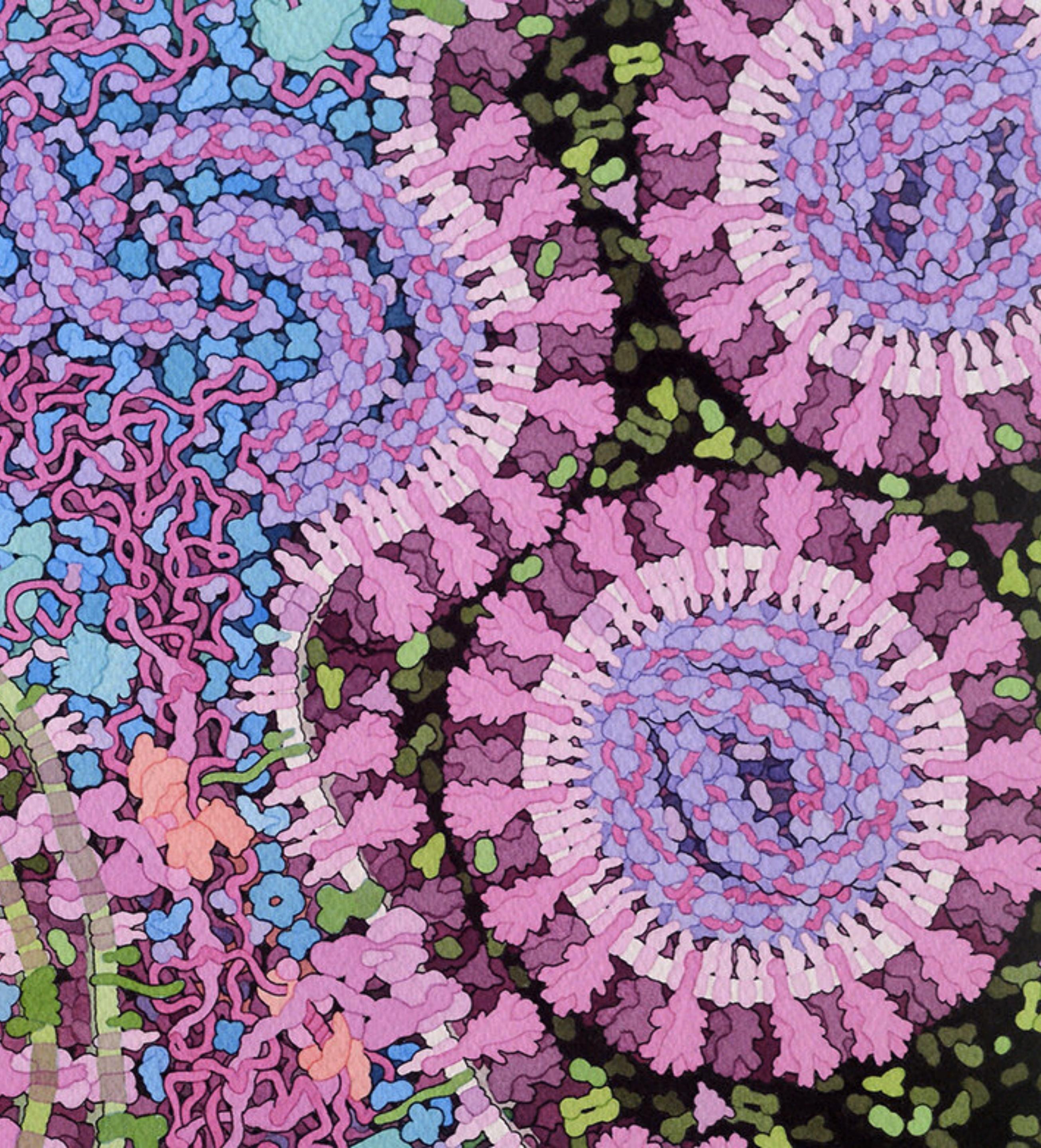


بسم الله الرحمن الرحيم

ڙنو ميڪ محاسباتي

جلسه ٨: هم ترازي چندگانه (۱)

١٤٠١-١٤٠٠
ترم پايز



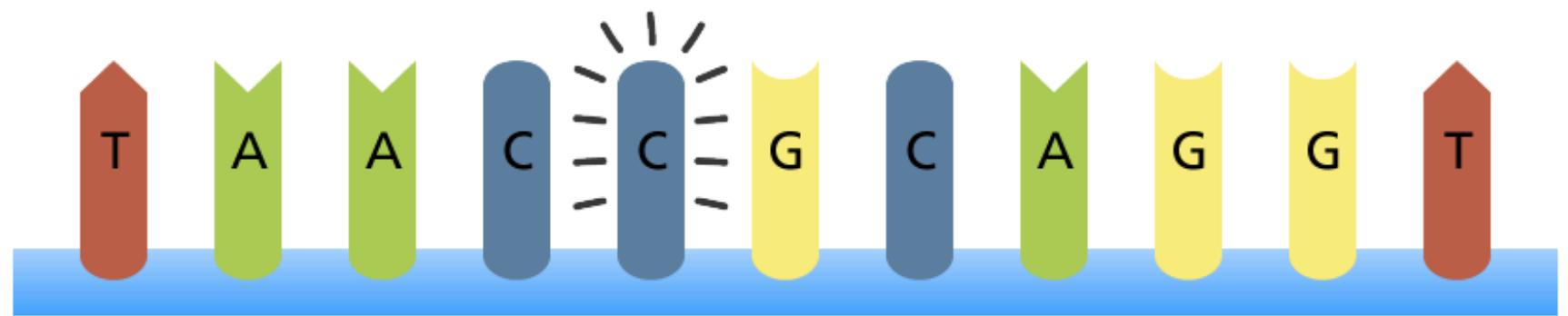
هم ترازی، چرا؟

● تغییرات زیستی،

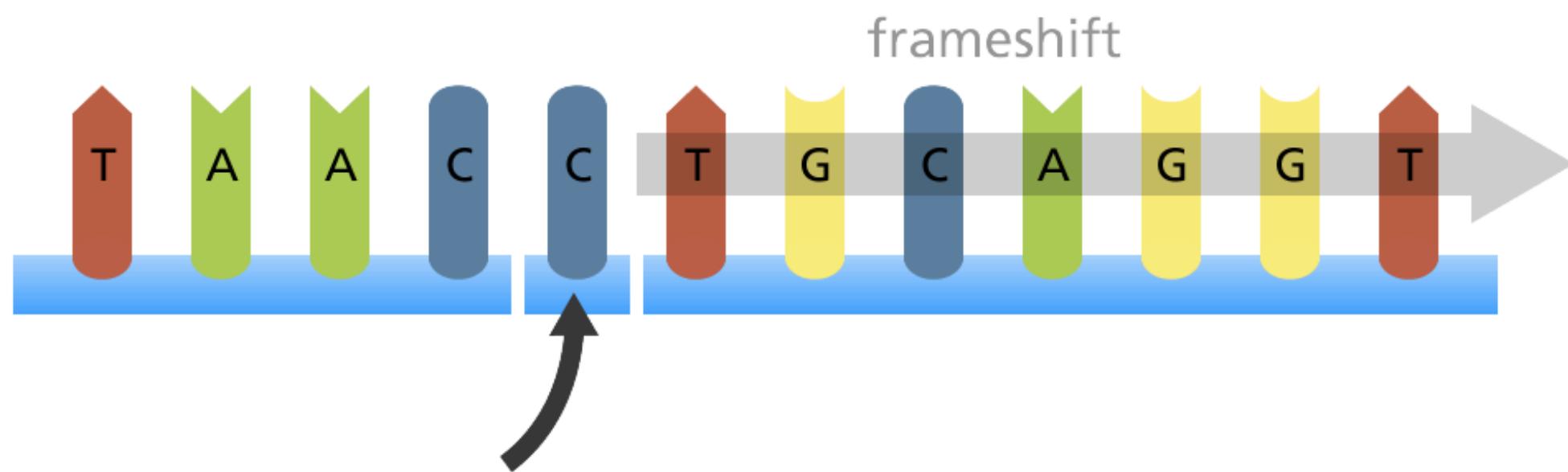
● تغییر، حذف، اضافه



Base substitution

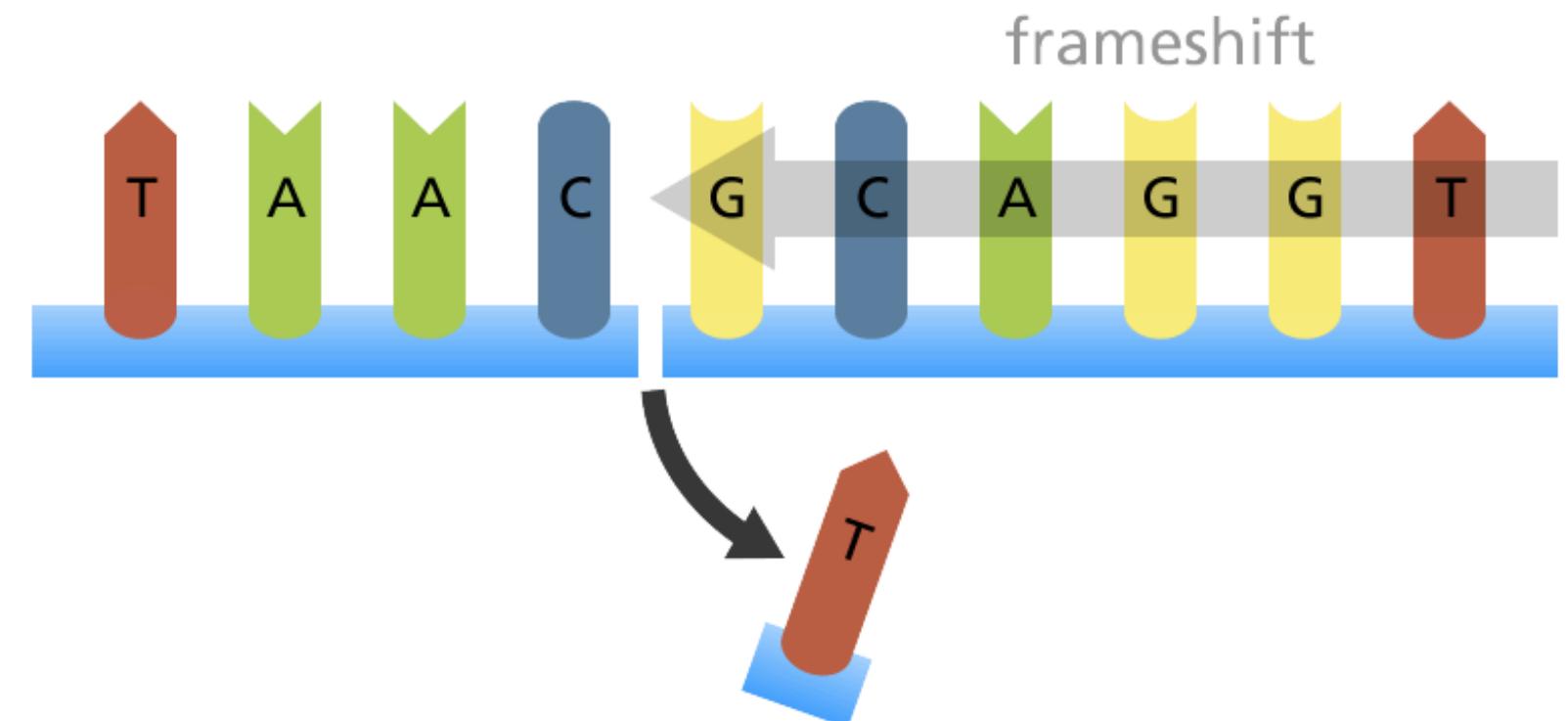


Base addition



Original sequence
تغيرات DNA

Base deletion



هم ترازی، چرا؟

- تغییرات زیستی،
- تغییر، حذف، اضافه
- کدام مهم است؟
- الف) مکان اسید آمینه در توالی،
- ب) نوع اسید آمینه؟
- چ) پس چی؟

یادآوری: مسئله هم ترازی دو رشته

چند رشته

● ورودی: دو رشته S و T و ماتریس مشابهت حروف

$$\delta : (\Sigma \cup \{-\})^2 \rightarrow \mathbb{R}$$

● خروجی:

● یک جدول

● سطر اول S که تعدادی «-» به آن اضافه شده

● سطر اول T که تعدادی «-» به آن اضافه شده

s1	-	s2	s3	s4	-	-	-	s5	-
-	T1	T2	T3	-	-	T4	T5	T6	T7

چند رشته

● دارای بیشترین امتیاز

امتیاز چند سطری؟

● امتیاز = جمع امتیاز ستون‌ها (بر اساس δ)

δ

یادآوری: مسئله همترازی دو رشته

چند رشته

	-	A	C	G	T
-	-1	-1	-1	-1	-1
A	-1	2	-1	-1	-1
C	-1	-1	2	-1	-1
G	-1	-1	-1	2	-1
T	-1	-1	-1	-1	2

مثال:

 $S = \text{ACAAT CC}$ $T = \text{AGCAT GC}$

ورودی: دو رشته S و T و ماتریس مشابهت حروف
 $\delta : (\Sigma \cup \{-\})^2 \rightarrow \mathbb{R}$

خروجی:

یک جدول

سطر اول S که تعدادی «-» به آن اضافه شدهسطر اول T که تعدادی «-» به آن اضافه شده

s1	-	s2	s3	s4	-	-	-	s5	-
-	T1	T2	T3	-	-	T4	T5	T6	T7

چند رشته

دارای بیشترین امتیاز

امتیاز چند سطری؟

امتیاز = جمع امتیاز ستون‌ها (بر اساس δ)

δ

یادآوری: مسئله همترازی دو رشته

چند رشته

	-	A	C	G	T
-	-1	-1	-1	-1	-1
A	-1	2	-1	-1	-1
C	-1	-1	2	-1	-1
G	-1	-1	-1	2	-1
T	-1	-1	-1	-1	2

مثال:

 $S = \text{ACAAT CC}$ $T = \text{AGCAT GC}$

مثال همترازی:

 $S = \text{A-CAATCC}$ $T = \text{AGCA-TGC}$ ورودی: دو رشته S و T و ماتریس مشابهت حروف

$$\delta : (\Sigma \cup \{-\})^2 \rightarrow \mathbb{R}$$

خروجی:

یک جدول

سطر اول S که تعدادی «-» به آن اضافه شدهسطر اول T که تعدادی «-» به آن اضافه شده

s1	-	s2	s3	s4	-	-	-	s5	-
-	T1	T2	T3	-	-	T4	T5	T6	T7

چند رشته

امتیاز چند رشته؟

امتیاز = جمع امتیاز ستون‌ها (بر اساس δ)

δ

یادآوری: مسئله همترازی دو رشته

چند رشته

	-	A	C	G	T
-	-1	-1	-1	-1	-1
A	-1	2	-1	-1	-1
C	-1	-1	2	-1	-1
G	-1	-1	-1	2	-1
T	-1	-1	-1	-1	2

مثال:

 $S = \text{ACAAT CC}$ $T = \text{AGCAT GC}$

مثال همترازی:

 $S = \text{A-CAATCC}$ $T = \text{AGCA-TGC}$

امتیاز: ???

ورودی: دو رشته S و T و ماتریس مشابهت حروف

$$\delta : (\Sigma \cup \{-\})^2 \rightarrow \mathbb{R}$$

خروجی:

یک جدول

سطر اول S که تعدادی «-» به آن اضافه شدهسطر اول T که تعدادی «-» به آن اضافه شده

s1	-	s2	s3	s4	-	-	-	s5	-
-	T1	T2	T3	-	-	T4	T5	T6	T7

چند رشته

دارای بیشترین امتیاز

سطری؟

امتیاز = جمع امتیاز ستون‌ها (بر اساس δ)

هم ترازی چندگانه

● ورودی: چند رشته S_i و ماتریس مشابهت حروف

$$\delta : (\Sigma \cup \{ - \})^2 \rightarrow \mathbb{R}$$

● خروجی:

● یک جدول

● سطر i : شامل S_i که تعدادی «_» به آن اضافه

شده

$S_1 = ACG_GAGA$
 $S_2 = _CGTTGACA$
 $S_3 = AC_T_GA_A$
 $S_4 = CCGTTCAC_$

S_{11}	-	S_{12}	S_{13}	S_{14}	-	-	-	S_{15}	-
-	S_{21}	S_{22}	S_{23}	-	-	S_{24}	S_{25}	S_{26}	S_{27}
S_{31}	S_{32}	-	S_{33}	-	-	S_{34}	S_{35}	-	S_{36}

● دارای بیشترین امتیاز

مثال:

● یک گزینه امتیاز = جمع امتیاز دو به دوی سطرها
(بدون فاصله)

هم ترازی چندگانه

δ

	-	A	C	G	T
-	-1	-1	-1	-1	-1
A	-1	2	-1	-1	-1
C	-1	-1	2	-1	-1
G	-1	-1	-1	2	-1
T	-1	-1	-1	-1	2

- ورودی: چند رشته S_i و ماتریس مشابهت حروف

$$\delta : (\Sigma \cup \{ - \})^2 \rightarrow \mathbb{R}$$

- خروجی:

- یک جدول

- سطر i : شامل S_i که تعدادی «_» به آن اضافه

شده

$S_1 = ACG_GAGA$
 $S_2 = _CGTTGACA$
 $S_3 = AC_T_GA_A$
 $S_4 = CCGTTCAC_$

S_{11}	-	S_{12}	S_{13}	S_{14}	-	-	-	S_{15}	-
-	S_{21}	S_{22}	S_{23}	-	-	S_{24}	S_{25}	S_{26}	S_{27}
S_{31}	S_{32}	-	S_{33}	-	-	S_{34}	S_{35}	-	S_{36}

- دارای بیشترین امتیاز

مثال:

- یک گزینه امتیاز = جمع امتیاز دو به دوی سطرها
(بدون فاصله)

مسئله هم ترازی چندگانه

- الگوریتم؟

- سخت_NP

- برنامه ریزی پویا؟!

- الگوریتم های تقریبی

- الگوریتم های مکاشفه ای

تلاش: برنامه‌ریزی پویا برای هم‌ترازی چندگانه

هم‌ترازی k رشته:

$$n_1 \cdot n_2 \cdot \dots \cdot n_k \rightarrow V(i_1, i_2, \dots, i_k) =$$

2^k

$$\max_{\substack{(b_1, \dots, b_k) \in \{0,1\}^k - \{0^k\} \\ \{V(i_1 - b_1, \dots, i_k - b_k) + \text{SP-score}(i_1 b_1, i_2 b_2, \dots, i_k b_k)\}}} V(i_1, i_2, \dots, i_k)$$

$O(k^2)$

هم‌ترازی دو رشته:

$$V(i_1, i_2) = \max \begin{cases} V(i_1 - 1, j_2 - 1) + \delta(S_1[i_1], S_2[j_2]) \\ V(i_1 - 1, j_2) + \delta(S_1[i_1], -) \\ V(i_1, j_2 - 1) + \delta(-, S_2[j_2]) \end{cases}$$

روش ستاره مرکز

- ورودی: $\{S_1, \dots, S_k\}$
- ایده:
- فاصله (D): فاصله بهترین هم ترازی بین S_i و S_j
- رشته مرکز S_C : جمع فاصله اش تا همه کمینه باشد
- همه S_i ها را با S_C هم تراز کن.
- اجتماع هم ترازی: فاصله به همه اضافه می کنیم تا هم ترازی بین S_C و S_i ها حفظ شود.

Center_Star_Method

Require: A set \mathcal{S} of sequences

Ensure: A multiple alignment of M with sum of pair distances at most twice that of the optimal alignment of \mathcal{S}

- 1: Find $D(S_i, S_j)$ for all i, j .
- 2: Find the center sequence S_c which minimizes $\sum_{i=1}^k D(S_c, S_i)$.
- 3: For every $S_i \in \mathcal{S} - \{S_c\}$, choose an optimal alignment between S_c and S_i .
- 4: Introduce spaces into S_c so that the multiple alignment \mathcal{M} satisfies the alignments found in Step 3.

S_1 : CCTGCTGCAG
 S_2 : GATGTGCCG
 S_3 : GATGTGCAG
 S_4 : CCGCTAGCAG
 S_5 : CCTGTAGG

	S_1	S_2	S_3	S_4	S_5
S_1	0	4	3	2	4
S_2		0	1	6	5
S_3			0	5	5
S_4				0	4
S_5					0

$$\begin{aligned}\sum_{i=1..k} D(S_1, S_i) &= 13 \\ \sum_{i=1..k} D(S_2, S_i) &= 16 \\ \sum_{i=1..k} D(S_3, S_i) &= 14 \\ \sum_{i=1..k} D(S_4, S_i) &= 17 \\ \sum_{i=1..k} D(S_5, S_i) &= 18\end{aligned}$$

S_1 : CCTGCTGCAG
 S_2 : GATG-TGCCG

S_1 : CCTGCTGCAG
 S_3 : GATG-TGCAG
 S_1 : CCTGCT-GCAG
 S_4 : CC-GCTAGCAG

S_1 : CCTGCT-GCAG
 S_2 : GATG-T-GCCG
 S_3 : GATG-T-GCAG
 S_4 : CC-GCTAGCAG
 S_5 : CCTG-TAG--G

S_1 : CCTGCT-GCAG
 S_5 : CCTG-TAG--G

S_1 : CCTGCTGCAG
 S_2 : GATG-TGCCG

S_1 : CCTGCTGCAG
 S_3 : GATG-TGCAG

S_1 : CCTGCTGCAG
 S_2 : GATG-TGCCG
 S_3 : GATG-TGCAG

S_1 : CCTGCT-GCAG
 S_4 : CC-GCTAGCAG

S_1 : CCTGCTG-CAG
 S_2 : GATG-TG-CCG
 S_3 : GATG-TG-CAG
 S_4 : CC-GCTAGCAG

S_1 : CCTGCT-GCAG
 S_5 : CCTG-TAG--G

S_1 : CCTGCT-GCAG
 S_2 : GATG-T-GCCG
 S_3 : GATG-T-GCAG
 S_4 : CC-GCTAGCAG
 S_5 : CCTG-TAG--G

اثبات ضریب تقریب:

همترازی ستاره مرکز M

همترازی بهینه M^*

$$\begin{aligned}
 &= \sum_{1 \leq i < j \leq k} d_{\mathcal{M}}(i, j) &= \sum_{1 \leq i < j \leq k} d_{\mathcal{M}^*}(i, j) \\
 &= \frac{1}{2} \sum_{i=1}^k \sum_{j=1}^k d_{\mathcal{M}}(i, j) &\geq \sum_{1 \leq i < j \leq k} D(S_i, S_j) \\
 &\leq \frac{1}{2} \sum_{i=1}^k \sum_{j=1}^k [D(S_c, S_i) + D(S_c, S_j)] &= \frac{1}{2} \sum_{i=1}^k \sum_{j=1}^k D(S_i, S_j) \\
 &= \frac{k}{2} \sum_{i=1}^k D(S_c, S_i) + \frac{k}{2} \sum_{j=1}^k D(S_c, S_j) &\geq \frac{1}{2} \sum_{i=1}^k \sum_{j=1}^k D(S_c, S_j) \\
 &= k \sum_j D(S_c, S_j) &= \frac{k}{2} \sum_{j=1}^k D(S_c, S_j)
 \end{aligned}$$

فاصله همтраزی ستاره مرکز \Rightarrow نصف فاصله همترازی بهینه

زمان اجرا

Center_Star_Method

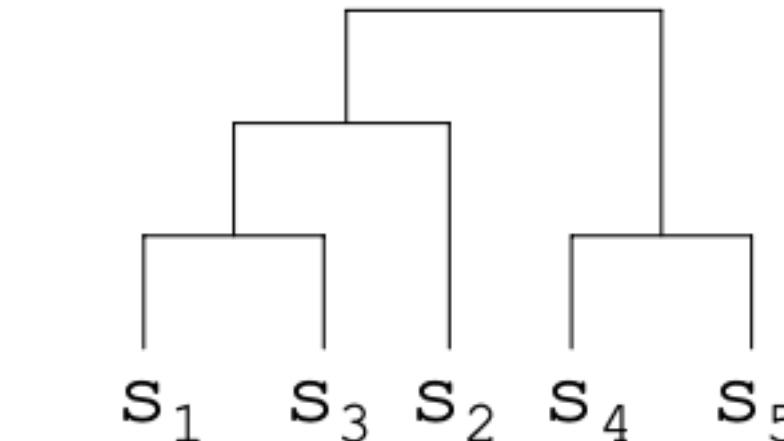
Require: A set \mathcal{S} of sequences

Ensure: A multiple alignment of M with sum of pair distances at most twice that of the optimal alignment of \mathcal{S}

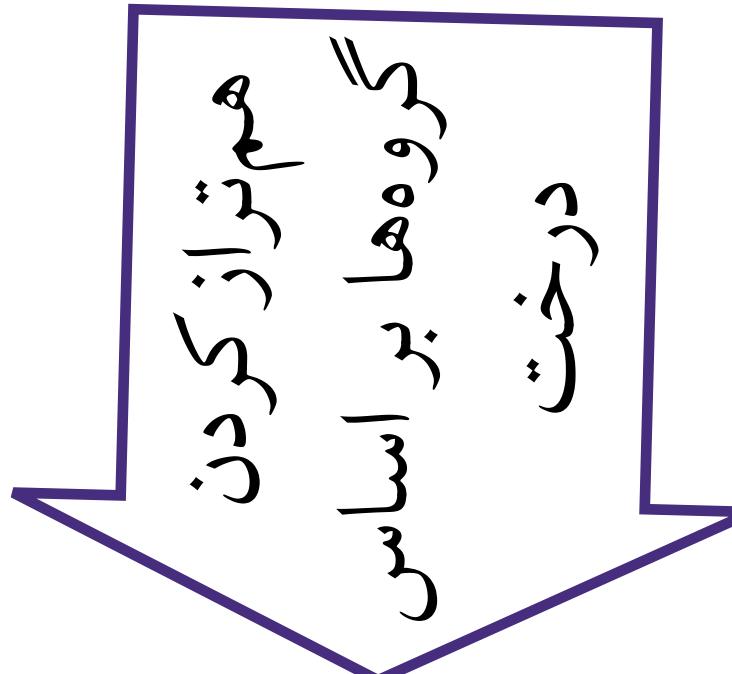
- 1: Find $D(S_i, S_j)$ for all i, j .
- 2: Find the center sequence S_c which minimizes $\sum_{i=1}^k D(S_c, S_i)$.
- 3: For every $S_i \in \mathcal{S} - \{S_c\}$, choose an optimal alignment between S_c and S_i .
- 4: Introduce spaces into S_c so that the multiple alignment \mathcal{M} satisfies the alignments found in Step 3.

$O(k^2n^2)$

روش‌های پیش‌رونده



محاسبه
درخت راهنمای



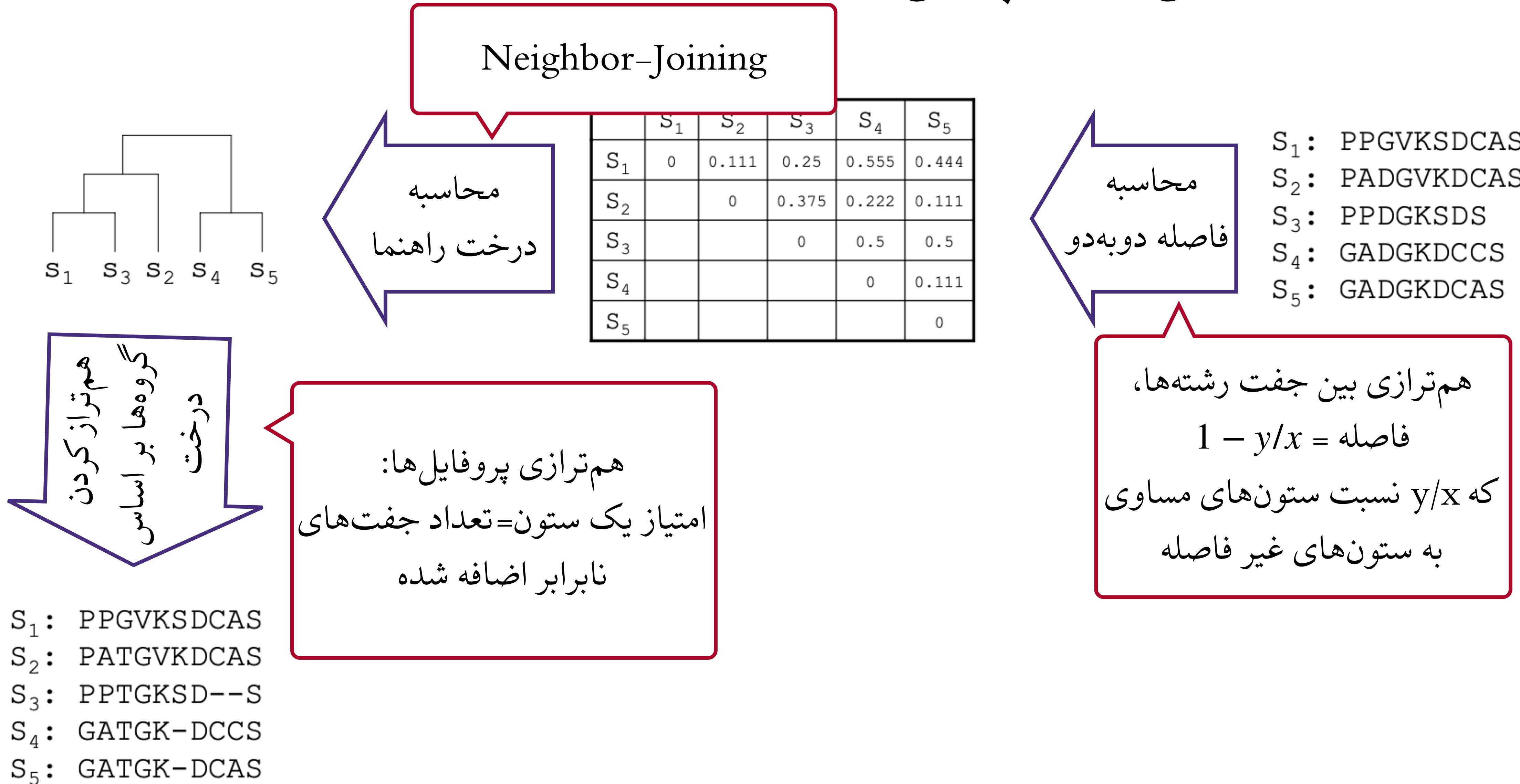
	S ₁	S ₂	S ₃	S ₄	S ₅
S ₁	0	0.111	0.25	0.555	0.444
S ₂		0	0.375	0.222	0.111
S ₃			0	0.5	0.5
S ₄				0	0.111
S ₅					0

محاسبه
فاصله دوبه دو

- S₁: PPGVKSDCAS
S₂: PATGVKDCAS
S₃: PPTGKSD--S
S₄: GATGK-DCCS
S₅: GATGK-DCAS

- S₁: PPGVKSDCAS
S₂: PADGVKDCAS
S₃: PPDGKSDS
S₄: GADGKDCCS
S₅: GADGKDCAS

روش‌های پیش‌رونده: مثال ClustalW



مثال هم ترازی پروفایل در ClustalW

S1:	PPGVKSEDCAS	S1:	PPGVKSEDCAS
S2:	PATGVKEDCAS	S2:	PATGVKEDCAS
S3:	PPDGKSED--S	S3:	PPDGKSED--S
		S4:	GATGKDCCS
		S5:	GATGKDCAS
(a)		(b)	(c)

روش‌های تکراری:

اپدھ

- ۱) تولید هم ترازی چندگانه اولیه
- ۲) تلاش برای بهبودهای کوچک
- باز درخت بسازیم
- باز هم ردیف کنیم

MUSCLE: مثال:

● ۱) پیش روی اولیه

● ۲) پیش روی ارتقاء دهنده

● ۳) اصلاح

MUSCLEمثال:

$$(1 - f_G^i)(1 - f_G^j) \log \sum_{x,y \in \mathcal{A}} f_x^i f_y^j \delta(x, y)$$

۱) پیش روی اولیه

روش های پیش رونده

برای هم ترازی پروفایل باتابع امتیاز لگاریتم امید

برای محاسبه فاصل، تعداد k -تایی ها را می شمارد

درخت راهنمای با UPGMA می سازد (سریع تر است)

۲) پیش روی ارتقاء دهنده

بازسازی درخت راهنمای

فاصله دو گونه = $\ln(1 - D - D^2/5)$ که D تعداد برابرهاست

بازسازی همردیفی چندگانه

۳) اصلاح

به ازای هر یال درخت، پروفایل دو قسمت را دوباره هم تراز می کند، اگر بهتر شد.

سؤال؟

