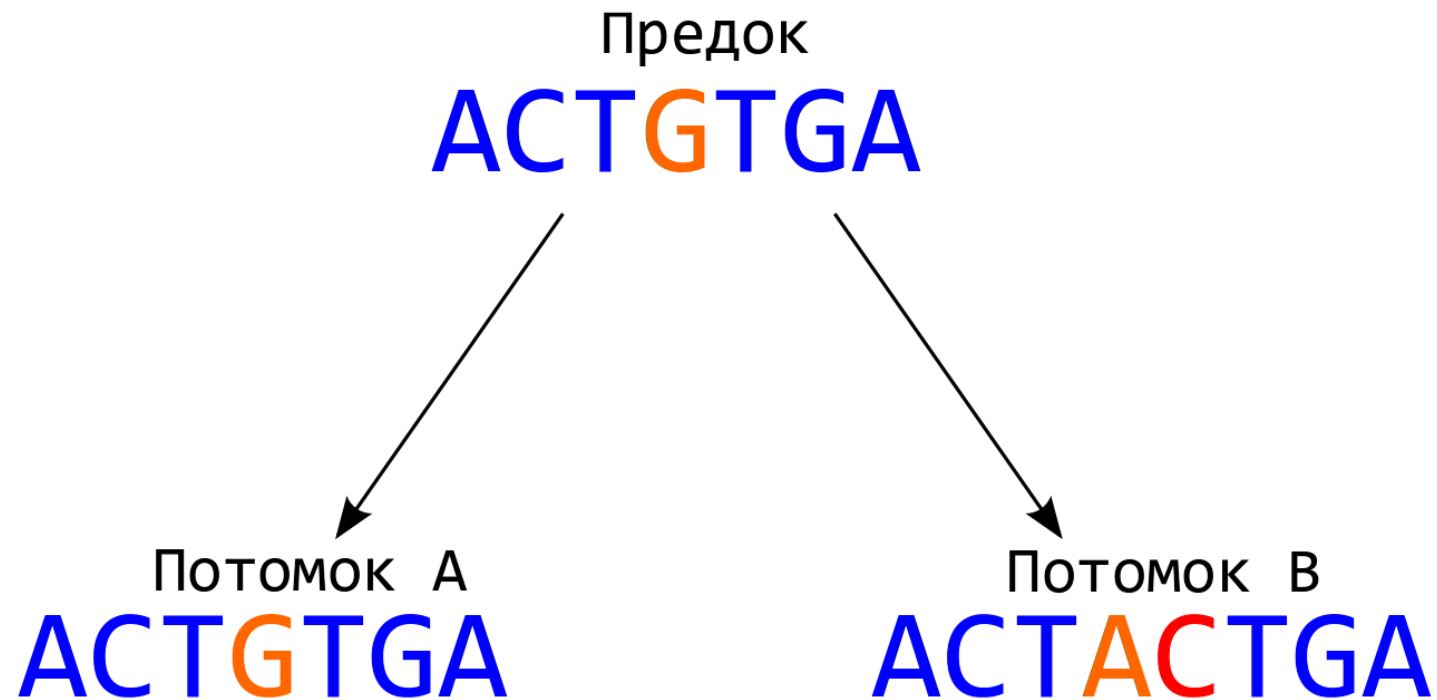


# Множественное выравнивание кодирующих последовательностей с учётом сдвигов рамки считывания

Студент: Батусов П. В.  
Руководитель: Страшнов П. В.

# Поиск гомологий в биологических последовательностях



# Задача выравнивания

ДНК предка:

...AA**A**CTGAT**G**CAAC**G**TGA...

ДНК потомка:

...AAT**C****T****T**TGATAC**C**TGA...

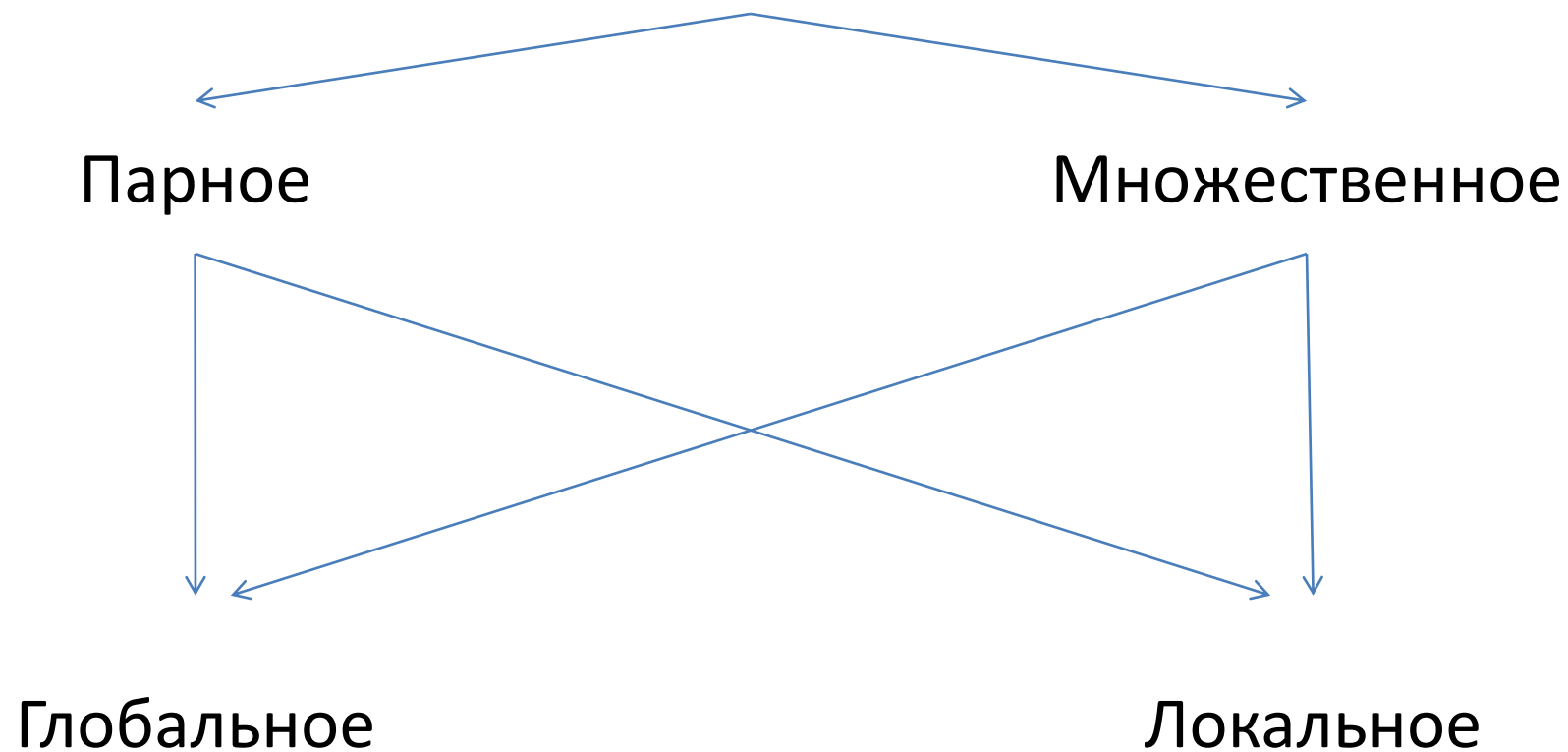


Выравнивание – общепринятый способ отражения “родства” двух последовательностей нуклеотидов или аминокислотных остатков

Seq1 :     ...AAAC--TGATGCAACGTGA...

Seq2 :     ...AATCTTTGAT---ACCTGA...

# Выравнивание



# Классические методы поиска гомологий

## Алгоритм Смита-Ватермана

- Строит локальное выравнивание двух последовательностей
- $O(\text{len}(S_1) \cdot \text{len}(S_2))$

## Алгоритм Нидлмана-Вунша

- Строит глобальное выравнивание двух последовательностей
- $O(\text{len}(S_1) \cdot \text{len}(S_2))$

# Алгоритм Нидлмана-Вунша

- $S(a,b)$  – похожесть символов  $a$  и  $b$
- Линейный штраф за разрыв  $d$

Базис:

- $F_{0,j} = d \cdot j$
- $F_{i,0} = d \cdot i$

Итерационная формула:

$$F_{i,j} = \max \begin{cases} F_{i-1,j-1} + S_{A_i,B_j} \\ F_{i-1,j} + d \\ F_{i,j-1} + d \end{cases}$$

	А	Г	Т	Ц
А	10	-1	-4	-3
Г	-1	7	-3	-5
Т	-4	-3	8	0
Ц	-3	-5	0	9

# Алгоритм Смита-Ватермана

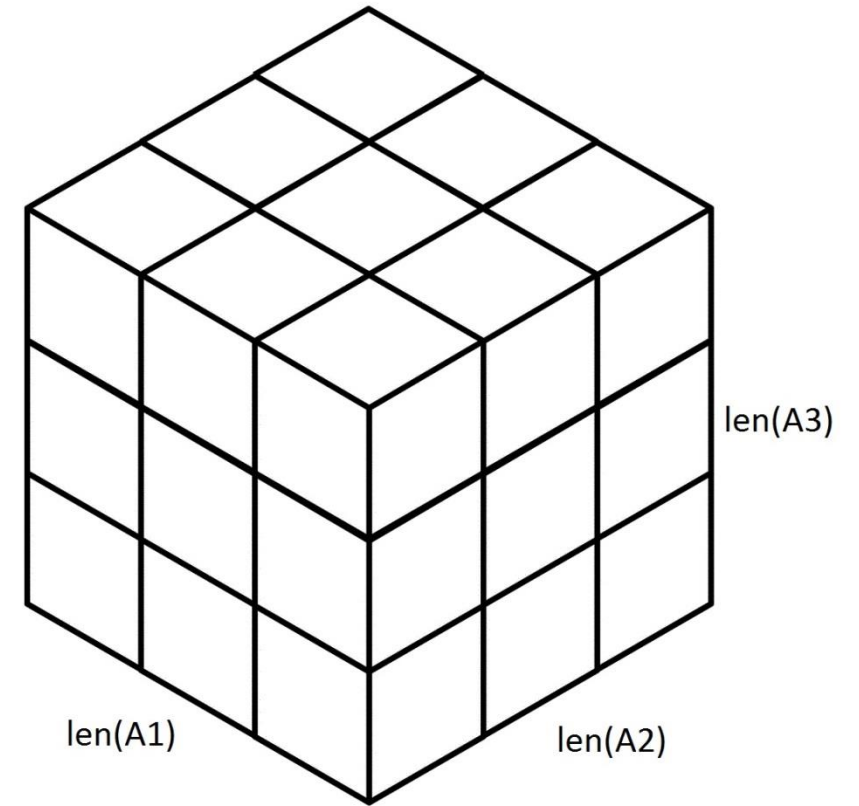
$$F_{i,j} = \max \begin{cases} F_{i-1,j-1} + S_{A_i,B_j} \\ F_{i-1,j} + D_{A_i} \\ F_{i,j-1} + I_{B_j} \\ 0 \end{cases}$$

# Множественное выравнивание

## Выравнивание в кубе

Сложность алгоритма для n-мерного случая:

$$O((2^n - 1) \prod_{i=1}^n \text{len}(A_i))$$

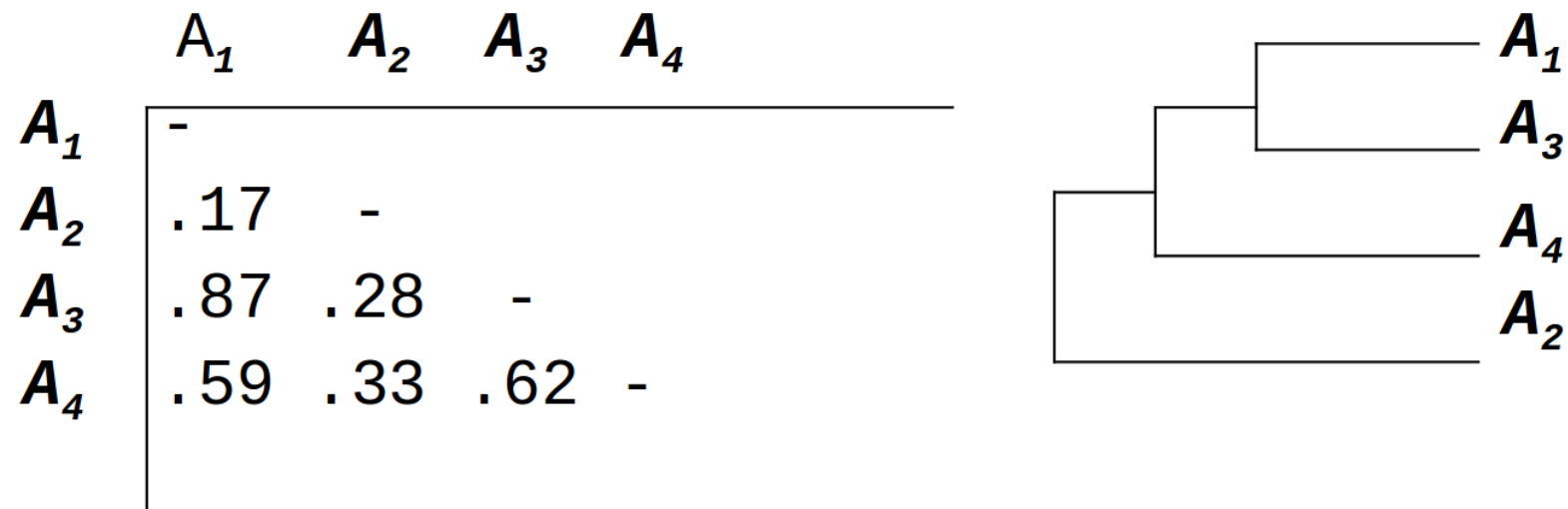




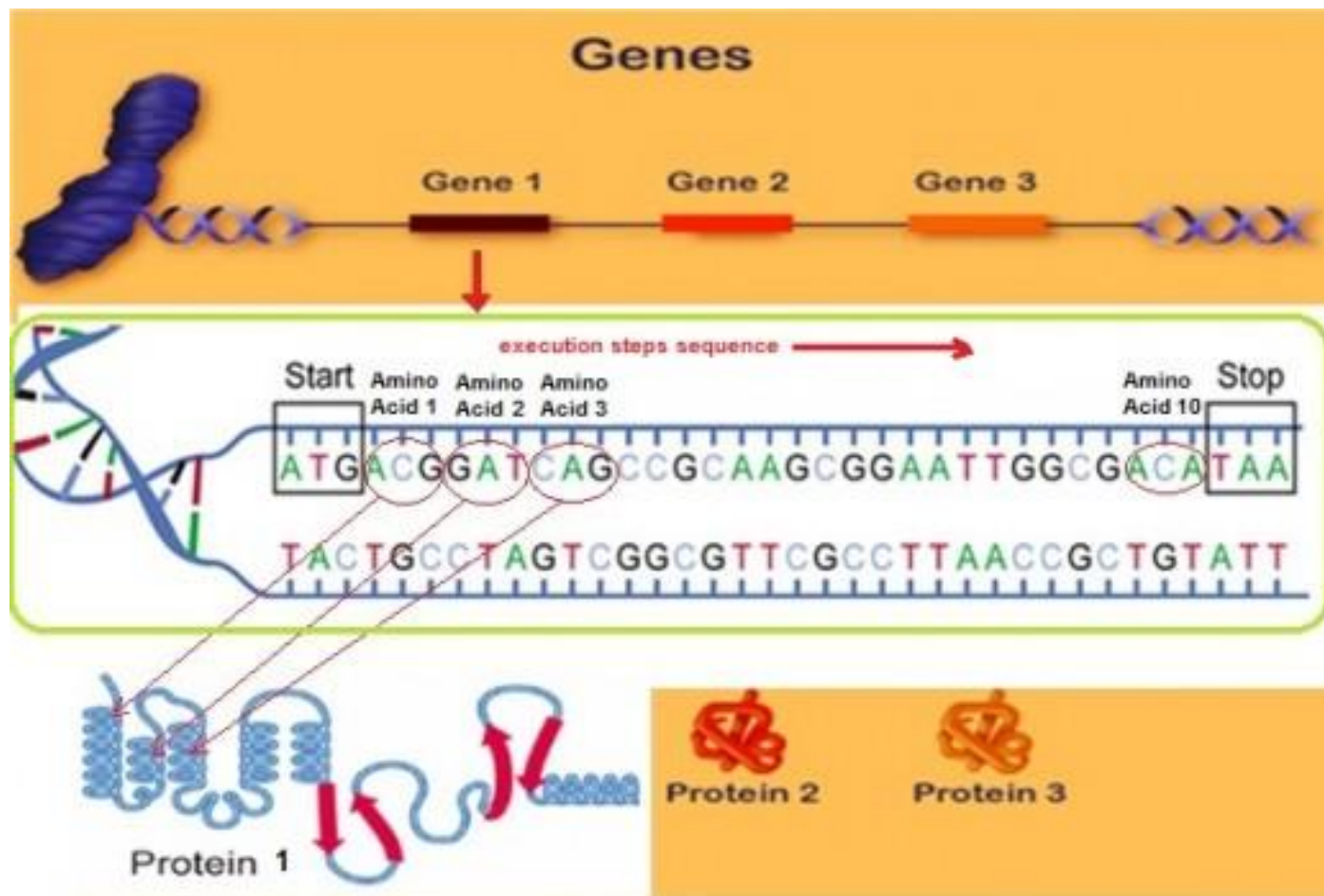
# Выравнивание выравниваний

## Алгоритм Clustal

$$f(f(f(\dots f(f(A_1, A_2), A_3) \dots), A_{n-1}), A_n)$$



# Открытые рамки считывания



# Трёхступенчатый подход

## Идея алгоритма

- Трансляция исходной последовательности нуклеотидов по всем возможным рамкам считывания
- Выравнивание последовательности аминокислот «классическими» алгоритмами
- Трансляция полученного белка обратно в последовательность нуклеотидов

## Проблемы

- Алгоритм не учитывает возможные изменения рамки считывания
- Невозможность расширения до задачи множественного выравнивания

# Двухуровневое выравнивание

## Идея алгоритма

- штраф за выравнивание является сочетанием двух штрафов: на аминокислотном и нуклеотидном уровнях
- инсерции допустимы только на аминокислотном уровне (запрет на сдвиг рамки считывания)

## Проблемы

- Высокая вычислительная сложность

# MACSE

Алгоритм основан на идее двухуровневого выравнивания, но имеет меньшую вычислительную сложность и позволяет строить множественные выравнивания, с учетом открытых рамок считывания.

MACSE производит выравнивание выравниваний, выбирая порядок через дерево-подсказку, как и алгоритм Clustal.

