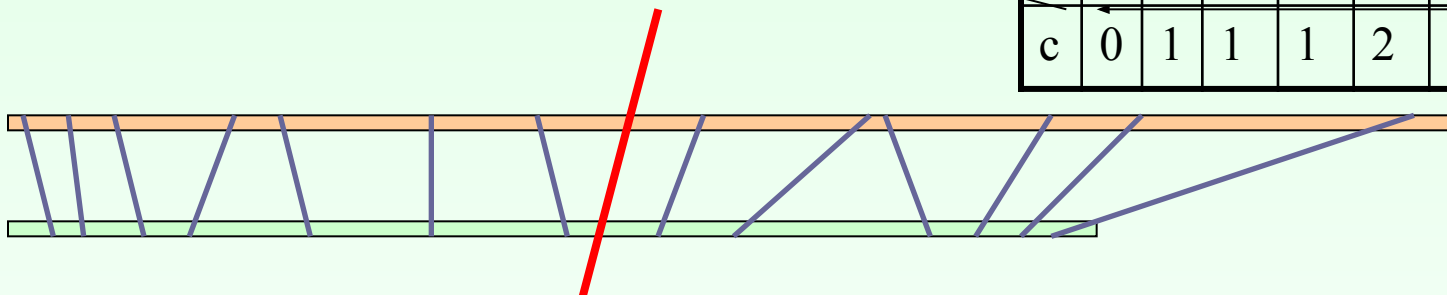
	ε	a	b	a	b	c
ε	0	0	0	0	0	0
a	0	1	1	1	1	1
a	0	1	1	2	2	2
c	0	1	1	2	2	3
a	0	1	1	2	2	3
c	0	1	1	2	2	3
b	0	1	2	2	3	3
b	0	1	2	2	3	3

# Алгоритм Хишберга (Hirschberg D.S.) – линейная память

Пусть  $L^*(i, j)$  – длина МДП текстов  $A[i + 1 : m]$  и  $B[j + 1 : n]$ .

Тогда для любого  $i$ :  $L(m, n) = \max_j \{L(i, j) + L^*(i, j)\}$

<p><math>A = \text{aaca'cbb}</math></p> <p><math>B = \text{ababc}</math></p> <p><math>L(4,0) + L^*(4,0) = 0 + 2</math></p> <p><math>L(4,1) + L^*(4,1) = 1 + 2</math></p> <p><math>L(4,2) + L^*(4,2) = 1 + 1</math></p> <p><math>L(4,3) + L^*(4,3) = \underline{2 + 1}</math></p> <p><math>L(4,4) + L^*(4,4) = 2 + 1</math></p> <p><math>L(4,5) + L^*(4,5) = 3 + 0</math></p>	<p><math>A1 = \text{aa'ca}</math></p> <p><math>B1 = \text{aba}</math></p> <p><math>L(2,0) + L^*(2,0) = 0 + 1</math></p> <p><math>L(2,1) + L^*(2,1) = \underline{1 + 1}</math></p> <p><math>L(2,2) + L^*(2,2) = 1 + 1</math></p> <p><math>L(2,3) + L^*(2,3) = 2 + 0</math></p>	<p><math>A2 = \text{cb'b}</math></p> <p><math>B2 = \text{bc}</math></p> <p><math>L(2,0) + L^*(2,0) = 0 + 1</math></p> <p><math>L(2,1) + L^*(2,1) = \underline{1 + 0}</math></p> <p><math>L(2,2) + L^*(2,2) = 1 + 0</math></p> <table border="1"> <tr> <td></td> <td><math>\epsilon</math></td> <td>c</td> <td>b</td> <td>a</td> <td>c</td> <td>a</td> </tr> <tr> <td><math>\epsilon</math></td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>b</td> <td>0</td> <td>0</td> <td>1</td> <td>1</td> <td>1</td> <td>1</td> </tr> <tr> <td>b</td> <td>0</td> <td>0</td> <td>1</td> <td>1</td> <td>1</td> <td>1</td> </tr> <tr> <td>c</td> <td>0</td> <td>1</td> <td>1</td> <td>1</td> <td>2</td> <td>2</td> </tr> </table>		$\epsilon$	c	b	a	c	a	$\epsilon$	0	0	0	0	0	0	b	0	0	1	1	1	1	b	0	0	1	1	1	1	c	0	1	1	1	2	2
	$\epsilon$	c	b	a	c	a																															
$\epsilon$	0	0	0	0	0	0																															
b	0	0	1	1	1	1																															
b	0	0	1	1	1	1																															
c	0	1	1	1	2	2																															



# Адаптивные алгоритмы вычисления длины МПД: Hunt - Shimanski

$Q_{ik}$  – наим.  $j$ , такое что  $(L(A[1 : i], B[1 : j])) = k$

$$Q_{i+1,k} = \begin{cases} \text{наим. } j, & \text{такое что } Q_{i,k-1} < j \leq Q_{ik} \text{ и } a_{i+1} = b_j \\ Q_{ik}, & \text{если такого } j \text{ не существует} \end{cases}$$

Начальные условия:

$$Q_{i,0} = 0, \quad 0 \leq i \leq n;$$

$$Q_{0k} = n + 1, \quad 1 \leq k \leq \min(m,n);$$

**Пример.**  $A = \text{a a c a s c b b c}$ ,

$B = \text{a b a b c}$ .

Трудоёмкость:  $O(r \times \log n)$ ,

где 
$$r = \sum_{i=1}^{|\Sigma|} f_i(A) \cdot f_i(B)$$

число потенциально возможных парных соответствий символов

	0	1	2	3	4	5
ε	0	6	6	6	6	6
a	0	1	6	6	6	6
a	0	1	3	6	6	6
c	0	1	3	5	6	6
a	0	1	3	5	6	6
c	0	1	3	5	6	6
b	0	1	2	4	6	6
b	0	1	2	4	6	6
c	0	1	2	4	5	6

# Адаптивные алгоритмы вычисления длины МПД: Nakatsu-Kambayashi-Yajima

$R_{i,k}$  – max  $j$ , такое что  $(L(A[i : m], B[j : n])) = k$

$$R_{i,k} = \begin{cases} \max j, & \text{такое что } R_{i+1,k} \leq j < R_{i+1,k-1} \text{ и } a_i = b_j \\ R_{i+1,k}, & \text{если такого } j \text{ не существует} \end{cases}$$

Начальные условия:

$$R_{i,0} = n + 1, \quad 0 \leq i \leq m;$$

$$R_{m+2-k,k} = 0, \quad 1 \leq k \leq \min(m,n);$$

**Пример.**  $A = \text{a a c a s c b b}$ ,

$B = \text{a b a b c}$ .

Трудоёмкость:  $O(n \times (m - L))$ ,

R	0	1	2	3	4
a	6	5	3	1	0
a	6	5	3	1	0
c	6	5	3	1	0
a	6	5	3	1	0
c	6	5	2	0	0
b	6	4	2	0	
b	6	4	0		
ε	6	0			

- **Близкие задачи:**
- задача о наикратчайшей **надпоследовательности**
- задача о **медиане** (string merging): построение текста T3, сумма переходов к которому от T1 и T2 минимальная. В зависимости от весов ред. операция в качестве T3 можно получить МДП, наименьшую надпоследовательность, один из текстов (T1 или T2), наиболее вероятного предка и т.п.
- поиск максимально длинной общей подпоследовательности для группы текстов.
- задача о максимально длинной возрастающей (убывающей) подпоследовательности для числовой перестановки

## Расстояния и меры сходства, отличные от ред. расстояния

Пусть  $T_1$  и  $T_2$  два текста.

Назовем **совместной частотной характеристикой**  $l$ -го порядка текстов  $T_1$  и  $T_2$  совокупность элементов

$$\Phi_l(T_1, T_2) = \{\varphi_{l1}(T_1, T_2), \varphi_{l2}(T_1, T_2), \dots, \varphi_{lM_l}(T_1, T_2)\}$$

где  $M_l = M_l(T_1, T_2)$  — число  $l$ -грамм, общих для обоих текстов,

а элемент  $\varphi_{li}$  ( $1 \leq i \leq M_l$ ) есть тройка:

$\ll$   $i$ -я общая  $l$ -грамма —  $x_i$ ,

частота ее встречаемости в  $T_1$  —  $F(T_1, x_i)$  и в  $T_2$  —  $F(T_2, x_i)$   $\gg$ .

Простейший **набор мер сходства**, упорядоченный по возрастанию  $l$  имеет вид:

$$q_l(T_1, T_2) = \frac{M_l(T_1, T_2)}{M_l(T_1) + M_l(T_2) - M_l(T_1, T_2)}$$

$l = 1, 2, \dots, l_{\max}(T_1, T_2)$

Мера сходства, учитывающая частоты встречаемости  $l$ -грамм:

$$\lambda(T_1, T_2) = \frac{\sum_{\alpha} \min\{F(T_1, \alpha), F(T_2, \alpha)\} \cdot |\alpha|}{\sum_{\alpha} \max\{F(T_1, \alpha), F(T_2, \alpha)\} \cdot |\alpha|}$$

где  $\alpha$  – произвольная цепочка текстов  $T_1$  и (или)  $T_2$ ,  
 $|\alpha|$  – ее длина.

(Findler N.V., Van Leeuten, 1979)



## Ранговые меры близости. Коэффициент корреляции $\tau$

Пусть  $l$ -граммы в  $\Phi_l(T_1)$  и  $\Phi_l(T_2)$  упорядочены по убыванию частот. Порядковое место  $l$ -граммы  $x_i$  в упорядочении определяет ее **ранг** –  $r(T_1, x_i)$  (соответственно,  $r(T_2, x_i)$ ).

$$\tau = \frac{S}{\frac{1}{2}n(n-1)}$$

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n g(i, j)$$

$$g(i, j) = \begin{cases} 1, & \text{если } r(T_1, x_i) < r(T_1, x_j) \text{ \& } r(T_2, x_i) < r(T_2, x_j) \\ 1, & \text{если } r(T_1, x_i) > r(T_1, x_j) \text{ \& } r(T_2, x_i) > r(T_2, x_j) \\ -1, & \text{в остальных случаях} \end{cases}$$

$x_i$	a	c	g	t
$f(T_1, x_i)$	123	101	98	37
$r(T_1, x_i)$	1	2	3	4
$f(T_2, x_i)$	147	211	988	1137
$r(T_2, x_i)$	4	3	2	1

$$\tau = \frac{-6}{\frac{1}{2}4(4-1)} = -1$$

## Ранговые меры близости. Коэффициент корреляции $\tau$

$$\tau = \frac{S}{\frac{1}{2}n(n-1)} \quad S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n g(i, j) \quad \tau = \frac{-3}{\frac{1}{2}10 \times 9} = -0,07$$

$$g(i, j) = \begin{cases} 1, & \text{если } r(T_1, x_i) < r(T_1, x_j) \ \& \ r(T_2, x_i) < r(T_2, x_j) \\ 1, & \text{если } r(T_1, x_i) > r(T_1, x_j) \ \& \ r(T_2, x_i) > r(T_2, x_j) \\ -1, & \text{в остальных случаях} \end{cases}$$

ученики	A	B	C	D	E	F	G	H	I	J
математика	7	4	3	10	6	2	9	8	1	5
музыка	5	7	3	10	1	9	6	2	8	4

ученики	I	F	C	B	J	E	A	H	G	D
математика	1	2	3	4	5	6	7	8	9	10
музыка	8	9	3	7	4	1	5	2	6	10

$$S = 2 - 7 + 1 - 7 + 5 - 2 + 1 - 5 + 3 - 2 + 4 - 0 + 2 - 1 + 2 - 0 + 1 = -3$$

## Ранговые меры близости. Коэффициент Спирмэна

Пусть  $l$ -граммы в  $\Phi_l(T_1)$  и  $\Phi_l(T_2)$  упорядочены по убыванию частот; порядковое место  $l$ -граммы  $x_i$  в упорядочении определяет ее ранг –  $r(T_1, x_i)$  (соответственно,  $r(T_2, x_i)$ ).

Группы равночастотных  $l$ -грамм представляются усредненным рангом. Введем  $l$ -граммный аналог расстояния:

$$S_l(T_1, T_2) = \sum_{x_i \in \Sigma_l} (r(T_1, x_i) - r(T_2, x_i))^2$$

где  $\Sigma_l$  – совокупность всевозможных цепочек длины  $l$ ;  $|\Sigma_l| = R_l = n^l$ . Аналогом коэффициента Спирмэна для характеристики  $l$ -го порядка ( $l = 1, 2, \dots$ ) является

$$\rho_l(T_1, T_2) = 1 - \frac{6S_l(T_1, T_2)}{n(n^2 - 1)}$$

При наличии равночастотных  $l$ -грамм в (\*\*\*) вносится поправка на "связанность" рангов. (Кендел М. Ранговые корреляции, М., Статистика, 1975)

## Ранговые меры близости. Коэффициент Спирмэна

$$\rho_l(T_1, T_2) = 1 - \frac{6S_l(T_1, T_2)}{n(n^2 - 1)}; \quad S_l(T_1, T_2) = \sum_{x_i \in \Sigma_l} (r(T_1, x_i) - r(T_2, x_i))^2$$

ученики	A	B	C	D	E	F	G	H	I	J
математика	7	4	3	10	6	2	9	8	1	5
музыка	5	7	3	10	1	9	6	2	8	4
Разности $d$	2	-3	0	0	5	-7	3	6	-7	1
$d^2$	4	9	0	0	25	49	9	36	49	1

$$\rho_l(T_1, T_2) = 1 - \frac{6 \times 182}{990} = -0,103;$$

## Ранговые меры близости. Коэффициент конкордации $W$

$W$  используется для сравнения  $m$  последовательностей.

Вычисляем суммы рангов каждого объекта  $SR(x_i) = \sum_{k=1}^m r(T_k, x_i)$

Среднее значение суммы рангов одного объекта составляет  $ES = m(n+1)/2$ .

$S$  – сумма квадратов отклонений от  $ES$   $S = \sum_{i=1}^n (SR(x_i) - ES)^2$

$$W = \frac{12S}{m^2(n^3 - n)}$$

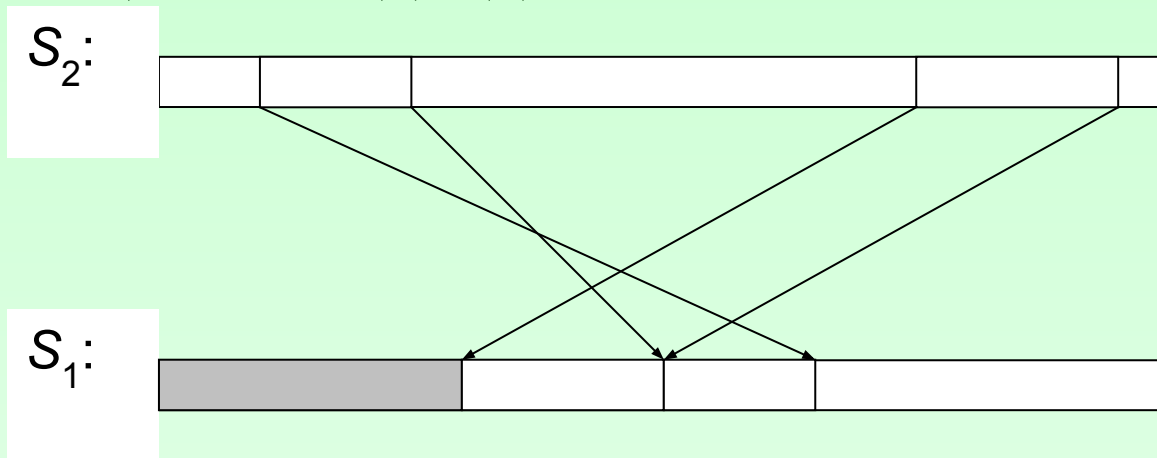
Равночастотным  $l$ -граммам назначаются усредненные ранги и в определении  $\tau$ ,  $\rho$ ,  $W$  вносится поправка на "связанность".

$t$  – число элементов в одной группе.

$$W = \frac{S}{\frac{1}{12}m^2(n^3 - n) - m \sum_{T'} T'}$$
$$T' = \frac{1}{12} \sum_t (t^3 - t)$$

$W$ , в отличие от  $\tau$  и  $\rho$  изменяется от 0 до 1.

## Обобщение подхода Лемпеля-Зива



- Представление  $S_1$  в виде конкатенации фрагментов из  $S_2$  назовем **сложностным разложением**  $S_1$  по  $S_2$ .
- На каждом шаге копируется максимальный фрагмент  $S_2$ , совпадающий с префиксом непокрытого участка  $S_1$
- Если такого фрагмента нет, используется операция генерации символа
- $c(S_1 / S_2)$  – сложность  $S_1$  относительно  $S_2$  определяется числом компонентов в разложении  $S_1$  по  $S_2$

# Относительная сложность и редакционное расстояние

$S_2$  = аааа а сссс с тттттттттттт - асасасас а атататат

$S_1$  = аааа г сссс г тттттттттттт г асасасас г атататат

$d(S_1, S_2)$  = 4

$H(S_1/S_2)$  = аааа\*г\*сссс\*г\*тттттттттттт\*г\*асасасас\*г\*атататат

$c(S_1/S_2)$  = 9

- $c(S_1/S_2) \leq 2d(S_1, S_2) + 1$

- $S_2$  = -----тттттттттттттттттттаааааааа

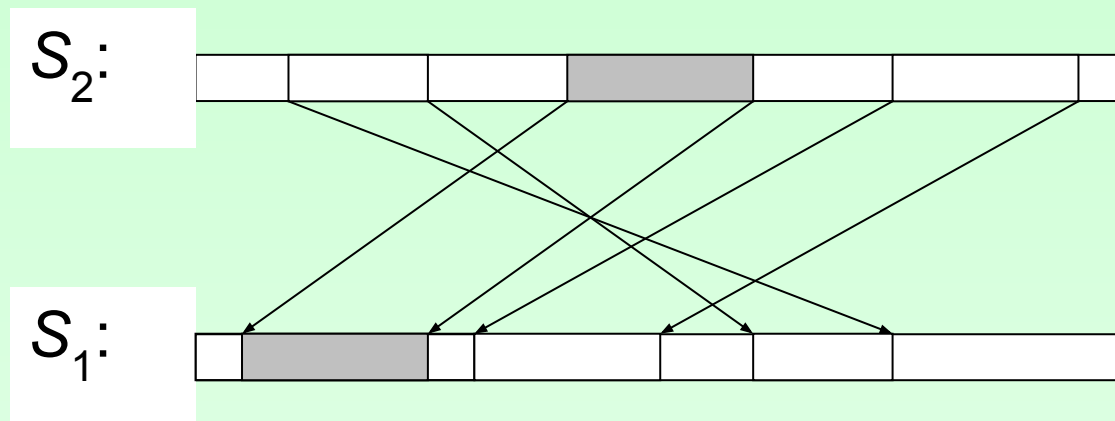
$S_1$  = ааааааааттттттттттттттттттт-----

$d(S_1, S_2) = 16$

$H(S_1 / S_2) = ааааа * ттттттттттттттттттт$

$c(S_1 / S_2) = 2$

# Трансформационное расстояние



- Трансформационное расстояние и относительная сложность идейно близки.
- Операция «вставка сегмента» используется, если посимвольная генерация фрагмента «дешевле» его копирования.
- Порядок покрытия  $S_1$  предполагает оптимизацию по всем парам межтекстовых повторов и промежуткам между ними.  $O(N^6)$ .

J.-S.Varré, J.-P.Delahaye, E. Rivals: Transformation Distances: a Family of Dissimilarity Measures Based on Movements of Segments. // Bioinformatics 15(3): 194-202 (1999)



## Инверсионное расстояние

$$\left( \begin{array}{cccccccc} 1 & 2 & \boxtimes & i-1 & i & i+1 & \boxtimes & j-1 & j & j+1 & \boxtimes & N \\ & & & & \longleftarrow & \longrightarrow & & & & & & \\ 1 & 2 & \boxtimes & i-1 & j & j-1 & \boxtimes & i+1 & i & j+1 & \boxtimes & N \end{array} \right)$$

- Инверсионное расстояние  $d_I(\pi, \sigma)$  между последовательностями  $\pi$  и  $\sigma$  определяется минимальным числом инверсий, переводящих одну из них в другую  
Задача вычисления инверсионного расстояния для перестановок является NP-полной
- В случае "знаковых" перестановок существуют полиномиальные решения

$$+1 \ [ +2 \ -4 \ -5 \ ] \ +3 \ +6 \quad \rightarrow \quad +1 \ +5 \ +4 \ -2 \ +3 \ +6$$

Hannenhalli, S. and Pevzner, P. Transforming Cabbage into Turnip (Polynomial Algorithm for Sorting Signed Permutation by Reversals). Proc. 27th Ann. ACM Symposium on the Theory of Computing, 1995, pp. 178–189

## Точки разрыва

$$\pi_0 = 0 \text{ and } \pi_{N+1} = N + 1$$

$\pi$  and  $\sigma$  – произвольные перестановки.

Разрыв между  $\pi_i = a$  и  $\pi_{i+1} = b$  фиксируется, если в  $\sigma$  нет биграмм  $ab$  и  $ba$ .

$$\pi = 0 | 6 4 | 1 8 5 | 3 | 2 9 7 | 10 \quad r(\pi, \sigma) = 5$$

$$\sigma = 0 | 5 8 1 | 2 9 7 | 6 4 | 3 | 10$$

$\sigma$  – тождественная перестановка (1 2 ... N).

Число точек разрывов (**breakpoint distance**)  $r(\pi, \sigma)$

определяется количеством позиций  $\pi$  таких что  $|\pi_i - \pi_{i+1}| \neq 1$ .

$$1 2 3 | 8 7 6 | 4 5 | 9$$

## Инверсионное расстояние, число точек разрыва и относительная сложность

- $r(\pi, \sigma) \leq 2d_I(\pi, \sigma)$
- Точки разрыва однозначно соответствуют границам компонентов сложностного разложения

$$r(\pi, \sigma) = c(\pi / \sigma) - 1$$

$$\sigma = 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9$$

$$H(\pi / \sigma) = 1\ 2\ 3\ * 8\ 7\ 6\ * 4\ 5\ * 9$$

- Сложностные разложения позволяют перейти от исходных перестановок к "знаковым" и сократить размерность задачи вычисления инверсионного расстояния