



Алгоритмы выравнивания биологических последовательностей

Дарья Владимировна Диброва
(к.б.н., с.н.с)

Факультет биоинженерии и биоинформатики
МГУ имени М.В. Ломоносова

2022 год

План лекции

1. Почему последовательности генов/белков из разных организмов похожи, но не одинаковы?
2. Что такое «парное выравнивание»? Оценка качества.
3. Матрицы аминокислотных замен BLOSUM
4. Алгоритмы парного выравнивания белков
5. Связанные методы и инструменты

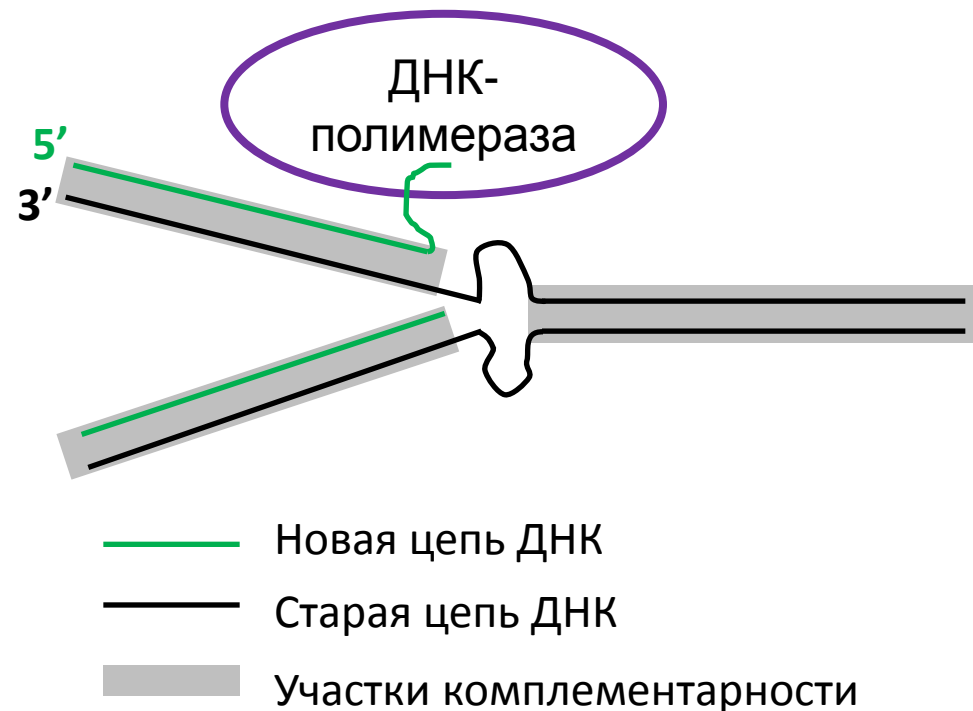
Часть 1

**Почему последовательности
генов/белков из разных организмов
похожи, но не одинаковы?**

Как может изменяться ДНК?

Генетическая информация передается от поколения к поколению за счет процесса **репликации ДНК**, после которого клетка делится

Однако ни одна ДНК-полимераза (и даже система последующей репарации) **не может** обеспечить 100% точности при копировании



Наиболее частые «рутинные» ошибки ДНК-полимеразы:

- вставка **неправильного** (нарушающего комплементарность) основания;
- **пропуск** одного или нескольких оснований;
- ошибочное **добавление** одного или нескольких лишних оснований.

Точечные мутации: замена, вставка, делеция

Met Leu His
ATGCTT**T**CAT...

ATGCT**C**CAT...
Met Leu His

1

Met Leu **His**
ATGCTT**T**CAT...

ATGCTT**CAG**...
Met Leu **Glu**

2

Met **Thr Ser**
ATG**A**CTT**C**CAT...

ATG-CTT**C**CAT...
Met **Leu His**

3

Последовательность кодируемого геном белка за счет свойств генетического кода и системы трансляции может:

1. не измениться вообще (синонимичная замена);
2. измениться на один аминокислотный остаток;
3. измениться чрезвычайно за счет сдвига рамки (англ. *frameshift*). При сдвиге рамки обычно очень скоро встречается стоп-кодон, и белок выходит дефектным.

Три фактора эволюции

Наследственность

ДНК (и белки)
за счет
процессов
репликации и
репарации
остаётся
практически
неизменным

Изменчивость

За счет
мутаций ДНК
(и белки) могут
изменяться

Отбор

Мутации в ДНК,
приводящие к
изменению белков,
зачастую **вредны**
или даже **летальны**

Некоторые
мутации, однако,
могут оказаться
нейтральны или
даже **полезны**

Наблюдаемое разнообразие последовательностей,
произошедших когда-то от одной предковой
последовательности: **гомологов**

Часть 2

**Что такое парное выравнивание и как
можно оценить его качество?**

«Парное выравнивание»

Это *любое* сопоставление двух последовательностей символов одинаковой длины, допускающее вставку в обе последовательности специального знака пропуска (*гэпа*, от англ. gap), обозначаемого обычно символом «—»

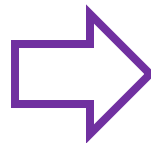
вчера вече-ром мама помыла раму
* * * *
м-ама мыла раму-----

Очевидно, что сопоставление может быть
«удачным» и «неудачным»

вчера вечером мама помыла раму
* * * * * * * * * * * * * * * *
-----мама --мыла раму

Зачем измерять «удачность» выравнивания?

Человек может оценить «удачность» на глаз;
программе это не под силу



требуется численная
мера, оценивающая
качество выравнивания

Используются *аддитивные* и *симметричные* меры:

- 1) значение меры для всего выравнивания получается сложением мер для его элементов;
- 2) значение меры не должно зависеть от того, какая последовательность считается первой, а какая второй

M – значение меры

L – длина

выравнивания

$$M = \sum_{i=1}^L m(A_{i_1}, A_{i_2})$$

Функция,
зависящая от
сопоставления
i-той пары
остатков

Как измерять «удачность» выравнивания?

1 По % совпадения (англ. *identity*);

Строгая мера, активно используемая для оценки качества уже построенного выравнивания

2 По % сходства (англ. *similarity*);

Учитывается «**химия**»: аминокислоты группируются по природе бокового радикала (например, «положительно заряженные», «гидрофобные алифатические» и т.п.)
Можно считать, что сопоставление разных аминокислот из одной и той же группы равносильно строгому совпадению

3 Используя специфические веса за замену каждой пары аминокислотных остатков

Учитывается «**биология**»: как часто определенные замены закрепляются в белках в ходе отбора

$$M = 100\% \cdot \frac{1}{L} \sum_{i=1}^L m(A_{i1}, A_{i2})$$
$$m(A_{i1}, A_{i2}) = \begin{cases} 1, & A_{i1} = A_{i2} \\ 0, & A_{i1} \neq A_{i2} \end{cases}$$

Матрицы аминокислотных замен (на примере серии BLOSUM): получение и применение

Матрица весов аминокислотных замен BLOSUM

Треугольная (симметричная) матрица

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

(Henikoff&Henikoff, 1992, PNAS)

Ключевая работа:

Amino acid substitution matrices from protein blocks

(Steven Henikoff & Jorja Henikoff, *PNAS*, 1992)



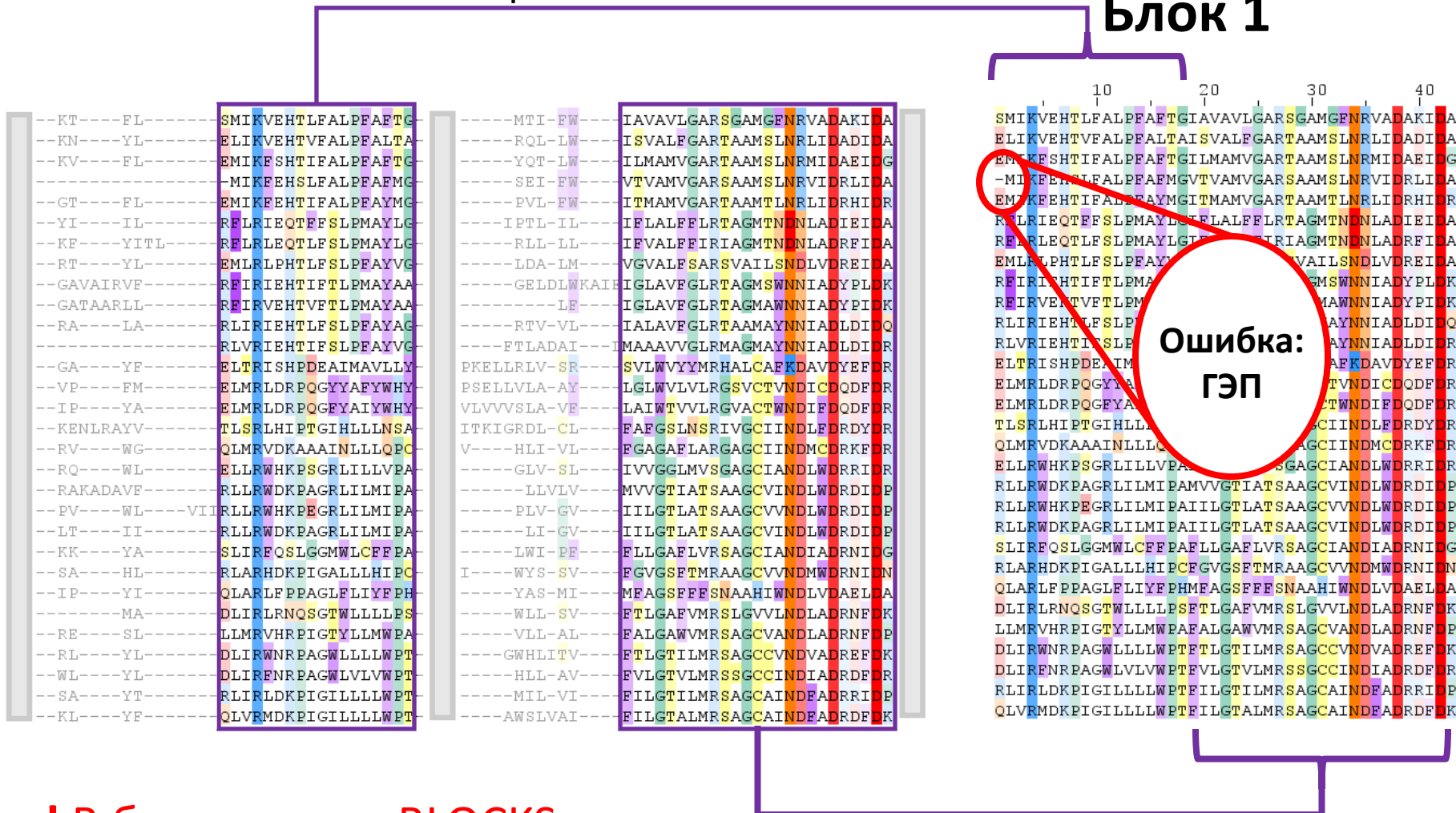
- Матрицы построены с использованием составленной авторами ранее базы данных **BLOCKS**
- На момент публикации в этой базе содержались сопоставления белков из **нескольких сотен** семейств



Каковы исходные данные, или что содержится в базе данных **BLOCKS**?

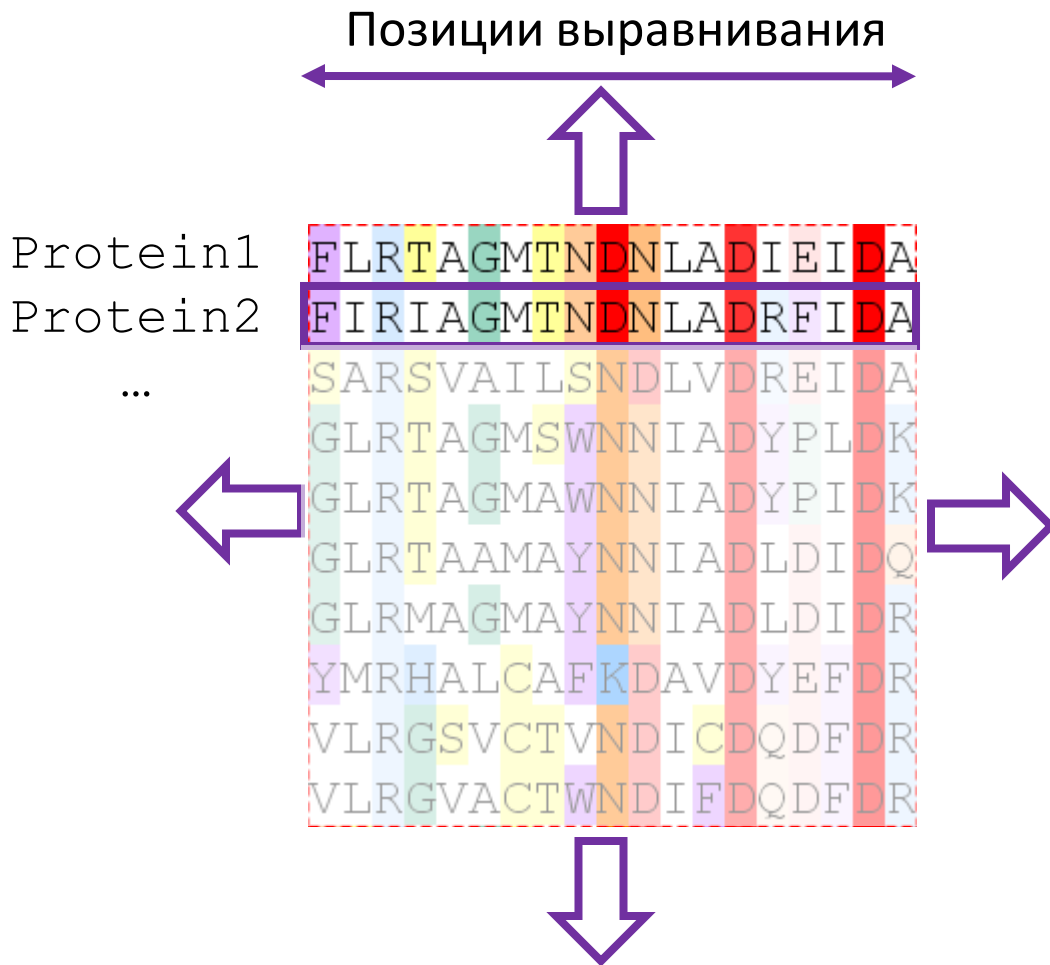
Что такое блоки (BLOCKS) ?

Множественное выравнивание



! В базе данных BLOCKS нет колонок, содержащих гэпы

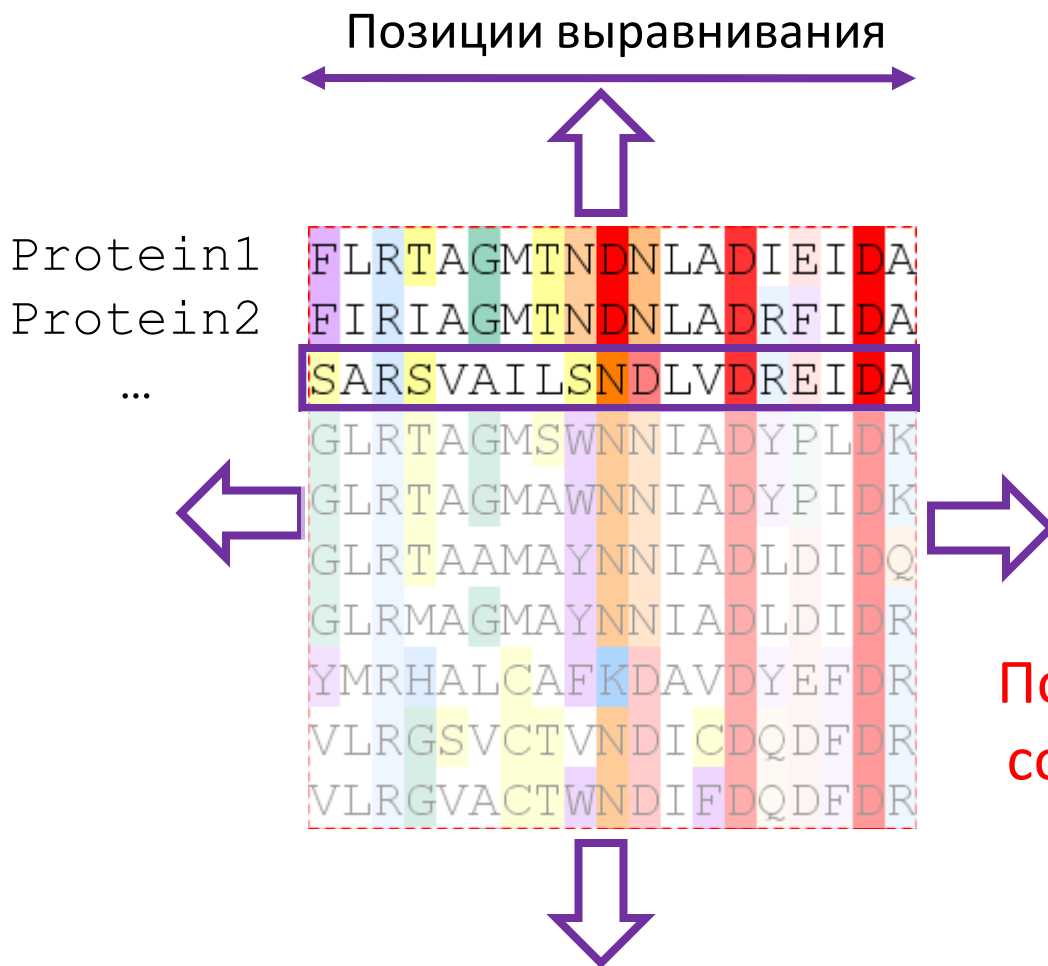
Таблица частот всех пар аминокислот



- Берем первые **две** последовательности
- Учитываем каждое сопоставление аминокислот между второй и первой

F → F = 1	T → T = 1
I → L = 1	N → N = 2
R → R = 1	D → D = 3
I → T = 1	L → L = 1
A → A = 3	R → I = 1
G → G = 1	F → E = 1
M → M = 1	I → I = 1

Таблица частот всех пар аминокислот



- Берем **третью** последовательность
- Учитываем каждое сопоставление аминокислот между ней и верхними



Получаем общее количество сопоставлений для всех пар аминокислот
(frequency table)

$$f_{ij}$$

Вычисление матрицы BLOSUM

$$q_{ij} = \frac{f_{ij}}{\sum_{i=1}^{20} \sum_{j=1}^{20} f_{ij}}$$

Наблюдаемая вероятность встречаемости каждой пары аминокислот (*observed probability of occurrence for each i,j pair*)

$$p_i = q_{ii} + \sum_{j \neq i} \frac{q_{ij}}{2}$$

Наблюдаемая частота встречаемости каждой аминокислоты (*observed probability of occurrence for amino acid i*)

$$p_i = \frac{1}{2} q_{ii} + \frac{1}{2} \sum_{j=1}^{20} q_{ij}$$

$$e_{ij} = \begin{cases} p_i p_j, & i = j \\ 2 p_i p_j, & i \neq j \end{cases}$$

Ожидаемая вероятность встречаемости каждой пары аминокислот (*expected probability of occurrence for each i,j pair*)

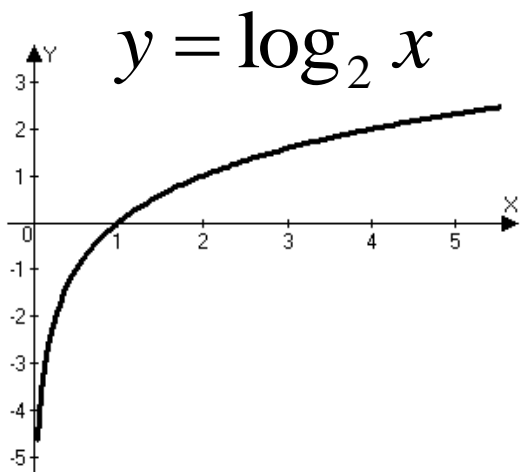
Вычисление матрицы BLOSUM

$$s_{ij} = 2 \cdot \log_2 \frac{q_{ij}}{e_{ij}}$$

Наблюдаемая вероятность для пары i, j

Ожидаемая вероятность для пары i, j

Значение в
матрице замен



$$\frac{q_{ij}}{e_{ij}} = 1$$

Частота встречаемости пары не отличается от ожидаемой случайно

$$\frac{q_{ij}}{e_{ij}} > 1$$

Пара встречается чаще, чем ожидалось бы случайно – «хорошая» замена

$$\frac{q_{ij}}{e_{ij}} < 1$$

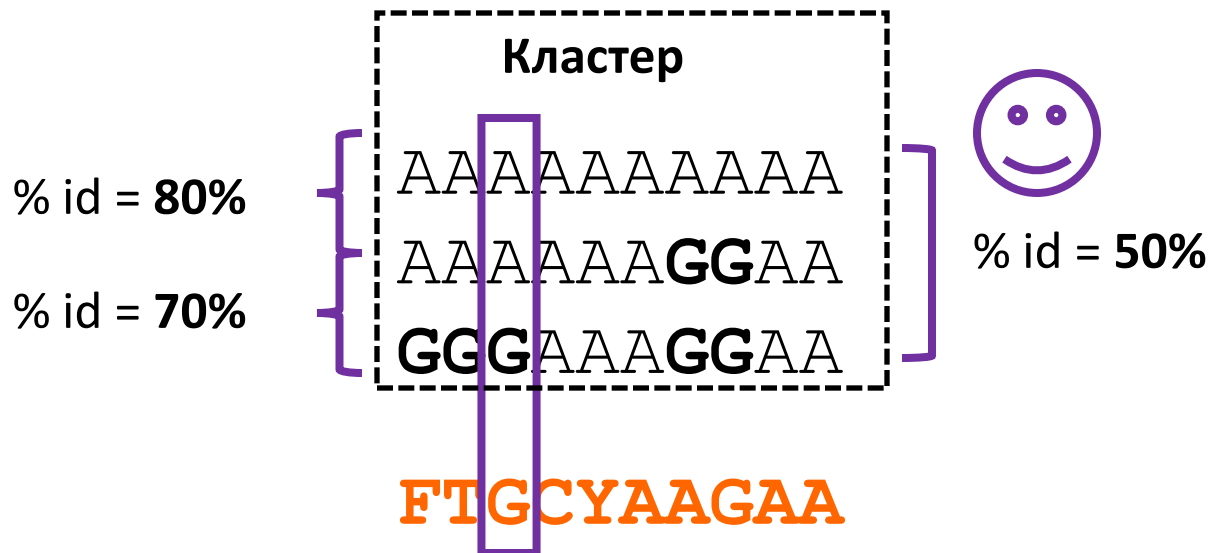
Пара встречается реже, чем ожидалось бы случайно – «плохая» замена

Что означает число 62 в «BLOSUM62»?

В блоках могут встречаться **практически идентичные последовательности**, за счет которых частоты пар могут быть ошибочно «сдвинуты»

Чтобы этого избежать, последовательности на заданном % идентичности **кластеризуют**, а порог указывают в названии матрицы

Пример: порог 62%



БЕЗ кластеризации

$$G \rightarrow G = \underline{1}$$

$$G \rightarrow A = 1 + 1 = \underline{2}$$

С кластеризацией

$$G \rightarrow G = 1 \cdot 0.33 = \underline{0.33}$$

$$G \rightarrow A = 1 \cdot 0.33 + 1 \cdot 0.33 = \underline{0.66}$$

Применение матриц серии BLOSUM

Компьютер может получить биологически осмысленную оценку качества построенного выравнивания

M – значение меры

L – длина выравнивания

$$M = \sum_{i=1}^L m(A_{i1}, A_{i2})$$

Функция, зависящая от сопоставления i -той пары остатков

Например, $m(A_{i2}, A_{i2})$ = соответствующий вес из матрицы BLOSUM.

Алгоритм парного выравнивания белковых последовательностей

Сколько существует выравниваний?

Если m и n – длины последовательностей, то число возможных выравниваний будет равно:

$$\sum_{k=0}^{\min\{m,n\}} C_n^k C_m^k = C_{n+m}^n = \frac{(n+m)!}{n!m!}$$

Для двух последовательностей длины n :

$$\frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{\pi n}} \approx 10^{77} \leftarrow n = 100$$

Число частиц во Вселенной $\sim 10^{80}$

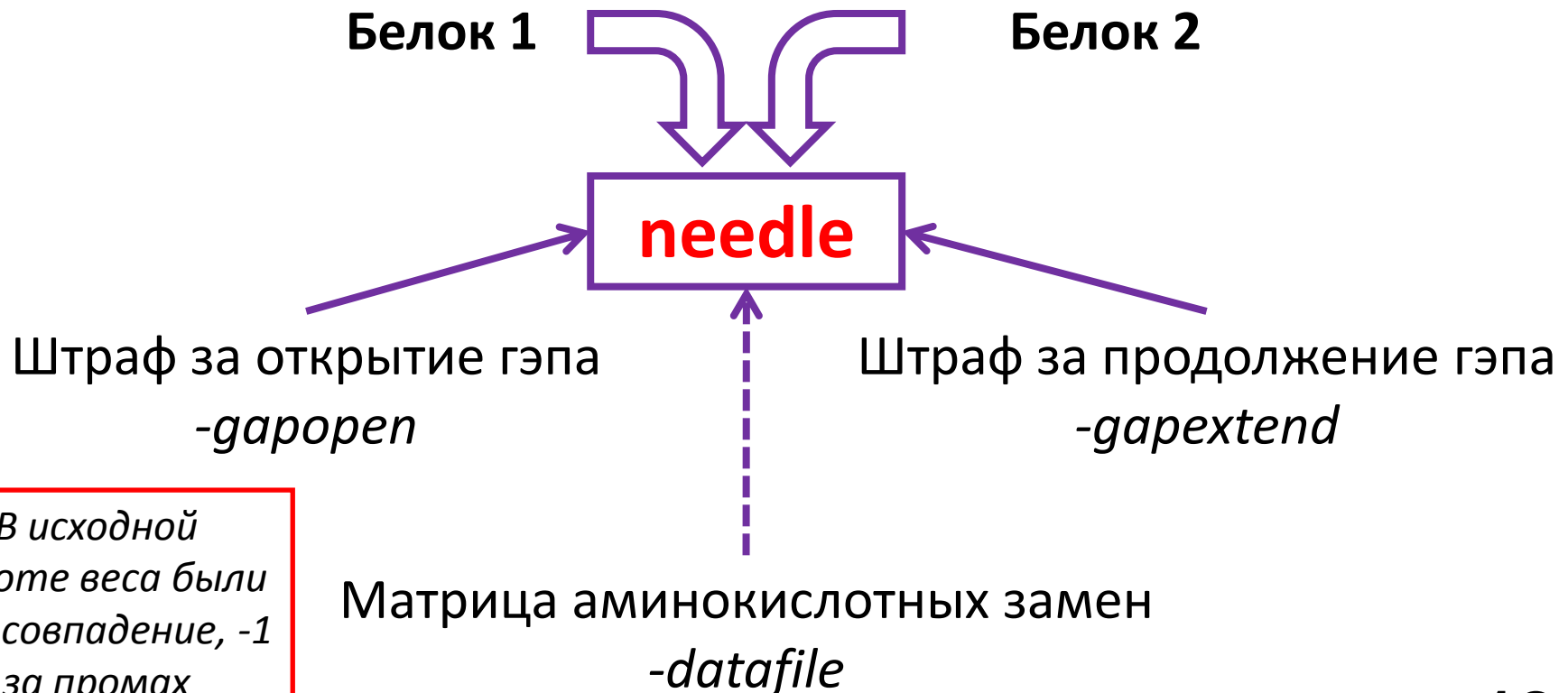
(<http://ru.wikipedia.org/wiki/Гугол>)

Глобальное парное выравнивание

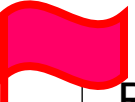
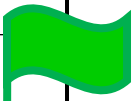
A general method applicable to the search for similarities in the amino acid sequence of two proteins

(Saul Needleman & Christian Wunsch, *J Mol Biol*, 1970)

https://www.ebi.ac.uk/Tools/psa/emboss_needle/





Динамическое программирование

	Старт			Белок 1				
			R	L	K	M	T	
		0						
A								
C								
Белок 2	L							
	K							
	M							
N								

Всегда есть два варианта вставки гэта:

- 1) вставить гэта в белок 1
- штраф за гэта
- 2) вставить гэта в белок 2
- штраф за гэта

Белок 1 RLKMT  **R**LKMT

Белок 2 ACLKMN  -ACLKMN

 -RLKMT
ACLKMN

Любое выравнивание
однозначно задается путем в
данной таблице

Динамическое программирование: гэпы

		Белок 1					
		R	L	K	M	T	
Белок 2		0	-2	-4	-6	-8	-10
	A	-2					
	C	-4					
	L	-6					
	K	-8					
	M	-10					
	N	-12					

Пусть штраф за гэп
будет = -2

Верхняя строка и левый
столбец заполняются
штрафами за гэпы

«**Концевые гэпы**»: можно
дополнительно оштрафовать за
то, что «финиш» не в конце
матрицы (т.е. нужно вставить
концевые гэпы в одну из
последовательностей)

Параметры **needle**

-endweight

-endopen

-endextend

Белок 1 RLKMT
Белок 2 -----ACLKMN
 RLKMT-----
 -----ACLKMN

Правило заполнения таблицы

		Белок 1					
		$i \rightarrow$	R	L	K	M	T
Белок 2	$j \downarrow$	0	-2	-4	-6	-8	-10
	A	-2					
	C	-4					
	L	-6					
	K	-8					
	M	-10					
	N	-12					

Штраф за гэп $g = 2$

- Выбираем, из какой клетки перейти в данную, чтобы получился наибольший вес
- Запоминаем, из какой клетки перешли
- Величину m можно вычислить по матрице BLOSUM или просто как 1 за совпадение, -1 за несовпадение

$$F_{i,j} = \max \begin{cases} F_{i,j-1} - g & \downarrow \\ F_{i-1,j} - g & \rightarrow \\ F_{i-1,j-1} + m(A_{i1}, A_{j2}) & \swarrow \end{cases}$$

Пример заполнения таблицы

		Белок 1					
		R	L	K	M	T	
Белок 2		0	-2	-4	-6	-8	-10
	A	-2	-1	-3	-5	-7	-9
	C	-4	-3	-2	-4	-6	-8
	L	-6	-5	-2	-3	-5	-7
	K	-8	-7	-4	-1	-3	-5
	M	-10	-9	-6	-3	0	-4
	N	-12	-11	-8	-5	-2	-1

Штраф за гэп $g = 2$

Цена совпадения = 1

Цена несовпадения = -1

! После заполнения матрицы мы находим наилучшее число в нижней строке или правом столбце, и **вспоминаем** путь к нему

Белок 1 R-LKMT →

Белок 2 ACLKMN

Белок 1 -RLKMT →

Белок 2 ACLKMN

! Может быть несколько альтернативных путей

Локальное парное выравнивание

Identification of Common Molecular Subsequences

(Temple Smith & Michael Waterman, *J Mol Biol*, 1981)

https://www.ebi.ac.uk/Tools/psa/emboss_water/



Белок 1

Белок 2



water

Штраф за открытие гэпа
-gapopen

Штраф за продолжение гэпа
-gapextend

Матрица аминокислотных замен
-datafile

В исходной
работе веса были
1 за совпадение, -1
за промах

Отличия от глобального выравнивания

		Белок 1				
		R	L	K	M	T
Белок 2	0	0	0	0	0	0
	A	0	-1	-1	-1	-1
	C	0	-1			
	L	0	-1			
	K	0	-1			
	M	0	-1			
	N	0	-1			

Штраф за гэп $g = 2$

Цена совпадения = 1

Цена несовпадения = -1

- 1) Нули для начальных гэпов

Отличия от глобального выравнивания

		Белок 1				
		R	L	K	M	T
Белок 2	0	0	0	0	0	0
	A	0	0	0	0	0
	C	0	0	0	0	0
	L	0	0	1	0	0
	K	0	0	0	2	0
	M	0	0	0	0	3
	N	0	0	0	0	0

Белок 1 LKM

Белок 2 LKM

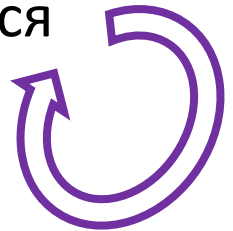


Штраф за гэп $g = 2$

Цена совпадения = 1

Цена несовпадения = -1

- 1) Нули для начальных гэпов
- 2) Все отрицательные числа в матрице приравниваются нулю
- 3) Находится самое большое число, от него восстанавливается обратный путь



Outlook: связанные и смежные методы и инструменты

BLAST: ПОИСК СХОДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

BLAST+ 2.13.0 is here!

Starting with this release, we are including the blastn_vdb and tblastn_vdb executables in the BLAST+ distribution.

Thu, 17 Mar 2022 12:00:00 EST

[More BLAST news...](#)

Web BLAST



Nucleotide BLAST
nucleotide ► nucleotide



blastx
translated nucleotide ► protein



tblastn
protein ► translated nucleotide



Protein BLAST
protein ► protein

Множественное выравнивание

менязовут-Даша-
менязовут-Женя-
менязовут-Вася-
тебязовут--Аня-
я---зовусьПетя-
меняз-валиЛизой

**Множественное
выравнивание:
одновременное
сопоставление
нескольких
последовательностей
СИМВОЛОВ**

親愛利子謂之慈，反慈為嚚；
子愛利親謂之孝，反孝為孽。
愛利出中謂之忠，反忠為倍。
心省恤人謂之惠；反惠為困。
兄敬愛弟謂之友，反友為虐。
弟敬愛兄謂之悌，反悌為敖。
接遇慎容謂之恭，反恭為媠。
接遇肅正謂之敬，反敬為嫚。
言行抱一謂之貞，反貞為偽。
期果言當謂之信，反信為慢。
衷理不辟謂之端，反端為跂。
據當不傾謂之平，反平為險。
行善決衷謂之清，反清為濁。
辭利刻謙謂之廉，反廉為貪。
兼覆無私謂之公，反公為私。
方直不曲謂之正，反正為邪。
以人自觀謂之度，反度為妄。
以己量人謂之恕，反恕為荒。
惻隱憐人謂之慈，反慈為忍。
厚志隱行謂之潔，反潔為汰。
施行得理謂之德，反德為怨。
放理潔靜謂之行，反行為污。

```
DVDNLYVVHVGVYK-GIT--LKRVPRAFQ
DVERLKI IHVA AHK-GIT--LKRWM PRAWG
DKEKLVIIHIAAHK-GIT--LKRYPMPRAFQ
DTERLRIVHIAVHK-GRV--LKRWM PRAFQ
DVDKLIIRHIAAHK-GIT--LQRYMPRAFQ
NVENLRIVHAAAHK-GMK--IRNYLPRAFQ
DVDRVVIVHAAAHK-GFK--IPNIMPRAFQ
NVDNVV I IHAAIHK-GRK--IKNYMPRAFQ
DVENLVIVHAQAQK-GRV--IERYPMPRAFQ
DADEMVIHVAPHKVGES---QGRKPRAMG
EGEEMVIHVAAHK-VGE--SPGQKPRAFQ
DVENLKLHVHCANR-GVI--IRGWT PRAFQ
NTEKLRIKHISTNK-GIT--IKRYPMPRAFQ
NTEKLRIKHISSNK-GFT--IKRHM PRAFQ
DTENMRIIHAASKK-GHV--TRGMMPRAFQ
DTRLYIKHIAAHK-GRV--IRGWI PRAFQ
DPDKLKI IHIAAHK-GPV--LRGWY PRAFQ
DPDRLRIIHVAHR-GPV--LRGYI PRAFQ
-SDSLEIVHIAVSK-GRM--IKKYTPKAYG
NTDNLVIKHIAANK-GRM--IKKYTPKAYG
DTDNLVIMHIAAHK-GR I--YKRYV PRAQG
DPDNLVVKHAAAQM-GMK--LRRFF PRA YG
DPNKLFIHVIAAHK-G---V RGYTS-VYG
DLDRLTIVGAVAHK-GIL--IKRFI PRAMG
DLDRLKIVNATVHK-GVI--VKRFI PRAMG
DGE SMTIDHVA AHKVGEQ--VGRQ- PRAFQ
DPNSLVVSEAFAD E-GPT--LKR FQ PRAQ G
DEDALYITEAFVDE-GPT--MKRFQ PRAQ G
DPATLVVATVYADQ-GPT--AKRIR PRAQ G
DADSLFISEAYVDE-GPT--LKRFR PRAQ G
RENDLVVKT TYVDE-GVT--LKRFRARAQ G
RPETLIVKA AFAD E-GPT--LKRIR PRAK G
DPELLFVEEVRVDE-GPT--IKRYR PRA LG
DLDRLYIKKAVADD-GPI--LKKVI PRAHG
NPEELVVRAYVDE-GPT--MKRVR PRA MG
DIDNLYIKEIRAED-GPI--LKRYI PRA YG
ESGELFVTKIFVDG-GAT--LKRMR PAPA QG
EEGELFVSRICVDG-AST--LKR LR PAPA QG
EDADLYIREIFVDG-GRV--LKR LR TAPA QG
EEAGLFVKEIRVDG-GMM--LKR LR PAPA QG
EDANLQVKTIFVDG-GR I--LKRMR PAPA QG
SDNELFVKAVFVDE-GPS--LKR T SAR TDG
```