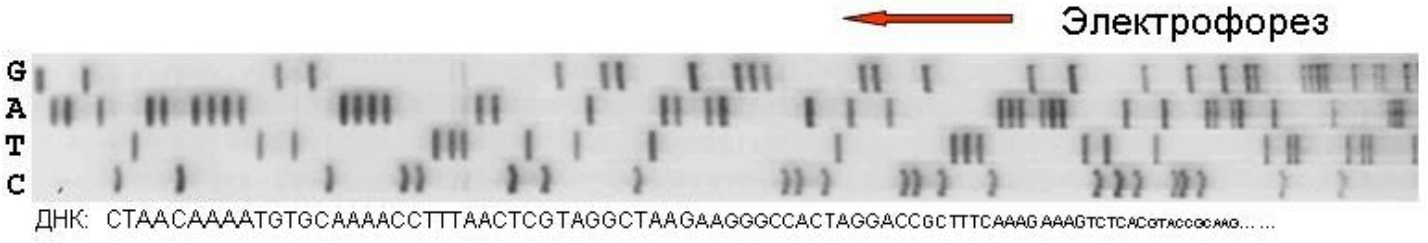
Анализ последовательностей

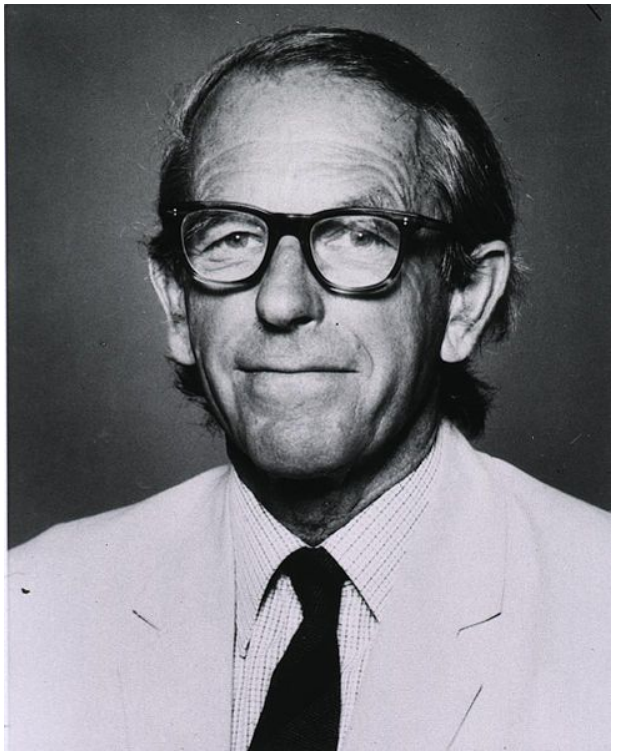
каф. биофизики
биологического ф-та МГУ

Зленко Д.В.

Введение

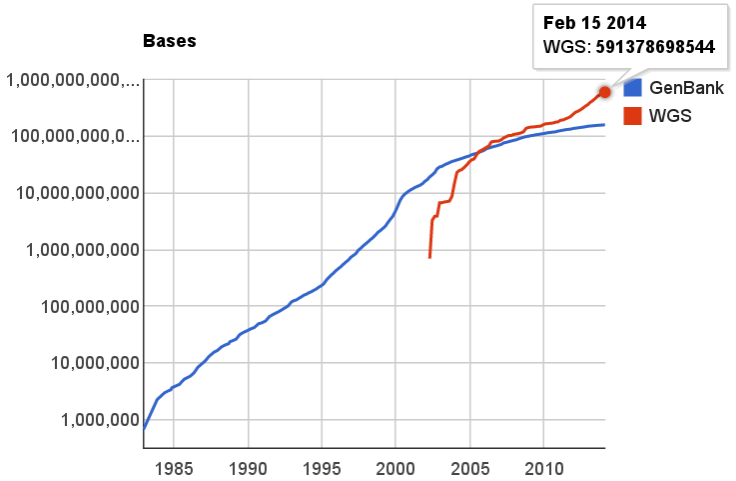
Проблема анализа последовательностей возникла с появлением быстрых методов секвенирования.

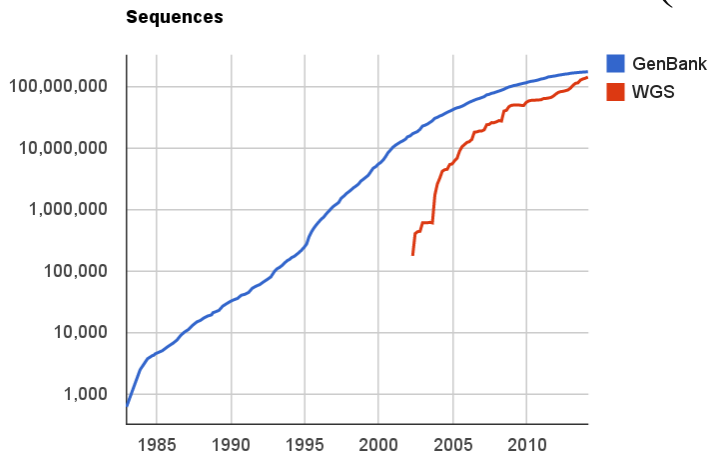




Фредерик Сэнгер
(1918-2013)

- Объем баз данных первичных последовательностей растет экспоненциально.
- База данных GeneBank (NCBI) содержит ~160 млрд. п.о., база данных WGS (Whole Genome Shotgun) ~560 млрд. п.о., а база данных UniProt ~17 млрд. а.к.





Что делать с этим колоссальным объемом данных?

Выравнивание последовательностей

Выравнивание - суть процедура попарного сопоставления остатков в последовательностях, направленное на поиск соответствия между ними. Выравнивание позволяет сопоставить и сравнить последовательности.

```
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEG
PQVGALELAGGPGAGGLEGPPQKRGIVEQCCASVCSLYQLENYCN
```

```
1    MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED   60  P01308   INS_HUMAN
1    MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEG   60  P01317   INS_BOVIN
     **** ** ********* * ** ********************************:***.*

61   LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110  P01308   INS_HUMAN
61   PQVGALELAGGPGAGGL-----EGPPQKRGIVEQCCASVCSLYQLENYCN  105  P01317   INS_BOVIN
      *** :**.******.*      **  **********:*:***********

1    MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED   60  P01308   INS_HUMAN
1    MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEG   60  P01317   INS_BOVIN
     **** ** ********* * ** ********************************:***.*

61   LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110  P01308   INS_HUMAN
61   PQVGALELAGGPGAGGL-----EGPPQKRGIVEQCCASVCSLYQLENYCN  105  P01317   INS_BOVIN
      *** :**.******.*      **  **********:*:***********
```

Считается, что белки с более чем 45% идентичных остатков в выравнивании, это ОЧЕНЬ похожие белки, более 25% - имеют сходную пространственную укладку, 18-25% - "переходная зона". Менее 18% - сходства, как правило, нет.

Тем не менее миоглобин кашалота и леггемоглобин люпина (15%) действительно гомологи, а химотрипсин и субтилизин (12%) родственниками не являются.

Оценка расстояний между последовательностями

• **p**-дистанция - доля не совпадающих в выравнивании остатков.

• Возможность возникновения множественных замен можно учесть, принимая во внимание, что вероятность замены подчиняется распределению Пуассона:

$$P(r,t,k) = \frac{(rt)^k}{k!} e^{-rt}$$

Для пары последовательностей вероятность не обнаружить мутацию:

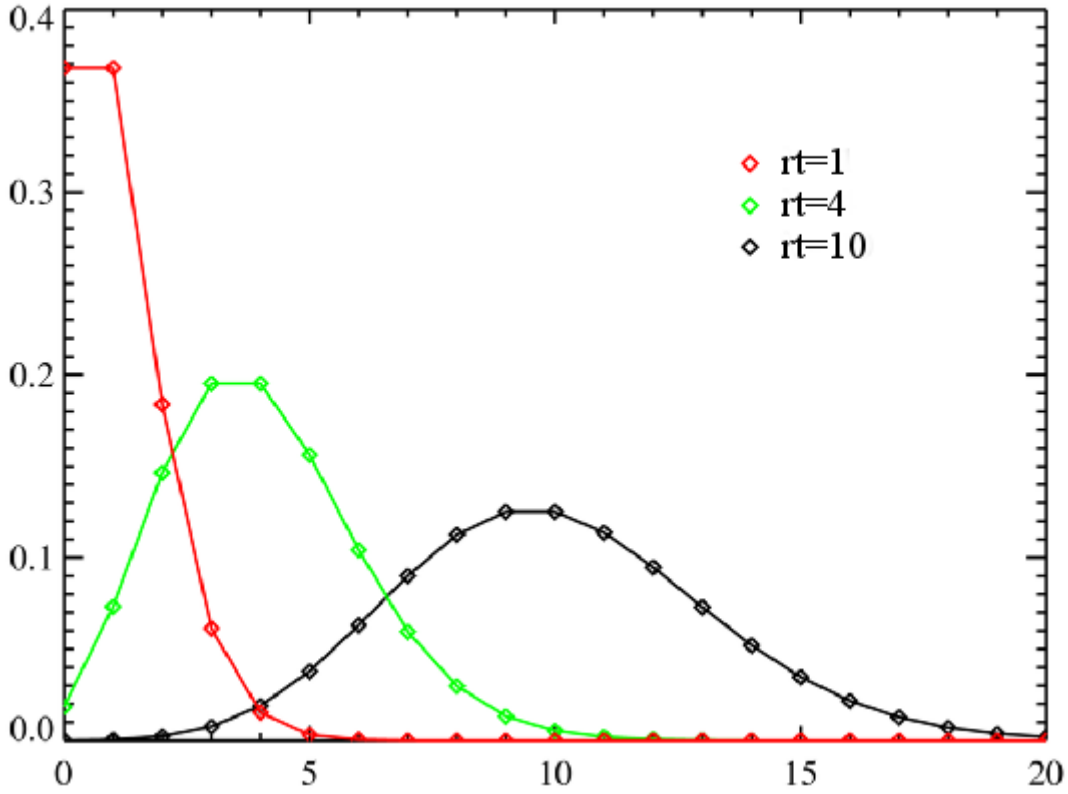
$$q = e^{-2rt}$$

Среднее число мутаций на сайт:

$$d = 2rt$$

Для Пуассон-корректированной дистанции получим:

$$d_{PC} = -ln(1-p)$$



Вариантом Пуассон-корректированной дистанции является дистанция Кимуры - эмпирическое правило для p < 0.7:

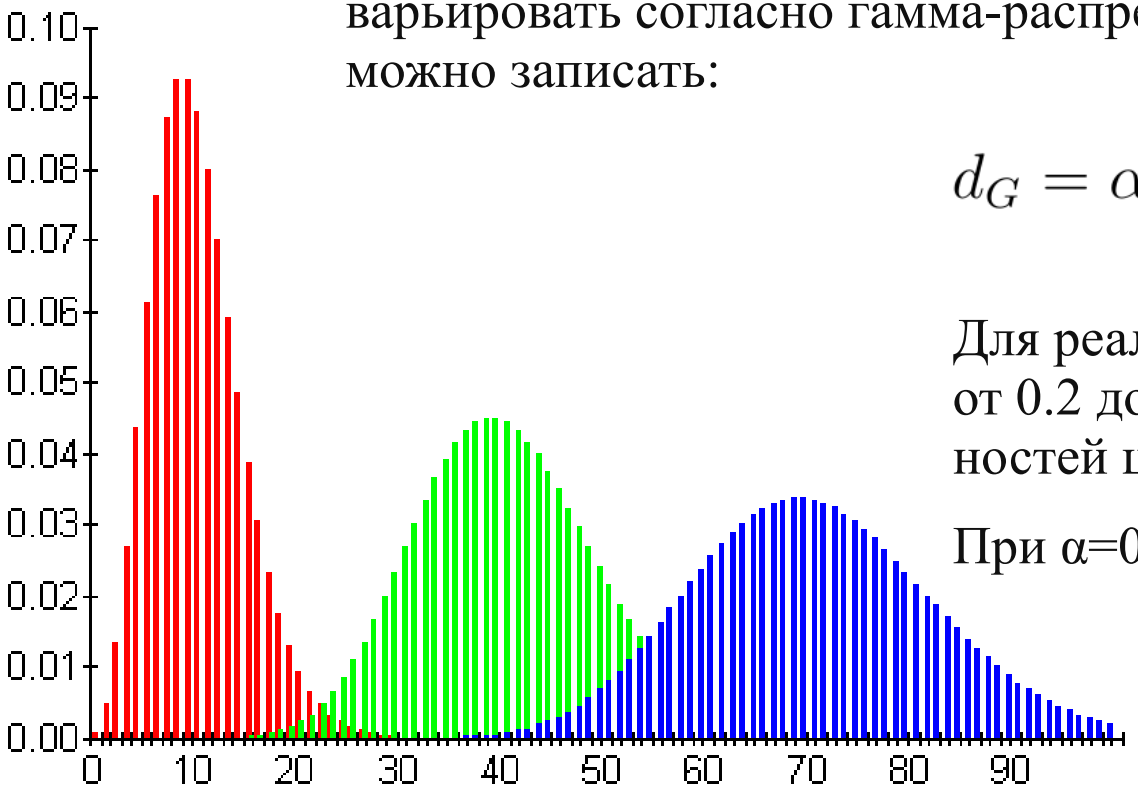
$$d_K = -ln(1-p-\frac{1}{5}p^2)$$

Оценка расстояний между последовательностями

В реальности наблюдается неоднородность в скорости замен от сайта к сайту. Дисперсия количества замен на сайт выше, чем дает распределение Пуассона и приближается к таковой распределения Паскаля:

$$P(k) = C_{k+r-1}^{k}(1-p)^{k}p^{r} \qquad C_{k+r-1}^{k} = \frac{(k+r-1)!}{k!(r-1)!}$$

В этом случае скорость накопления мутаций от сайта к сайту должна варьировать согласно гамма-распределению. Тогда для гамма-дистанции можно записать:



$$d_G = \alpha\left((1-p)^{-\frac{1}{\alpha}} - 1\right)$$

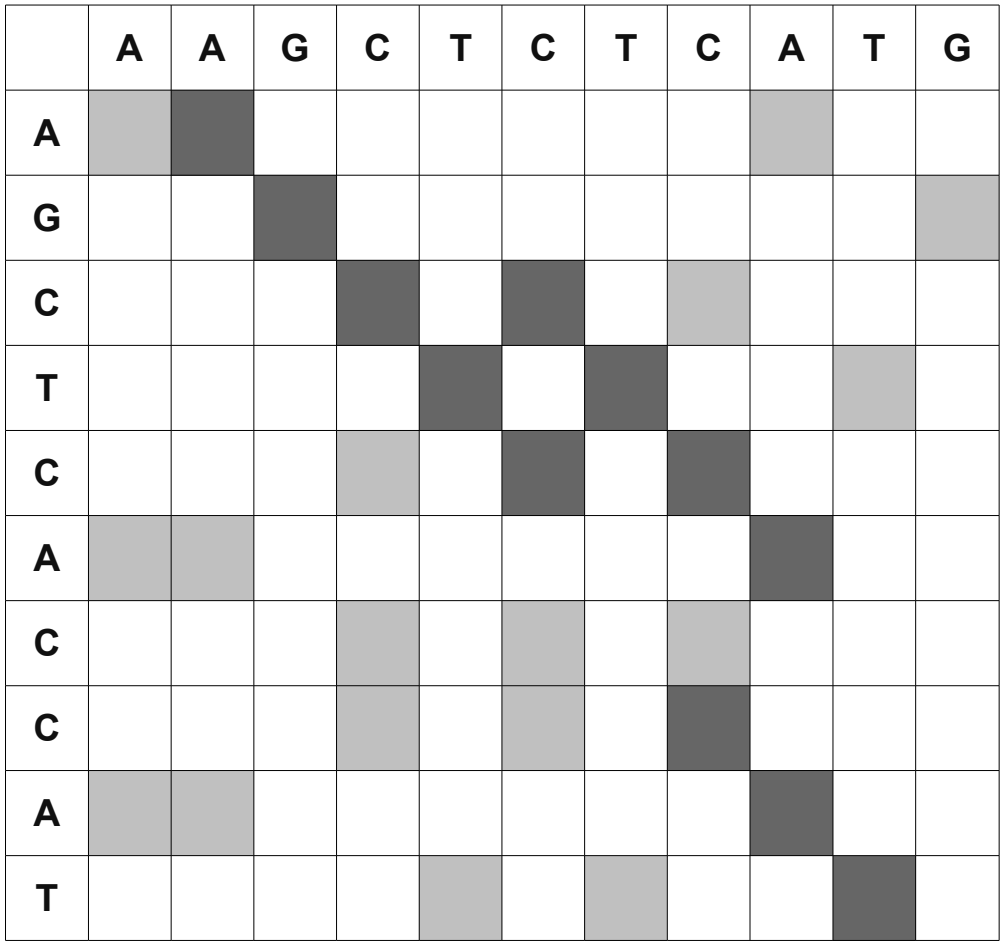
Для реальных белков α варьирует в пределах от 0.2 до 3.5, в частности для последовательностей цитохрома С позвоночных $\alpha \approx 2$.

При α=0.65 d_G переходит в дистанцию Гришина.

$$d_{gri} = 0.65\left((1-p)^{-\frac{1}{0.65}} - 1\right)$$

Точечная матрица сходства

Точечная матрица сходства есть матрица, в которой строки соответствуют одной последовательности, а столбцы другой. В простейшем варианте на пересечении совпадающих элементов выставляют единицы, а в остальных ячейках нули.



Позволяет

- установить участки локального совпадения.
- установить наличие повторов и палиндромов.
- сопоставить по крайней мере фрагменты последовательностей.

AAGCTCT-**CAT**G
-**AGCTC**AC**CAT**-

AAGCT**CTCA**TG--
-**AG**--**CTCA**CCAT

Матрицы аминокислотных замен

Вероятность замены одной аминокислоты на другую зависит от того, какие именно это аминокислоты. Об этом красноречиво свидетельствует статистика.

PAM (1-250) (**P**oint **A**ccepted **M**utations) - матрица аминокислотных замен, предложенная в 1972 году М. Дейхоф, на основании анализа 1572 белковых последовательностей.

Маргарет Дейхоф
(1925 - 1983)

$$M_{ij} = \frac{p_{ij}}{p_i}$$

$$\hat{M}^{nm} = \hat{M}^n \times \hat{M}^m$$

$$\begin{pmatrix} \dots & \dots & \dots \\ \dots & p_{ji} & \dots \\ \dots & \dots & \dots \end{pmatrix} \times \begin{pmatrix} \dots & \dots & \dots \\ \dots & p_{kj} & \dots \\ \dots & \dots & \dots \end{pmatrix} = \begin{pmatrix} \dots & \dots & \dots \\ \dots & +p_{kj} \cdot p_{ji} + & \dots \\ \dots & \dots & \dots \end{pmatrix}$$

C Cys	12																			
S Ser	0	2																		
T Thr	-2	1	3																	
P Pro	-3	1	0	6																
A Ala	-2	1	1	1	2															
G Gly	-3	1	0	-1	1	5														
N Asn	-4	1	0	-1	0	0	2													
D Asp	-5	0	0	-1	0	1	2	4												
E Glu	-5	0	0	-1	0	0	1	3	4											
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F Phe	-4	-3	-3	-5	-5	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

Dayhoff, M.O., Schwartz, R. and Orcutt, B.C. "A model of Evolutionary Change in Proteins". Atlas of protein sequence and structure (volume 5, supplement 3 ed.). Nat. Biomed. Res. Found. pp. 345–358. **1978**.

Матрицы **BLOSUM**
BLOck **SU**bstitution **M**atrix

Для построения белки были собраны в группы по гомологии, после чего были проанализированы замены в консервативных участках между такими группами. Таким образом были созданы матрицы, подходящие для сравнения неродственных белков.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

$$S_{ij} \sim \log\left(\frac{\mathrm{p_{ij}}}{\mathrm{q_i} \cdot \mathrm{q_j}}\right)$$

Henikoff, S.; Henikoff, J.G. "Amino Acid Substitution Matrices from Protein Blocks". PNAS. **1992**. 89(22): 10915–10919.

Цена делеции и вставки

Вопрос о том, насколько важна делеция, по сравнению с заменой остается открытым, хотя примерно величины штрафов можно оценить исходя из вероятностей появления делеций.

Более или менее однозначно установлено, что открытие делеции существенно дороже, чем ее продолжение.

Как правило для выравниваний последовательностей ДНК используют +1 для совпадения, 0 для замены, -10 для открытия и -0.1 для продолжения делеции. Для белков, при использовании матрицы BLOSUM62, штрафы составят -11 за открытие и -1 за продолжение делеции.

Алгоритм Смита-Ватермана

Пусть цена совпадения +2, замены -1, открытия и продолжения делеции -2.

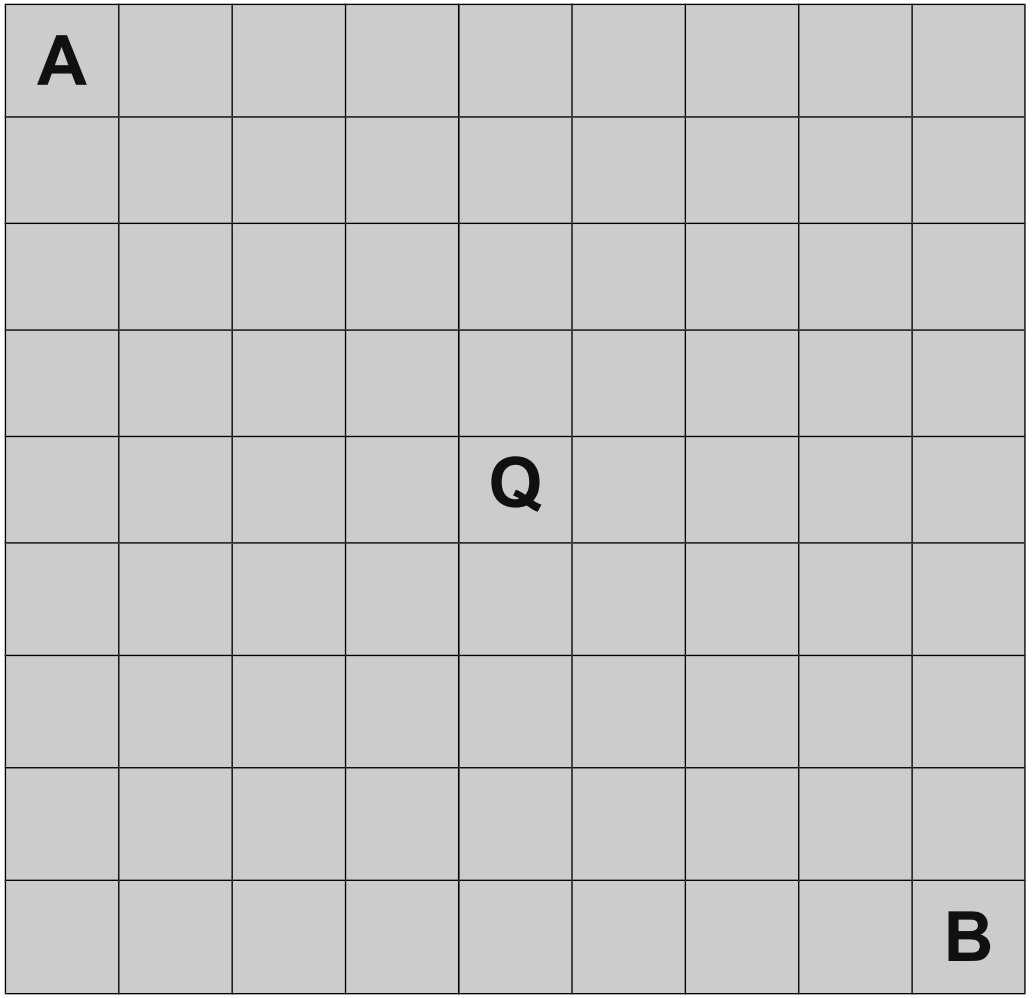
	A	A	G	C	T	A	T	G	C	A	T
G	-1	-1	2	0	-1	-1	-1	2	0	-1	-1
C	-1	-2	0	4	2	1	-2	0	4	2	0
T	-1	-2	-2	2	6	4	3	1	2	3	1
G	-1	-2	0	0	4	5	3	5	3	1	2
C	-1	-2	-2	2	2	3	4	3	7	5	3
T	-1	-2	-3	0	4	2	5	3	5	6	7

A A G C T A T G C A T
- - G C - - T G C - T

A A G C T A T G C A T
- - G C T - - G C - T

T.F. Smith and M.S. Waterman. "Identification of Common Molecular Subsequences". Journal of Molecular Biology. **1981.** 147:195–197.

Редукция задачи



Сколькими способами можно попасть из A в B?

$$N = \frac{(2(n-1))!}{((n-1)!)^2}$$

А если предположить, что путь ОБЯЗАТЕЛЬНО проходит через Q?

Всего путей обхода матрицы 12870, а путей, проходящих через Q только 70.

Эта незамысловатая идея лежит в основе всех методов **динамического программирования**

BLAST

BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) – это семейство алгоритмов, позволяющих выравнивать тестовую последовательность относительно другой последовательности (базы данных).

Алгоритм:

1. Разобьем тестовую последовательность на короткие (~3 а.к. или 11 п.о.) «слова» и будем искать их в базе.

2. Каждый участок локальной гомологии будем расширять до тех пор, пока вес участка локального выравнивания продолжает расти.

3. Для объединения участков локальной гомологии воспользуемся алгоритмом Смита-Ватермана.

В отличие от более ресурсоёмких алгоритмов поиск гомологии при помощи BLAST может привести к не оптимальному выравниванию

Altschul, S; Gish, W; Miller, W; Myers, E; Lipman, D. "Basic local alignment search tool". Journal of Molecular Biology. **1990**. 215 (3): 403–410.

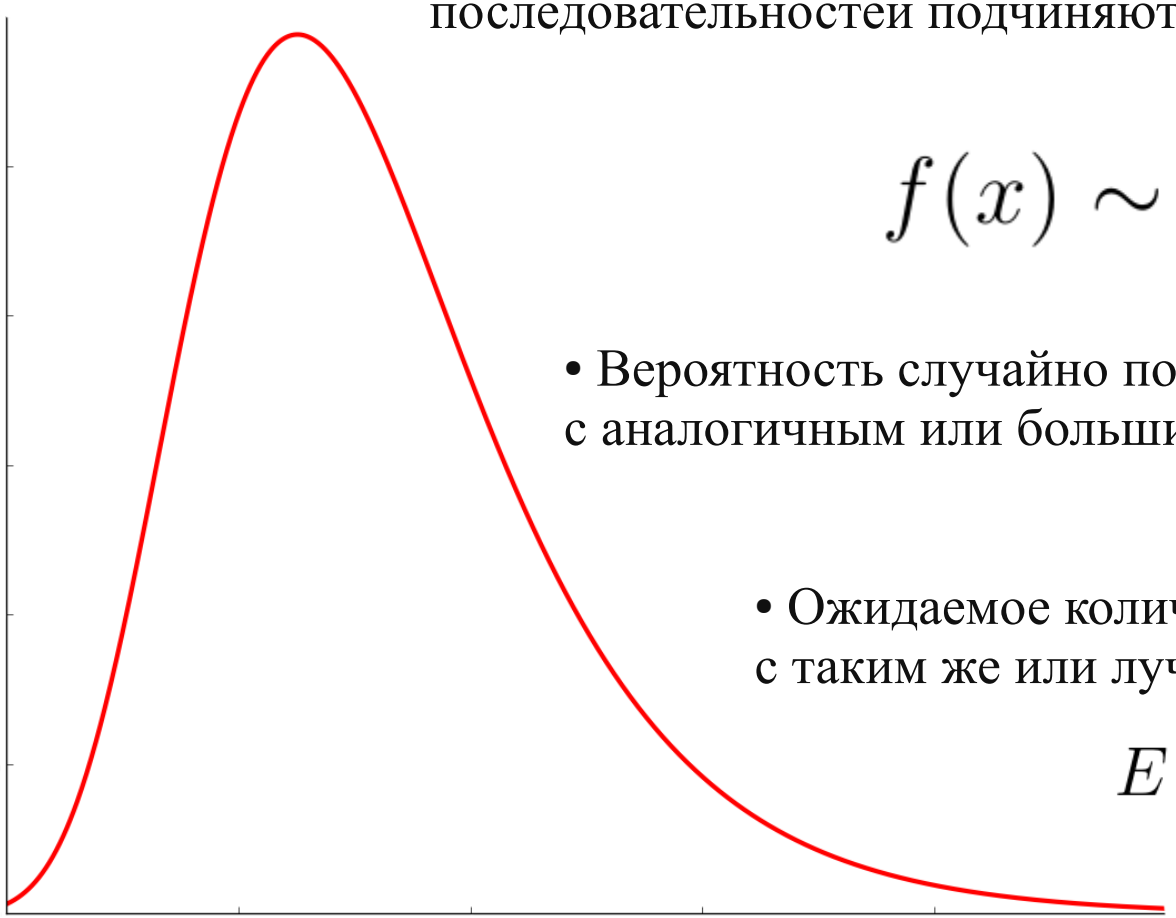
Значимость выравнивания

• Относительная значимость выравнивания может быть выражена Z-score:

$$Z_{score} = \frac{X - \bar{X}}{\sigma(X)}$$

• Веса оптимальных выравниваний для набора случайных последовательностей подчиняются распределению Гумбеля:

$$f(x) \sim e^{-\lambda(x-\mu)} \cdot e^{-Ke^{-\lambda(x-\mu)}}$$



• Вероятность случайно получить последовательность с аналогичным или большим весом:

$$P = 1 - e^{-Ke^{-\lambda(x-\mu)}}$$

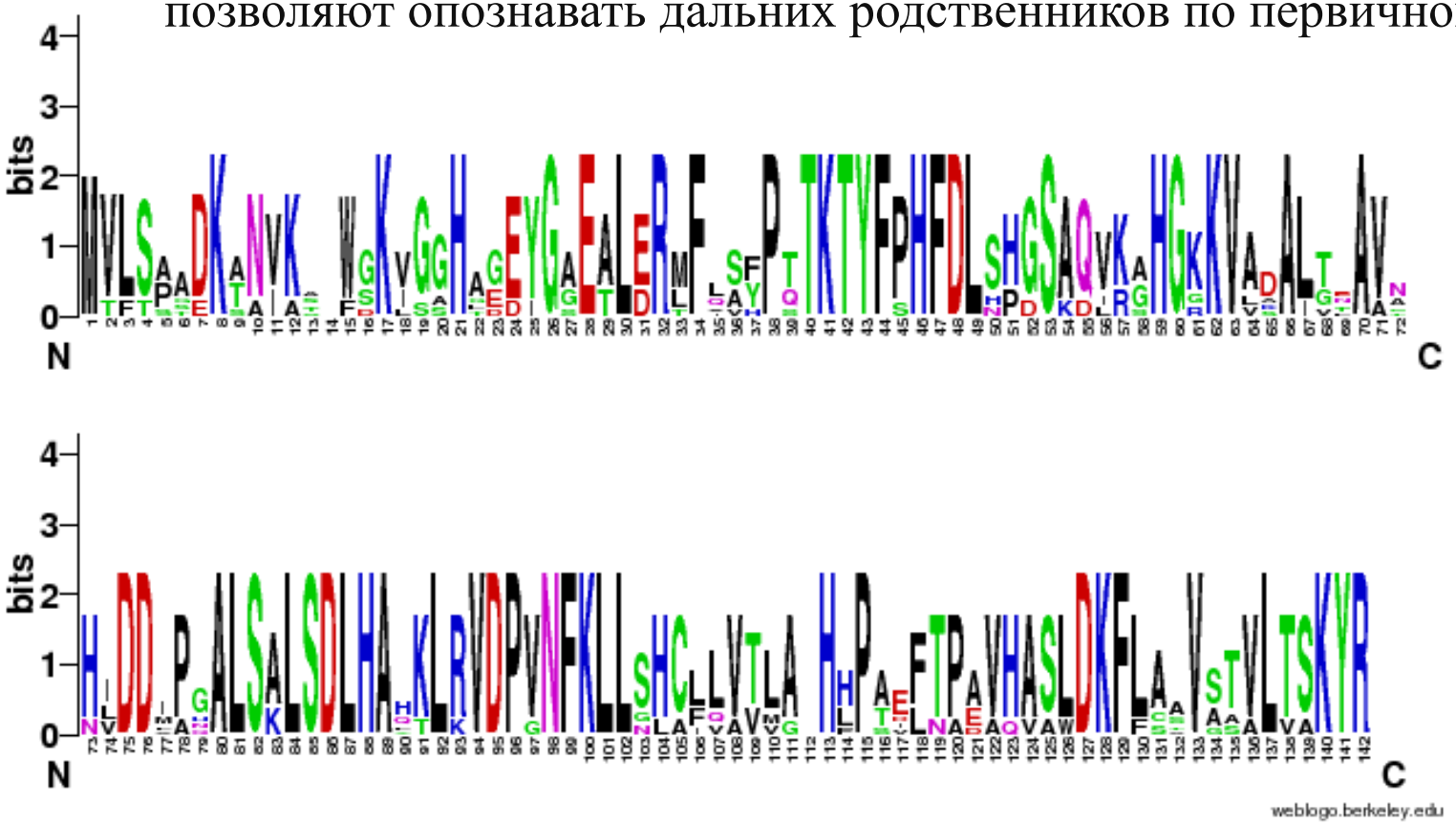
• Ожидаемое количество последовательностей с таким же или лучшим сходством:

$$E = 1 - (1 - P)^N \sim P \cdot N$$

Профили

Профиль - есть набор частот встречаемости остатков определенных позициях среди более или менее однородной выборки достаточно близких гомологов. Идея построения профиля сродни идее построения матрицы замен BLOSUM.

Опыт показывает, что наборы аминокислот высоконсервативных участков позволяют опознавать дальних родственников по первичной структуре.



Профиль для α-цепей гемоглобинов позвоночных (шпорцевой лягушки, слоновой черепахи, филина, ежа, собаки, макака резуса и человека).

http://weblogo.berkeley.edu

PSI-BLAST

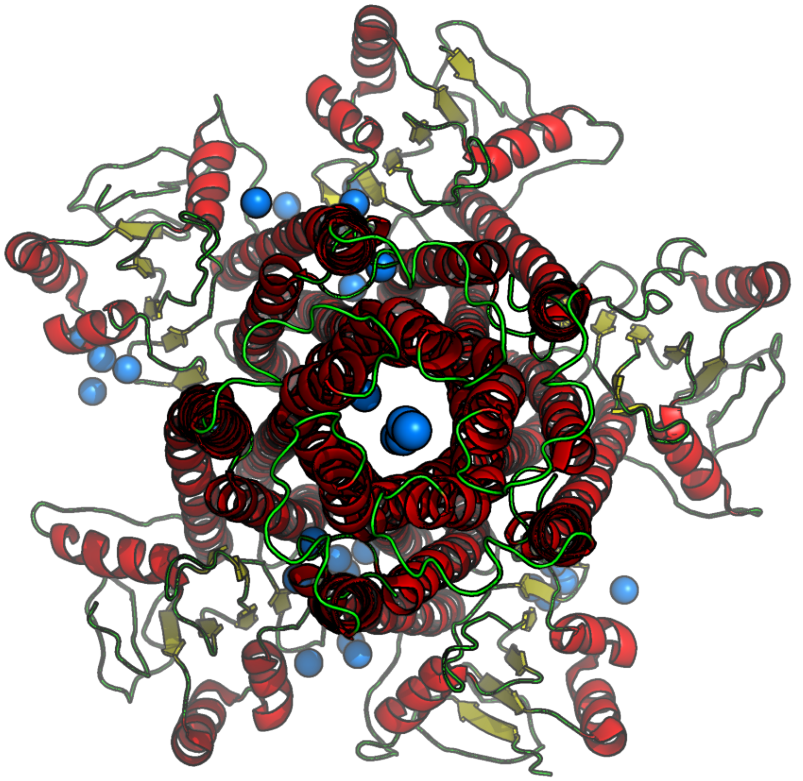
Position **S**pecific **I**terative **BLAST** - продукт совмещения идей динамического программирования и построения профилей:

Алгоритм:

1. Проводим стандартную процедуру BLAST.
2. Строим профиль для начала списка гомологов.
3. Проводим процедуру BLAST с учетом новых весовых коэффициентов.
4. Продолжаем, пока результат не перестанет меняться

PSI-BLAST - мощное средство поиска дальних гомологов.

www.uniprot.org



Methanococcus jannaschii

Alignments	Entry	Entry name	Status	Protein names	Organism	Length	Identity	Score	E-value	Gene names
	Q58439	CORA_METJA	★	**Magnesium transport protein CorA**	Methanocaldococcus jannaschii (strain ATCC 43067 / DSM 2661 / JAL-1 / JCM 10045 / NBRC 100440) (Methanococcus jannaschii)	317	100.0%	1,609	0.0	**corA** MJ1033
	D3S657	D3S657_METSF	☆	**Magnesium and cobalt transport protein CorA**	Methanocaldococcus sp. (strain FS406-22)	317	96.0%	1,562	0.0	MFS40622_1650
	C9RF15	C9RF15_METVM	☆	**Magnesium and cobalt transport protein CorA**	Methanocaldococcus vulcanius (strain ATCC 700851 / DSM 12094 / M7) (Methanococcus vulcanius)	317	91.0%	1,508	0.0	Metvu_0300
	C7P932	C7P932_METFA	☆	**Magnesium and cobalt transport protein CorA**	Methanocaldococcus fervens (strain DSM 4213 / JCM 157852 / AG86) (Methanococcus fervens)	317	90.0%	1,501	0.0	Mefer_1255
	F6BBW1	F6BBW1_METIK	☆	**Magnesium and cobalt transport protein CorA**	Methanotorris igneus (strain DSM 5666 / JCM 11834 / Kol 5)	317	70.0%	1,171	1.0×10^{-159}	Metig_1709
	H1KYJ9	H1KYJ9_9EURY	☆	**Magnesium and cobalt transport protein CorA**	Methanotorris formicicus Mc-S-70	317	69.0%	1,165	1.0×10^{-159}	MetfoDRAFT_0872
	D5VQH2	D5VQH2_METIM	☆	**Magnesium and cobalt transport protein CorA**	Methanocaldococcus infernus (strain DSM 11812 / JCM 15783 / ME)	314	65.0%	1,123	1.0×10^{-152}	Metin_0153
	F8ALJ9	F8ALJ9_METOI	☆	**Magnesium and cobalt transport protein CorA**	Methanothermococcus okinawensis (strain DSM 14208 / JCM 11175 / IH1)	315	66.0%	1,096	1.0×10^{-148}	Metok_1192
	D7DTW4	D7DTW4_METV3	☆	**Magnesium and cobalt transport protein CorA**	Methanococcus voltae (strain ATCC BAA-1334 / A3)	317	62.0%	1,067	1.0×10^{-144}	Mvol_0917
	A6VG53	A6VG53_METM7	☆	**Magnesium and cobalt transport protein CorA**	Methanococcus maripaludis (strain C7 / ATCC BAA-1331)	316	62.0%	1,064	1.0×10^{-143}	MmarC7_0360
	A6USX4	A6USX4_META3	☆	**Magnesium and cobalt transport protein CorA**	Methanococcus aeolicus (strain Nankai-3 / ATCC BAA-1280)	317	61.0%	1,060	1.0×10^{-143}	Maeo_0003
	A4FX61	A4FX61_METM5	☆	**Magnesium and cobalt transport protein CorA**	Methanococcus maripaludis (strain C5 / ATCC BAA-1333)	316	60.0%	1,053	1.0×10^{-142}	MmarC5_0476
	A9AAJ8	A9AAJ8_METM6	☆	**Magnesium and cobalt transport protein CorA**	Methanococcus maripaludis (strain C6 / ATCC BAA-1332)	316	60.0%	1,040	1.0×10^{-140}	MmarC6_1559
	Q6LY83	Q6LY83_METMP	☆	**Magnesium, nickel and cobalt transport protei...**	Methanococcus maripaludis (strain S2 / LL)	316	61.0%	1,032	1.0×10^{-138}	**corA** MMP1108
	G0H0Q5	G0H0Q5_METMI	☆	**Magnesium and cobalt transport protein CorA**	Methanococcus maripaludis (Methanococcus deltae)	316	61.0%	1,032	1.0×10^{-138}	GYY_06365
	F1ZWY2	F1ZWY2_THEET	☆	**Magnesium and cobalt transport protein CorA**	Thermoanaerobacter ethanolicus JW 200	319	29.0%	420	3.0×10^{-46}	TheetDRAFT_1821
	I8R2P7	I8R2P7_9THEO	☆	**Magnesium Mg(2+) and cobalt Co(2+) transport ...**	Thermoanaerobacter siderophilus SR4	319	29.0%	416	1.0×10^{-45}	ThesiDRAFT1_0649
	G2MTF4	G2MTF4_9THEO	☆	**Magnesium and cobalt transport protein CorA**	Thermoanaerobacter wiegelii Rt8.B1	319	29.0%	416	1.0×10^{-45}	Thewi_2044
	D7CN51	D7CN51_SYNLT	☆	**Magnesium and cobalt transport protein CorA**	Syntrophothermus lipocalidus (strain DSM 12680 / TGB-C1)	326	29.0%	401	2.0×10^{-43}	Slip_1373
	A5UVY2	A5UVY2_ROSS1	☆	**Magnesium and cobalt transport protein CorA**	Roseiflexus sp. (strain RS-1)	339	28.0%	398	9.0×10^{-43}	RoseRS_2408

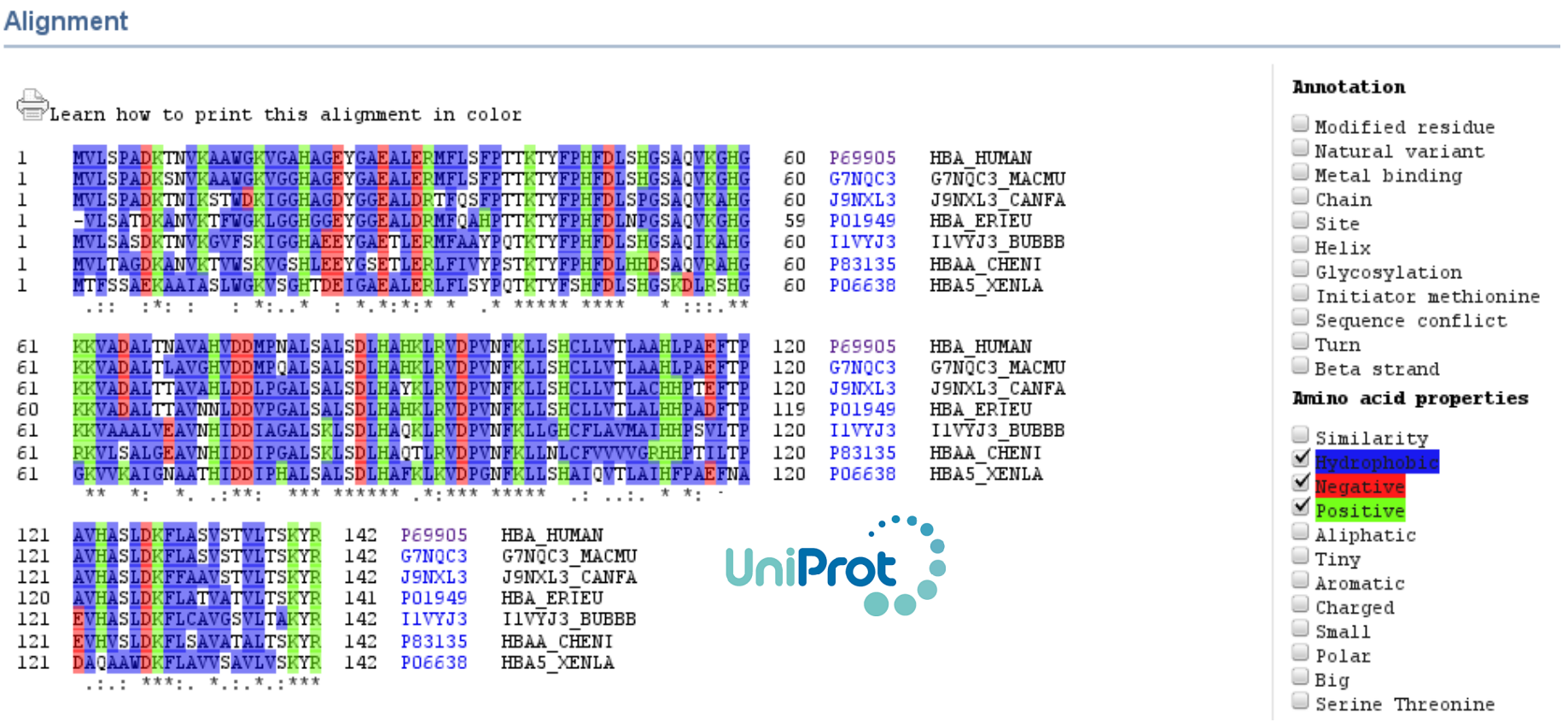
Множественное выравнивание (Clustal)

Множественное выравнивание - есть выравнивание нескольких последовательностей друг относительно друга.

1. Для множества последовательностей выполняют попарное выравнивание.

2. Пара наиболее близких последовательностей служат «затравкой».

3. Относительно затравки с использованием идеологии профилей последовательно выравнивают все остальные последовательности в порядке увеличения несхожести.



Sievers F et. all. "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega". **2011**. Mol Syst Biol 7 7 (539)

Спасибо за внимание!!!