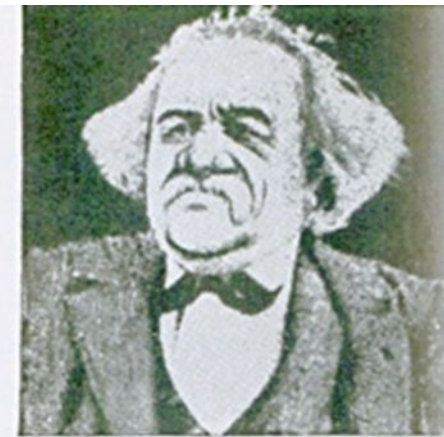


Парное выравнивание. Матрицы замен. BLAST

Лекция 2



Парное выравнивание является самой фундаментальной операцией биоинформатики

- Определяет связаны ли структурно или функционально два белка (или гена)
- Выявляет домены или мотивы, которые являются общими между белками
- Используется для анализа и аннотации генома (поиск и описание генов, участков кодирующих рРНК и тРНК, поиск регуляторных сигналов)

Парные выравнивания: белковые последовательности могут быть более информативными, чем ДНК

- Последовательность белка более информативна (20 против 4 символов); многие аминокислоты имеют общие физико-химические свойства
- Нуклеотидные кодоны вырождены: изменения в третьей позиции часто не приводит к изменению аминокислоты
- Последовательности ДНК могут быть переведены в белковые, и затем использоваться в парных выравниваниях

Принятые однобуквенные коды нуклеиновых кислот

A --> adenosine

C --> cytidine

G --> guanine

T --> thymidine

U --> uridine

R --> G A (purine)

Y --> T C (pyrimidine)

K --> G T (keto)

M --> A C (amino)

S --> G C (strong)

W --> A T (weak)

B --> G T C

D --> G A T

H --> A C T

V --> G C A

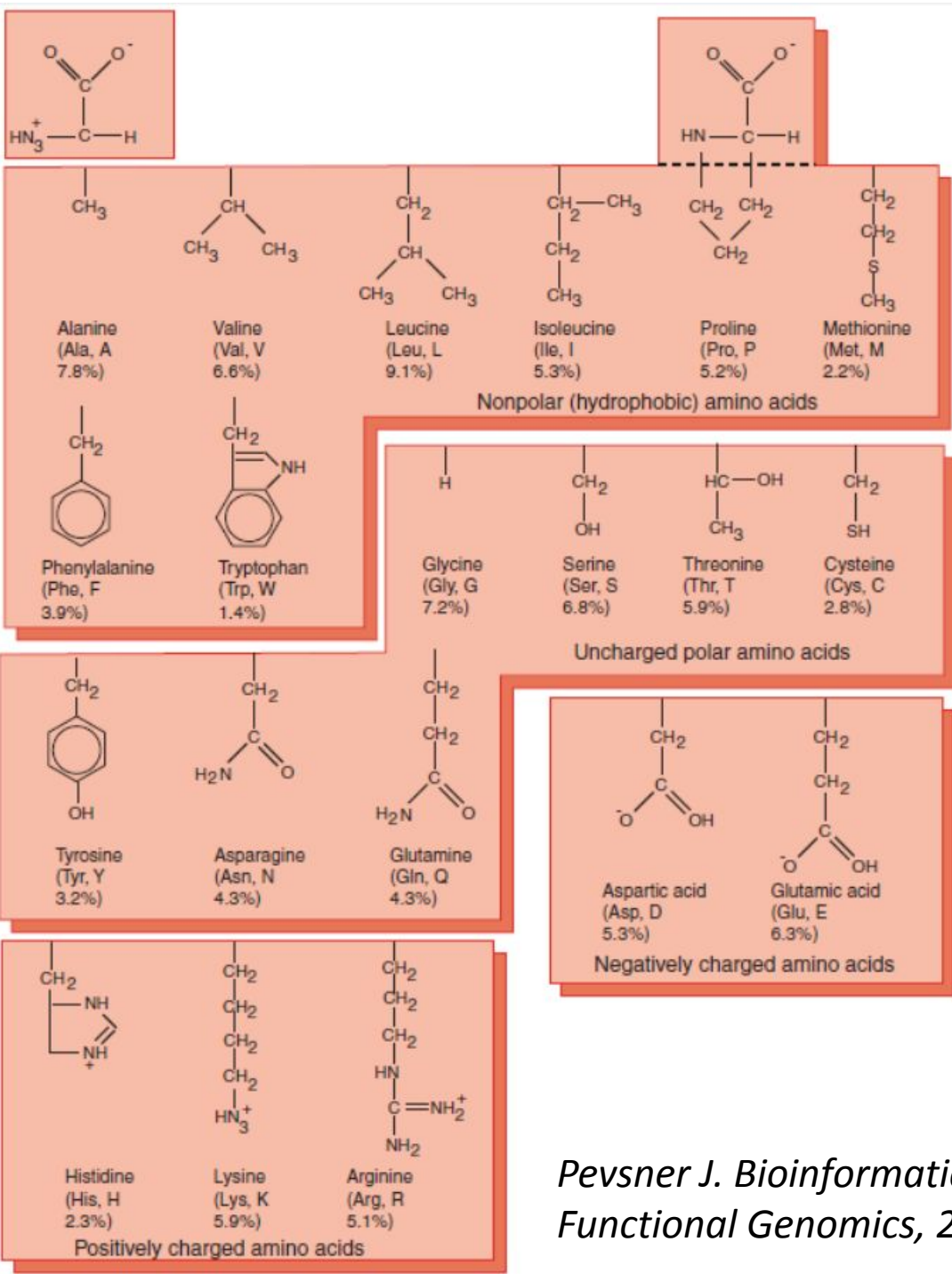
N --> A G C T (any)

– интервал

неопределенной

длины

		Second letter				
		U	C	A	G	
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	U C A G
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G
	A	AUU } AUC } Ile AUA } AUG Met	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U C A G



Pevsner J. Bioinformatics and Functional Genomics, 2009

Принятые однобуквенные коды нуклеиновых кислот

A alanine	P proline
B aspartate or asparagine	Q glutamine
C cystine	R arginine
D aspartate	S serine
E glutamate	T threonine
F phenylalanine	U selenocysteine
G glycine	V valine
H histidine	W tryptophan
I isoleucine	Y tyrosine
K lysine	Z glutamate or glutamine
L leucine	X any
M methionine	* translation stop
N asparagine	– интервал неопределенной длины

Парное выравнивание в 1950-х годах

β -corticotropin (sheep)
Corticotropin A (pig)



Oxytocin
Vasopressin



Определение парного выравнивания

Процесс выравнивания двух последовательностей для достижения максимальных уровней идентичности (и консервативности, в случае аминокислотных последовательностей) с целью оценки степени сходства и возможной гомологии.

Matrix: EBLOSUM62
Gap_penalty: 14
Extend_penalty: 4

Length: 169
Identity: 69/169 (40.8%)
Similarity: 104/169 (61.5%)
Gaps: 8/169 (4.7%)
Score: 282

```
=====
          10          20          30          40          50
KLYTKTGDKGQIGLVGG-RTDKDSLRLVESYGTIDELNSFIGLALAEISGQ
.....: : : : : : : : : : : : : : : : : : : : : : : : :
RIYTRTGDNGTIALFGGSRIDKDDIRVEAYGTVDELISQLGVCYASTRQA
          10          20          30          40          50

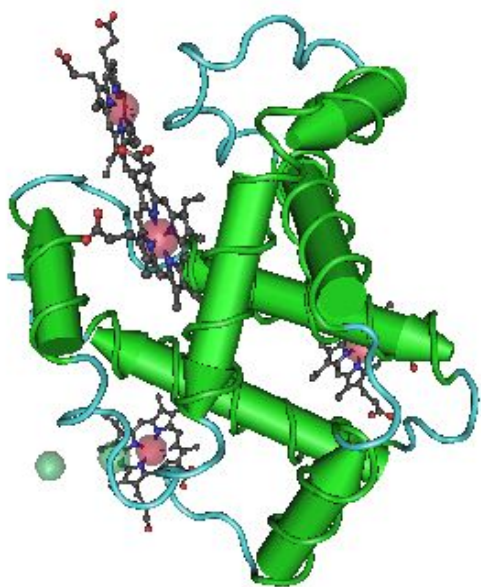
          60          70          80          90
PGFEDLTAELLLIQHELFDCCGDLAIVTE---RKDYKLTEESVSVFLETRI
.: : : : : : : : : : : : : : : : : : : : : : : : :
----ELRQELHAMQKMLFVLGAELASDQKGLTRLKQRIGEEDIQALEQLI
          60          70          80          90

100          110          120          130          140
DAYTAEAPELKKFILPGGSKCASLLHIARTITRRAERRVVALMKSEEIHE
: .. : : : : : . . . : : : : : : : : : : : : . . .
DRNMAQSGPLKEFVIPGKNLASAQLHVARTLTRLRLERILIAMGRITLRLD
100          110          120          130          140

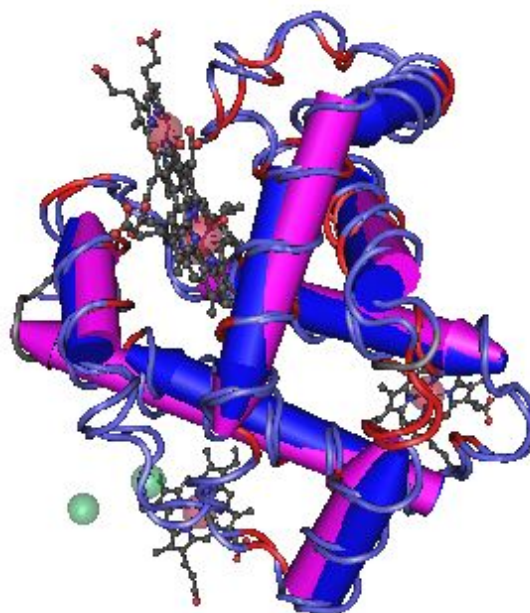
150          160
TVLRYLNRLSDYFFAAARV
.: : : : : : : : : : : : : : : : : : : : : : : : :
EARRYINRLSDALFSMARI
150          160
```

Гомология

Сходство между последовательностями связано с происхождением от общего

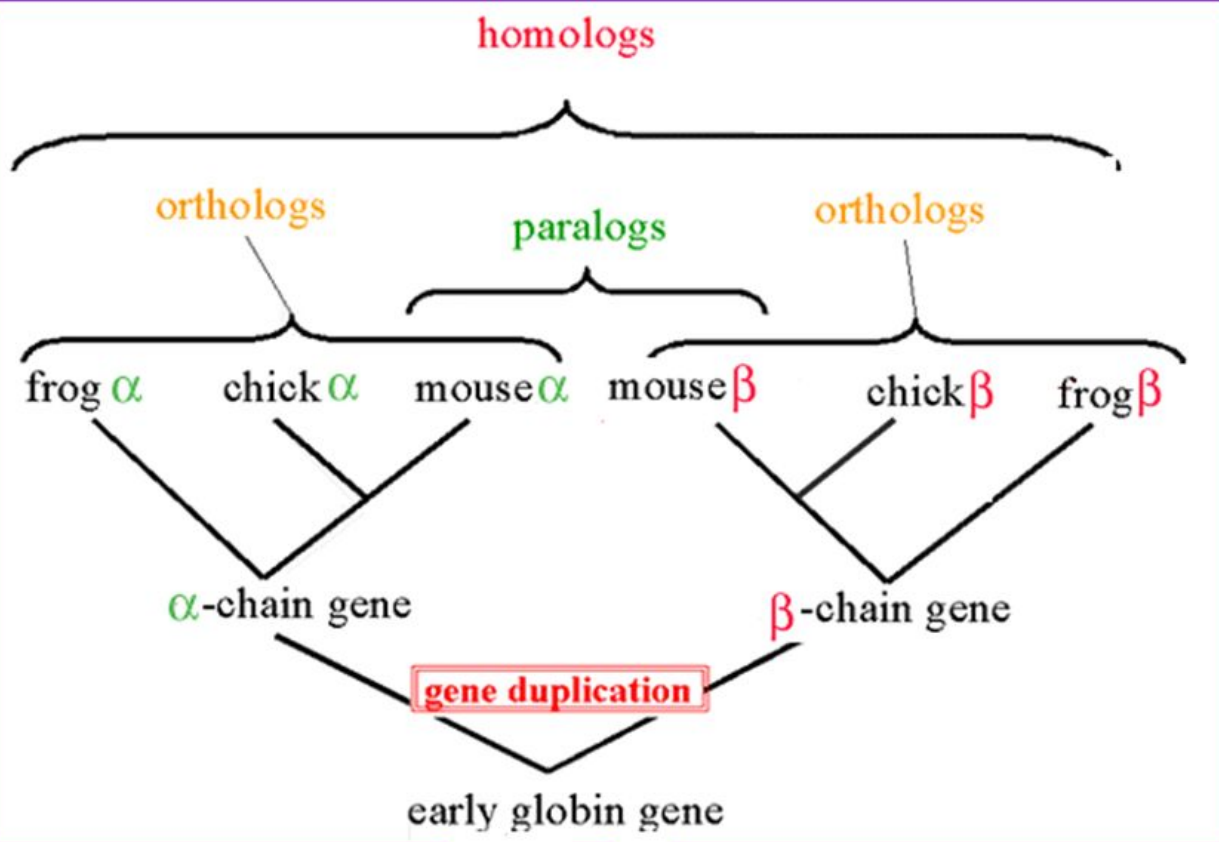


Beta globin
(NP_000509)
2HHB



myoglobin
(NP_005359)
2MM1

Два типа гомологии



Ортолог

и:

Гомологичные последовательности у разных видов, которые возникли из общего предкового гена во время видообразования; могут быть или не быть ответственным за аналогичные функции.

Паралоги: Гомологичные последовательности в пределах одного вида, которые возникли путем дупликации генов.

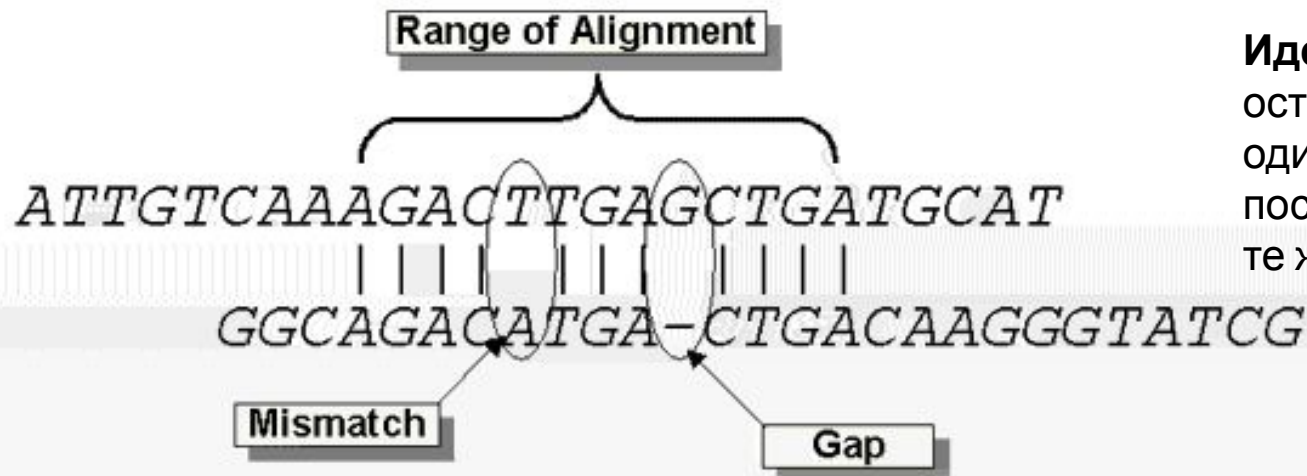
Общий подход к попарному выравниванию

- Выбрать две последовательности
- Выбрать алгоритм, который генерирует оценку сходства
- Определить условия (штраф) для пробелов (вставки, делеции) при выравнивании
- Счет отражает степень сходства
- Выравнивание может быть глобальными или локальными
- Оценить вероятность того, что выравнивание произошло случайно

Редакционное расстояние

- **Элементарное преобразование последовательности**: замена буквы или удаление буквы или вставка буквы.
- **Редакционное расстояние**: минимальное количество элементарных преобразований, переводящих одну последовательность в другую.
- **Формализация задачи сравнения последовательностей**: найти редакционное расстояние и набор преобразований, его реализующий

Расчёт оценки выравнивания (Score)



Идентичность (identity) – остатки (аминокислоты) в одинаковых позициях последовательностей одни и те же. «+» оценка

Несовпадение (mismatch) – остатки (аминокислоты) в одинаковых позициях последовательностей разные. «-» или «+» оценка

$$S = \sum(\text{identities, mismatches}) - \sum(\text{gap penalties})$$

$$\text{Score} = \text{Max}(S)$$

Штраф за пробел (gap penalty) – в одной из последовательностей произошла вставка или делеция, поэтому необходимо добавить пробел. Т.к. такое событие происходит реже, чем изменение остатка, то за это действие вводится штраф. Штрафы могут быть разные: за начало пробела (**gap opening**) и за продолжение пробела (**gap extension**). «-» оценка

http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Alignment_Scores2.html

Сходство последовательностей (Similarity)

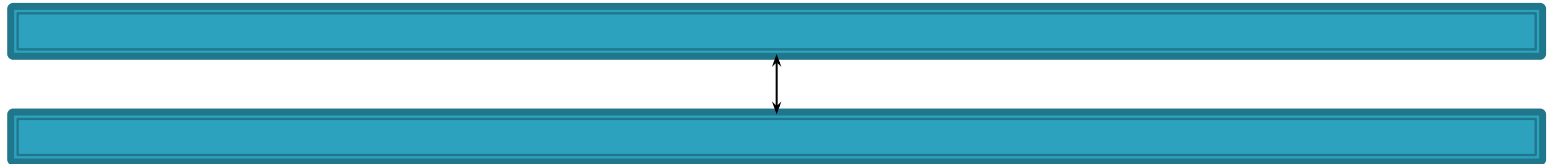
Степень, в которой нуклеотидные или аминокислотные последовательности связаны между собой. Она основана на идентичности и консервативности.

Идентичность (identity) : Степень, в которой две (нуклеотидные или аминокислотные) последовательности одинаковы.

Консервативность (conservation) : Изменения в определенном положении аминокислотного остатка или (реже, нуклеотидного) в последовательности, которые сохраняют физико-химические свойства исходного остатка.

Стратегии выравнивания

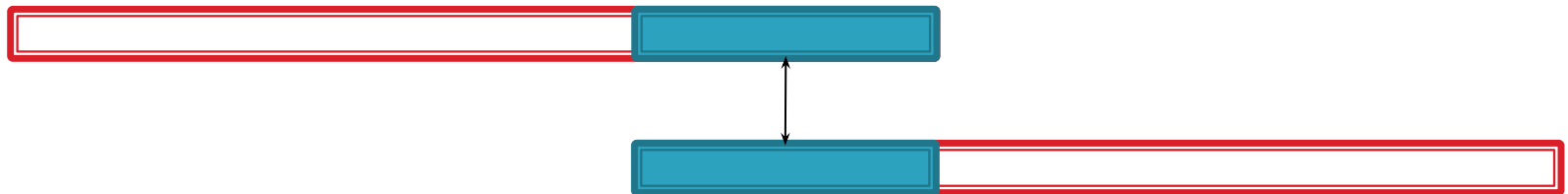
- Глобальное выравнивание



- Локальное выравнивание



- Поиск перекрывающихся последовательностей



NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

New DELTA-BLAST, a more sensitive protein-protein search

BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id—completions will be suggested

- [Human](#)
- [Rabbit](#)
- [Zebrafish](#)
- [Mouse](#)
- [Chimp](#)
- [Clawed frog](#)
- [Rat](#)
- [Guinea pig](#)
- [Arabidopsis](#)
- [Cow](#)
- [Fruit fly](#)
- [Rice](#)
- [Pig](#)
- [Honey bee](#)
- [Yeast](#)
- [Dog](#)
- [Chicken](#)
- [Microbes](#)

Basic BLAST

Choose a BLAST program to run.

[nucleotide blast](#)

Search a **nucleotide** database using a **nucleotide** query
Algorithms: blastn, megablast, discontinuous megablast

[protein blast](#)

Search **protein** database using a **protein** query
Algorithms: blastp, psi-blast, phi-blast, delta-blast

[blastx](#)

Search **protein** database using a **translated nucleotide** query

[tblastn](#)

Search **translated nucleotide** database using a **protein** query

Your Recent Results **New!**

News

[MOLE-BLAST](#)

MOLE-BLAST is a new tool to classify multiple query sequences and discover their relationship to each other.

Thu, 29 Jan 2015 10:00:00 EST

Tip of the Day

[How to Search Custom Databases in Web-Blast Using Entrez Queries.](#)

A powerful feature of the BLAST Web interface is the ability to limit BLAST searches to a subset of any database using a standard Entrez query.

BLAST *Basic Local Alignment Search Tool*

Home Recent Results Saved Strategies Help

NCBI/BLAST/blastp suite

blastn blastp blastx tblastn tblastx

BLASTP programs search protein databases using a protein query

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear Query subrange

From

To

Or, upload file Browse...

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database

Organism

Optional

Entrez Query

Optional

Program Selection

Algorithm

blastp (protein-protein BLAST)

PSI-BLAST (Position-Specific Iterated BLAST)

PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

BLAST Search database nr using Blastp (protein-protein BLAST)

Show results in a new window

Algorithm parameters

Выберем:
Align two or more
sequences...

BLAST Basic Local Alignment

Home Recent Results Saved Strategies Help

NCBI/BLAST/blastp suite

blastn **blastp** blastx tblastn tblastx

BLASTP programs search protein s

Enter Query Sequence

Enter accession number, gi, or FASTA sequence [Clear](#)

```
>gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]
MVHLTPPEKSAVLTALWGKVNVDVVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAH(
AFSDCLAHLDNLIKCTFATLSELHCDKLVDPENFRLLCGMVLVLCVLAHHFCREFTPPVQAAAYQKV
ALAHKYH
```

Or, upload file [Browse...](#)

Job Title [Clear](#)
Enter a descriptive title for your BLAST search

Align two or more sequences

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence [Clear](#)

```
np_005359
```

Or, upload file [Browse...](#)

Program Selection

Algorithm blastp (protein-protein BLAST)
[Choose a BLAST algorithm](#)

BLAST Search protein sequence using Blastp (protein-protein BL) Show results in a new window

Algorithm parameters

Введем две последовательности (accession numbers или в формате fasta format) и кликнем BLAST.

Выберем “Algorithm parameters” и обратим внимание на опцию Matrix.

BLAST Search protein sequence using Blastp (protein-protein BL) Show results in a new window

Algorithm parameters Note:

General Parameters

Max target sequences [Select the maximum number of aligned sequences to display](#)

Short queries Automatically adjust parameters for short input sequences

Expect threshold [?](#)

Word size [?](#)

Scoring Parameters

Matrix **BLOSUM45** [?](#)





Gap Costs [?](#)

Compositional adjustments [?](#)

Результаты парного выравнивания human beta globin и myoglobin

Myoglobin RefSeq

Информация о выравнивании:
score, expect value, identities,
positives, gaps...

```
>  ref|NP_005359.1|  myoglobin [Homo sapiens]
ref|NP_976311.1|  myoglobin [Homo sapiens]
ref|NP_976312.1|  myoglobin [Homo sapiens]
▶ ll more sequence titles
Length=154

GENE ID: 4151 MB | myoglobin [Homo sapiens] (Over 10 PubMed links)

Score = 47.4 bits (144), Expect = 8e-11, Method: Compositional matrix adjust.
Identities = 37/145 (25%), Positives = 57/145 (39%), Gaps = 2/145 (1%)

Query 4 LTPEEKSAVTALWGKVNVDVEVG--GEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPKV 61
L+ E V +WGKV D G E L RL +P T F+ F L + D + + +
Sbjct 3 LSDGEWQLVLMVWGKVEADIPGHGQEVLRIRLFKHPETLEKFDKFKHLKSEDEMKASEDL 62

Query 62 KAHGKKVLGAFSDGLAHLNLDLKGTFATLSELHCDKLHVDPENFRLLGNVLCVLAHHFGK 121
K HG VL A L ++ L++ H K + + + ++ VL
Sbjct 63 KKHGATVLTALGGILKKGHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG 122

Query 122 EFTPPVQAAYQKVVAGVANALAHKY 146
+F Q A K + +A Y
Sbjct 123 DFGADAQGAMNKALELFRKDMASNY 147
```

Query = HBB
Subject = MB

Средняя строка показывает
identities;

+ sign for similar matches

Результаты парного выравнивания human beta globin и myoglobin:

Score = сумма совпадений (match), несовпадений (mismatch), создание пробела (gap creation), и продолжение пробела (gap extension)

Score = 18.1 bits (35), Expect = 0.015, Method: Composition-based stats.
Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

Query	12	VTALWGKVNVD--EVGGEALGRLL	33
		V +WGKV D G E L RL	
Sbjct	11	VLNVWGKVEADIPGHGQEVLIIRLF	34
match		4 11 5 6 6 5 4 5	sum of matches: +60
		6 4 4	
mismatch		-1 1 0 -2 -2 -4 0	sum of mismatches: -13
		-2 0 -3 0	
gap open		-11	sum of gap penalties: -12
gap extend		-1	
			total raw score: 60 - 13 - 12 = 35

V matching V дает +4
T matching L дает -1

Эти оценки даны на основе матрицы замен “scoring matrix”!

Пробелы (gaps)

Score = 18.1 bits (35), Expect = 0.015, Method: Composition-based stats.
 Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

Query	12	VTALWGKVNVD--EVGGEALGRLL								33
		V	+WGKV	D		G	E	L	RL	
Sbjct	11	VLNVWGKVEADIPGHGQEVLRIF								34
match		4	11	5	6	6	5	4	5	sum of matches: +60
				6	4				4	
mismatch		-1	1	0	-2	-2	-4	0		sum of mismatches: -13
		-2		0	-3	0				
gap open					-11					sum of gap penalties: -12
gap extend					-1					
total raw score: 60 - 13 - 12 = 35										



First gap position scores -11

Second gap position scores -1

Создание пробела – большой штраф;

Расширение пробела – небольшой штраф.

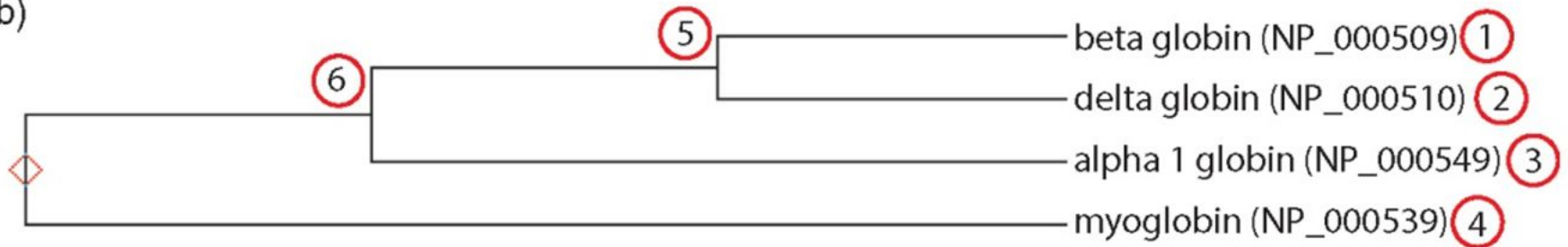
Нахождение предка

(a)

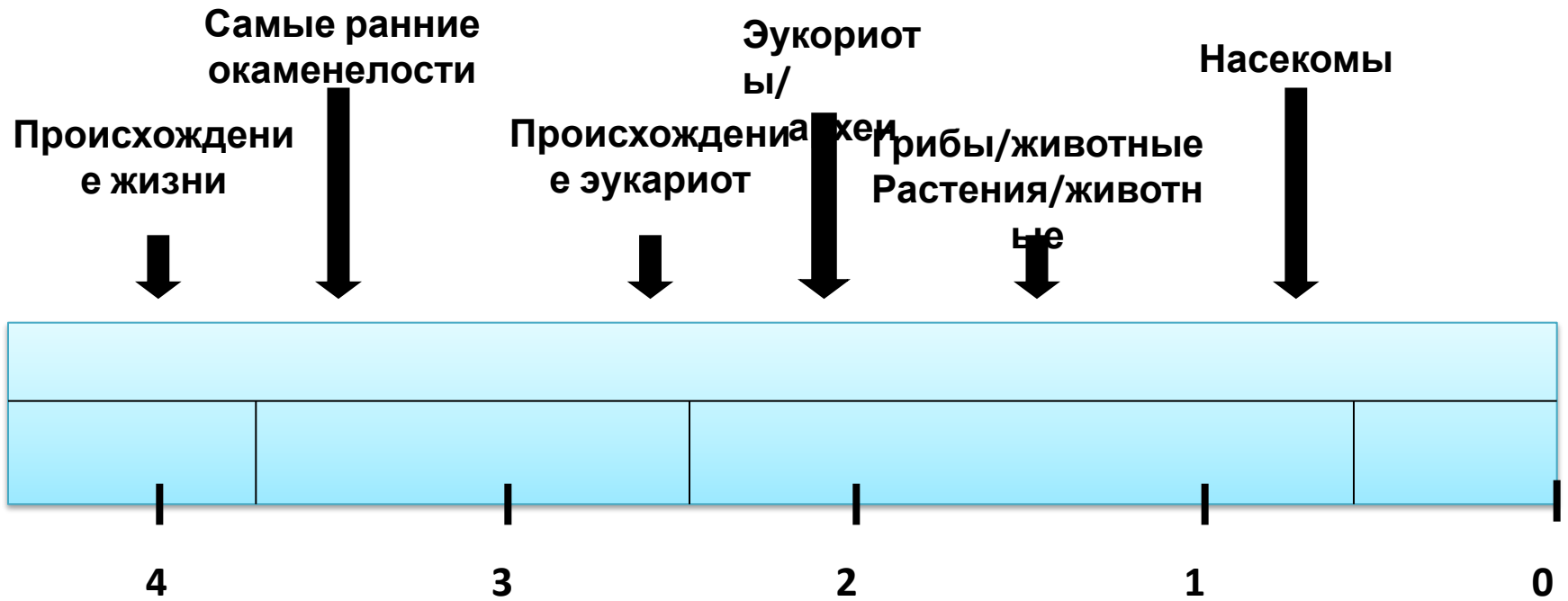
```
beta globin      MVHLTPEEKSAVTALWGKV
delta globin     MVHLTPEEKTAVNALWGKV
alpha 1 globin   MV.LSPADKTNVKAAWGKV
myoglobin        .MGLSDGEWQLVLNVWGKV
5                MVHLSPEEKTAVNALWGKV
6                MVHLTPEEKTAVNALWGKV
```



(b)



Выравнивание парных последовательностей позволяет нам вернуться на миллиарды лет назад



Когда вы делаете попарное выравнивание гомологичных белков человека и растений, вы изучаете последовательности общего предка, жившего 1500000000 лет назад!

Множественное выравнивание последовательностей глицеральдегид 3-фосфат дегидрогеназ: пример очень высокого консерватизма

```
fly          GAKKVIISAP SAD.APM..F VCGVNLDAYK PDMKVVSNAS
CTTNCLAPLA
human       GAKRVIISAP SAD.APM..F VMGVNHEKYD NSLKIISNAS
CTTNCLAPLA
plant       GAKKVIISAP SAD.APM..F VVGVNEHTYQ PNMDIVSNAS
CTTNCLAPLA
bacterium  GAKKVMTGP SKDNTPM..F VKGANFDKY. AGQDIVSNAS
CTTNCLAPLA
yeast       GAKKVVITAP SS.TAPM..F VMGVNEEKYT SDLKIVSNAS
CTTNCLAPLA
archaeon    GADKVLISAP PKGDEPVKQL VYGVNHDEYD GE.DVVSNAS
CTTNSITPVA
```

```
fly          KVINDNFEIV EGLMTTVHAT TATQKTVDGP SGKLWRDGRG
AAQNIIPAST
human       KVIHDNFGIV EGLMTTVHAI TATQKTVDGP SGKLWRDGRG
ALQNIIPAST
plant       KVVHEEFGIL EGLMTTVHAT TATQKTVDGP SMKDWRGGRG
ASQNIIPSST
bacterium  KVINDNFGII EGLMTTVHAT TATQKTVDGP SHKDWRGGRG
ASQNIIPSST
yeast       KVINDAFGIE EGLMTTVHSL TATQKTVDGP SHKDWRGGRT
ASGNIIPSST
archaeon    KVLDEEFGIN AGQLTTVHAY TGSQNLMDGP NGKP.RRRRA
```

Substituent residue
(Percentage of total residue sites at which the substituent occurs)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A				28			31	33								31				
R									50			58				25				
N	33								33			33				33	33			

lys обнаружен в 58% сайтов
arg

Emile Zuckerkandl и Linus Pauling (1965) посчитали частоту замен в 18 глобинах (миоглобины и гемоглобины от человека до миноги).

Черный: Идентичные

Серые: очень консервативные замены (частота >40%)

Белые: слабо консервативные замены (частота >21%)

Red: замен не наблюдалось

Два белка с 50% идентичностью могут иметь 80 изменений среди 100 остатков. (Почему? Потому что, любой остаток может быть предметом обратных мутаций.)

Substituent residue
(Percentage of total residue sites at which the substituent occurs)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A				28			31	33								31				
R								50			58				25					
N	33			47				33			33				33	33				
D	44		22				47	34	22			28				25				
C	(66)																			
Q				56			30		40			70								
E	50			44				38				41		24						
G	51			33			30					27				36				
H				26							26	30				22	22			
I	39										58									46
L	21								23		23		28							30
K	23	21		28			31	23			21					21				
M	22									22	89			22						45
F									22		61									
P	50			43			57	43			21									
S	49			24			24	36			24						40			
T	32						28	24			24					52				
W	(40)									(40)		(60)								
Y									(33)			(50)								
V	36									21	43	21								

Матрицы замен

- Матрица замен содержит значения, пропорциональные вероятности того, что аминокислота i мутирует в аминокислоту j для всех пар аминокислот.
- Матрицы замен строятся путем соединения большого и разнообразного набора проверенных попарных выравниваний (или множественных выравниваний) аминокислот.
- Матрицы замен должны отражать истинные вероятности мутаций, происходящих в течении эволюции.
- PAM и BLOSUM - два основных типа матриц замен.

Основные матрицы замен, применяемые в исследованиях

PAM (Percentage of Acceptable Point Mutations) или матрица Dayhoff. Исходная матрица PAM рассчитана по набору глобальных выравниваний близкородственных белков (>85% идентичность) со средней вероятностью мутации в 1%. Остальные матрицы получены путем возведения матрицы в соответствующую степень. Наиболее часто используется матрица PAM250.

Матрицы серии **BLOSUM** рассчитаны на основе блоков, составленных из непрерывных выравненных фрагментов. Матрица BLOSUM62 рассчитана по выравненным наборам с идентичностью не менее 62%.

Мы можем варьировать:
от $RAM250 = (RAM1)^{250}$,
оценочная матрица,
которая присваивает
баллы и прощает
несоответствия...

(+17 для замены W на W
или -5 для замены W на T)

A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	-2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

34 белковых надсемейства Dayhoff

<u>Белок</u>	<u>PAMs за 100 млн. лет</u>
Ig kappa chain	37
Kappa casein	33
luteinizing hormone b	30
lactalbumin	27
complement component 3	27
epidermal growth factor	26
proopiomelanocortin	21
pancreatic ribonuclease	21
haptoglobin alpha	20
serum albumin	19
phospholipase A2, group IB	19
prolactin	17
carbonic anhydrase C	16
Hemoglobin a	12
Hemoglobin b	12

34 белковых надсемейства Dayhoff

<u>Белок</u>	<u>PAMs за 100 млн. лет</u>
apolipoprotein A-II	10
lysozyme	9.8
gastrin	9.8
myoglobin	8.9
nerve growth factor	8.5
myelin basic protein	7.4
thyroid stimulating hormone b	7.4
parathyroid hormone	7.3
parvalbumin	7.0
trypsin	5.9
insulin	4.4
calcitonin	4.3
arginine vasopressin	3.6
adenylate kinase 1	3.2

34 белковых надсемейства Dayhoff

Белок

PAMs за 100 млн. лет

triosephosphate isomerase 1	2.8
vasoactive intestinal peptide	2.6
glyceraldehyde phosph. dehydrogease	2.2
cytochrome c	2.2
collagen	1.7
troponin C, skeletal muscle	1.5
alpha crystallin B chain	1.5
glucagon	1.2
glutamate dehydrogenase	0.9
histone H2B, member Q	0.9
ubiquitin	0

Парное выравнивание человеческого (NP_005203) и мышиноного (NP_031812) убиквитина

Score = 1316 bits (3407), Expect = 0.0

Identities = 681/685 (99%), Positives = 682/685 (99%), Gaps = 0/685 (0%)

Query	1	MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQLIFAGKQLEDGRTLSDYN	60
		MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQLIFAGKQLEDGRTLSDYN	
Sbjct	1	MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQLIFAGKQLEDGRTLSDYN	60
Query	61	IQKESTLHLVLRRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQLI	120
		IQKESTLHLVLRRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQLI	
Sbjct	61	IQKESTLHLVLRRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQLI	120
Query	121	FAGKQLEDGRTLSDYNIQKESTLHLVLRRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKA	180
		FAGKQLEDGRTLSDYNIQKESTLHLVLRRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKA	
Sbjct	121	FAGKQLEDGRTLSDYNIQKESTLHLVLRRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKA	180
Query	181	KIQDKEGIPSDQQLIFAGKQLEDGRTLSDYNIQKESTLHLVLRRLRGGMQIFVKTLTGKT	240
		KIQDKEGIP DQQLIFAGKQLE GRTLSDYNIQKESTLHLVLRRLRGGMQIFVKTLTGKT	
Sbjct	181	KIQDKEGIPPDQQLIFAGKQLEGGRTLSDYNIQKESTLHLVLRRLRGGMQIFVKTLTGKT	240
Query	241	ITLEVEPSDTIENVKAKIQDKEGIPPDQQLIFAGKQLEDGRTLSDYNIQKESTLHLVLR	300
		ITLEVEPSDTIENVKAKIQDKEGIPPDQQLIFAGKQLEDGRTLSDYNIQKESTLHLVLR	
Sbjct	241	ITLEVEPSDTIENVKAKIQDKEGIPPDQQLIFAGKQLEDGRTLSDYNIQKESTLHLVLR	300

Подход Dayhoff позволяет посчитать оценку замены для любых двух выровненных аминокислотных остатков

$$s_{i,j} = 10 \times \log \left(\frac{q_{i,j}}{p_{i,j}} \right)$$

Dayhoff определяет оценку двух выровненных остатков I, J, как 10 кратный логарифм отношения, частоты их совпадения в природе q (на основе известных последовательностей) на вероятность совпадения этих аминокислот случайно p .

Число "принимаемых точечных мутаций": какие аминокислотные замены происходят в белках?

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	y	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val

Относительная мутабельность аминокислотных остатков

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

Нормализованная частота аминокислотных замен

Gly	8.9%	Arg	4.1%
Ala	8.7%	Asn	4.0%
Leu	8.5%	Phe	4.0%
Lys	8.1%	Gln	3.8%
Ser	7.0%	Ile	3.7%
Val	6.5%	His	3.4%
Thr	5.8%	Cys	3.3%
Pro	5.1%	Tyr	3.0%
Glu	5.0%	Met	1.5%
Asp	4.7%	Trp	1.0%

- **синий** = 6 кодонов; **красный** = 1 кодон

РАМ1 (Point-Accepted Mutations) матрица частоты мутаций

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile
A	9867	2	9	10	3	8	17	21	2	6
R	1	9913	1	0	1	10	0	0	10	3
N	4	1	9822	36	0	4	6	6	21	3
D	6	0	42	9859	0	6	53	6	4	1
C	1	1	0	0	9973	0	0	0	1	1
Q	3	9	4	5	0	9876	27	1	23	1
E	10	0	7	56	0	35	9865	4	2	3
G	21	1	12	11	1	3	7	9935	1	0
H	1	8	18	3	1	20	1	0	9912	0
I	2	2	3	1	2	1	2	0	0	9872



РАМ1 - Встречается одно изменение аминокислоты на 100

РАМ1 (Point-Accepted Mutations) матрица вероятности мутаций


		Original amino acid											
		A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys
Replacement amino acid	A	98.7	0.0	0.1	0.1	0.0	0.1	0.2	0.2	0.0	0.1	0.0	0.0
	R	0.0	99.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.2
	N	0.0	0.0	98.2	0.4	0.0	0.0	0.1	0.1	0.2	0.0	0.0	0.1
	D	0.1	0.0	0.4	98.6	0.0	0.1	0.5	0.1	0.0	0.0	0.0	0.0
	C	0.0	0.0	0.0	0.0	99.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Q	0.0	0.1	0.0	0.1	0.0	98.8	0.3	0.0	0.2	0.0	0.0	0.1
	E	0.1	0.0	0.1	0.6	0.0	0.4	98.7	0.0	0.0	0.0	0.0	0.0
	G	0.2	0.0	0.1	0.1	0.0	0.0	0.1	99.4	0.0	0.0	0.0	0.0
	H	0.0	0.1	0.2	0.0	0.0	0.2	0.0	0.0	99.1	0.0	0.0	0.0
	I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	98.7	0.1	0.0
	L	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.2	99.5	0.0
	K	0.0	0.4	0.3	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0	99.3
	M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0
	F	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0
	P	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0
	S	0.3	0.1	0.3	0.1	0.1	0.0	0.1	0.2	0.0	0.0	0.0	0.1
	T	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1
	W	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	V	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.1	0.0

РАМ1 - Встречается одно изменение аминокислоты на 100

Множественное выравнивание последовательностей глицеральдегид 3-фосфат дегидрогеназ: колонки остатков могут иметь высокую или низкую консервативность

fly	GAKKVIISAP	SAD.APM..F	VCGVNLDAYK	PDMKVVSNAS	CTTNCLAPLA
human	GAKRVIISAP	SAD.APM..F	VMGVNHEKYD	NSLKIISNAS	CTTNCLAPLA
plant	GAKKVIISAP	SAD.APM..F	VVGVNEHTYQ	PNMDIVSNAS	CTTNCLAPLA
bacterium	GAKKVMT G P	SKDNTPM..F	VKGANFDKY.	AGQDIVSNAS	CTTNCLAPLA
yeast	GAKKVITAP	SS.TAPM..F	VMGVNEEKYT	SDLKIVSNAS	CTTNCLAPLA
archaeon	GADKVLISAP	PKGDEPVKQL	VYGVNHDEYD	GE.DVVSNAS	CTTNSITPVA
fly	KVINDNFEIV	EGLMTTVHAT	TATQKTVDGP	SGKLWRDGRG	AAQNIIPAST
human	KVIHDNFGIV	EGLMTTVHAI	TATQKTVDGP	SGKLWRDGRG	ALQNIIPAST
plant	KVVHEEFGIL	EGLMTTVHAT	TATQKTVDGP	SMKDWRGGRG	ASQNIIPSST
bacterium	KVINDNFGII	EGLMTTVHAT	TATQKTVDGP	SHKDWRGGRG	ASQNIIPSST
yeast	KVINDAFGIE	EGLMTTVHSL	TATQKTVDGP	SHKDWRGGRT	ASGNIIPSST
archaeon	KVLDEEFGIN	AGQLTTVHAY	T G SQNLMDGP	NGKP.RRRRA	AAENIIPST
fly	GAAKAVGKVI	PALNGKLTGM	AFRVPTPNVS	VVDLTVRLGK	G ASYDEIKAK
human	GAAKAVGKVI	PELNGKLTGM	AFRVPTANVS	VVDLTCRLEK	PAKYDDIKKV
plant	GAAKAVGKVL	PELNGKLTGM	AFRVPTSNVS	VVDLTCRLEK	G ASYEDVKAA
bacterium	GAAKAVGKVL	PELNGKLTGM	AFRVPTPNVS	VVDLTVRLEK	A ATYEQIKAA
yeast	GAAKAVGKVL	PELQGKLTGM	AFRVPTVDVS	VVDLTVKLNK	ETTYDEIKKV
archaeon	GAAQAATEVL	PELEGKLDGM	AIRVPVPNGS	ITEFVVDLDD	DVTESDVNAA



```

mouse      AIPNPSFLAMPTNENQDNTAIPITIDPITPIVST--PVPTM-----ESIVNTIVANPEAST
rabbit     S--HPFFMAILPNKMQDKAVTPTTNTIAAVEPT--PIPTT-----EPVVSTEVIAEASP
sheep      PHPHLSFMAIIPPKKDQDKTEIPAINTIASAEPTVHSTPTT-----EAVVNAVDNPEASS
cattle     PHPHLSFMAIIPPKKNQDKTEIPTINTIASGEPT--STPTT-----EAVESTVATLEDSP
pig        PRPHASFIAIIPPKKNQDKTAIPAINSIATVEPT--IVPATEPIVNAEPIVNAVVTPEASS
human      PNLHPSFIAIIPPKKIQDKIIIPTINTIATVEPT--PAPAT-----EPTVDSVVTPEAFS
horse      PCPHPSFIAIIPPKKLQEITVIPKINTIATVEPT--PIPTP-----EPTVNNAVIPDASS
.   :   *:*:  .:: *:   *   :.*:.  .*   *:   *   .   .   :   .

```

FIGURE 3.11 Multiple sequence alignment of seven kappa caseins, representing a protein family that is relatively poorly conserved. Only a portion of the entire alignment is shown. Note that just eight columns of residues are perfectly conserved (indicated with asterisks), and gaps of varying length form part of the alignment. In several columns, there are four different aligned amino acids (arrowheads); in two instances there are five different residues (double arrowheads). The sequences were aligned with MUSCLE 3.6 (see Chapter 6) and were human (NP_005203), equine (*Equus caballus*; NP_001075353), pig (*Sus scrofa* NP_001004026), ovine (*Ovis aries* NP_001009378), rabbit (*Oryctolagus cuniculus* P33618), bovine (*Bos taurus* NP_776719) and mouse (*Mus musculus* NP_031812).

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.
 © 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
 Companion Website: www.wiley.com/go/pevsnerbioinformatics

		original amino acid							
PAM ₀	A	R	N	D	C	Q	E	G	
A	100	0	0	0	0	0	0	0	
R	0	100	0	0	0	0	0	0	
N	0	0	100	0	0	0	0	0	
D	0	0	0	100	0	0	0	0	
C	0	0	0	0	100	0	0	0	
Q	0	0	0	0	0	100	0	0	
E	0	0	0	0	0	0	100	0	
G	0	0	0	0	0	0	0	100	

		original amino acid							
PAM _∞	A	R	N	D	C	Q	E	G	
A	8.7	8.7	8.7	8.7	8.7	8.7	8.7	8.7	
R	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	
N	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	
D	4.7	4.7	4.7	4.7	4.7	4.7	4.7	4.7	
C	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	
Q	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8	
E	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	
G	8.9	8.9	8.9	8.9	8.9	8.9	8.9	8.9	

FIGURE 3.12 Portion of the matrices for a zero PAM value (PAM₀; upper panel) or for an infinite PAM_∞ value (lower panel). At PAM_∞ (i.e., if the PAM1 matrix is multiplied by itself an infinite number of times), all the entries in each row converge on the normalized frequency of the replacement amino acid (see **Table 3.1**). A PAM2000 matrix has similar values that tend to converge on these same limits. In a PAM2000 matrix, the proteins being compared are at an extreme of unrelatedness. In contrast, at PAM₀ no mutations are tolerated and the residues of the proteins are perfectly conserved.

РАМ250 матрица вероятности мутаций

Встречается 250 изменений на 100 а.к. остатков

		Original amino acid																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Replacement amino acid	A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
	R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
	N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
	D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
	C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
	Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
	E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
	G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
	H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
	I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
	L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
	K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
	M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
	F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
	P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
	S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
	T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
	W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
	Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	5	31	2
	V	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

FIGURE 3.13 The PAM250 mutation probability matrix. At this evolutionary distance, only one in five amino acid residues remains unchanged from an original amino acid sequence (columns) to a replacement amino acid (rows). Note that the scale has changed relative to **Figure 3.11**, and the columns sum to 100.

РАМ250 логарифмов вероятности замен

A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	-2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Почему мы используем вместо матрицы вероятностей мутаций, матрицу логарифмической вероятности мутаций?

- Оценочная матрица должна быть удобной для попарного выравнивания (или поиска BLAST) и оценки двух выровненных аминокислотных остатков.
- Логарифмы легче использовать для системы оценки. Они позволяют нам суммировать баллы выровненных остатков вместо того, чтобы умножить их.

Переход от матрицы вероятности замен к логарифмической матрице

Оценка S для выравнивания остатков a, b :

$$S(a,b) = 10 \log_{10} (M_{ab}/p_b)$$

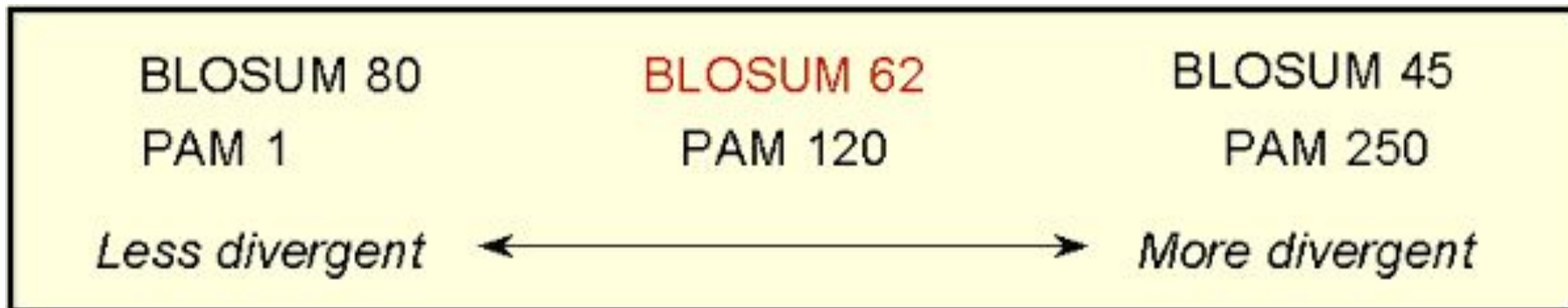
M_{ab} - вероятность замены a на b ; p_b - частота
замены а.к. b

Например, триптофан:

$$S(\text{trp}, \text{trp}) = 10 \log_{10} (0.55/0.010) = 17.4$$

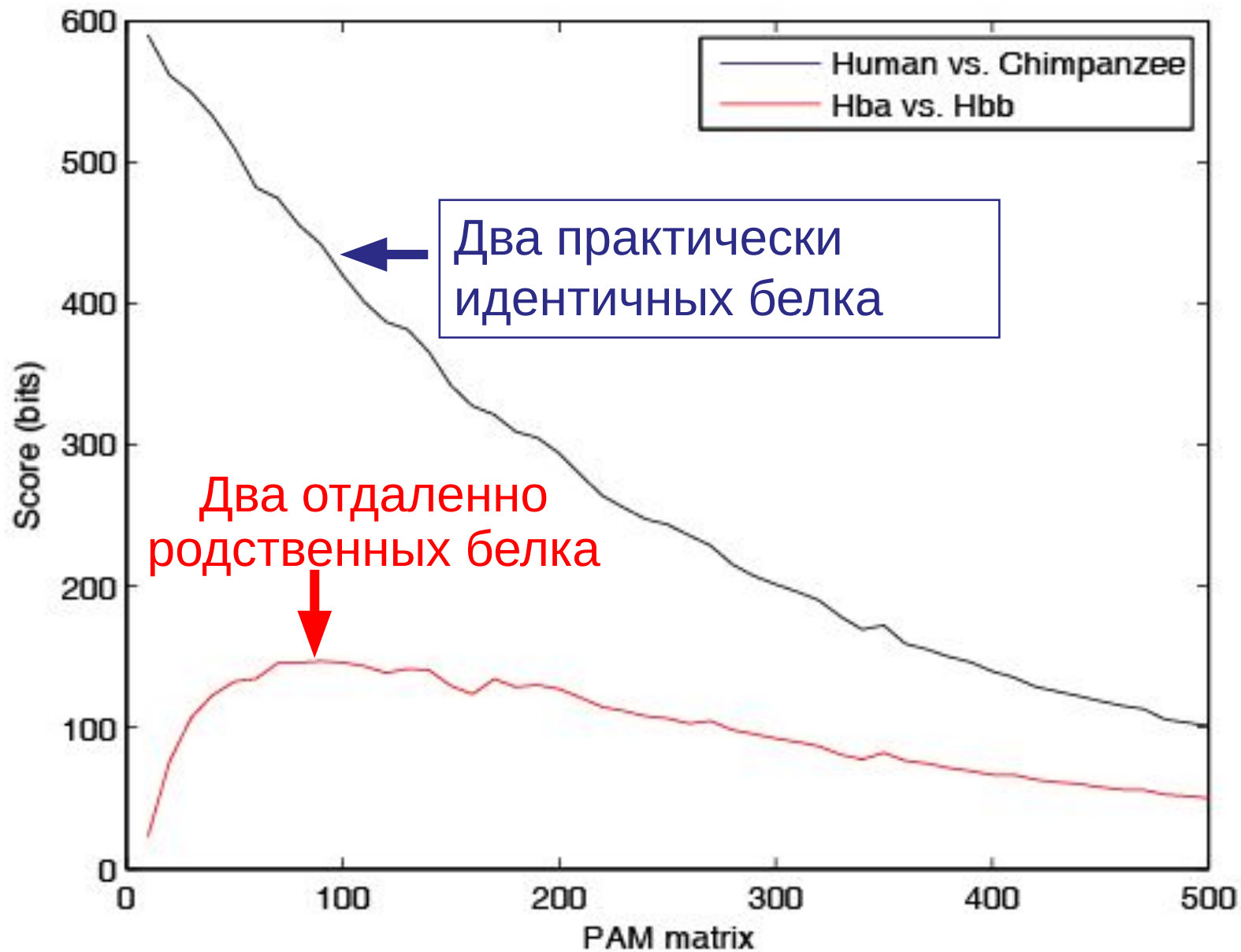
Что означают числа логарифмической матрицы?

- Счет 2 показывает, что замена аминокислоты происходит в 1,6 раза чаще, чем ожидалось случайно.
- Счет 0 является нейтральным.
- Счет -10 означает, что замена аминокислоты в выравнивании происходит в 10 раз медленней, чем ожидалось случайно.



Более консервативный
Глобин кролика и мыши

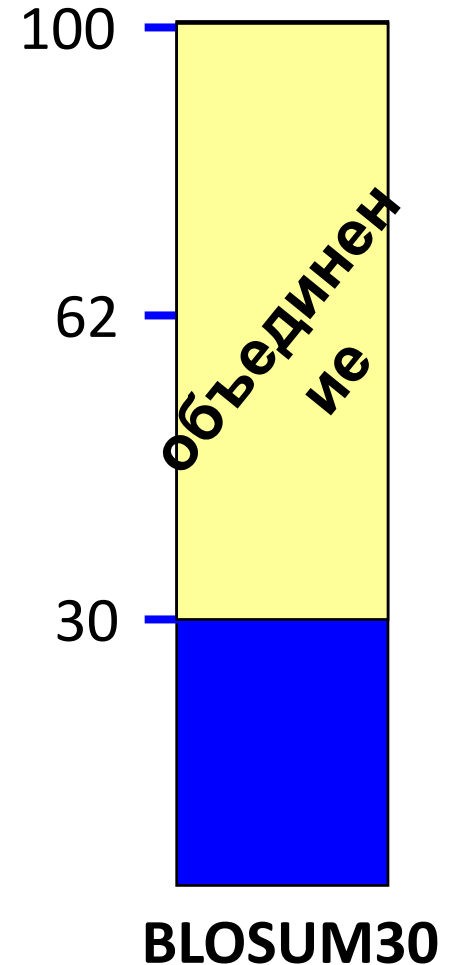
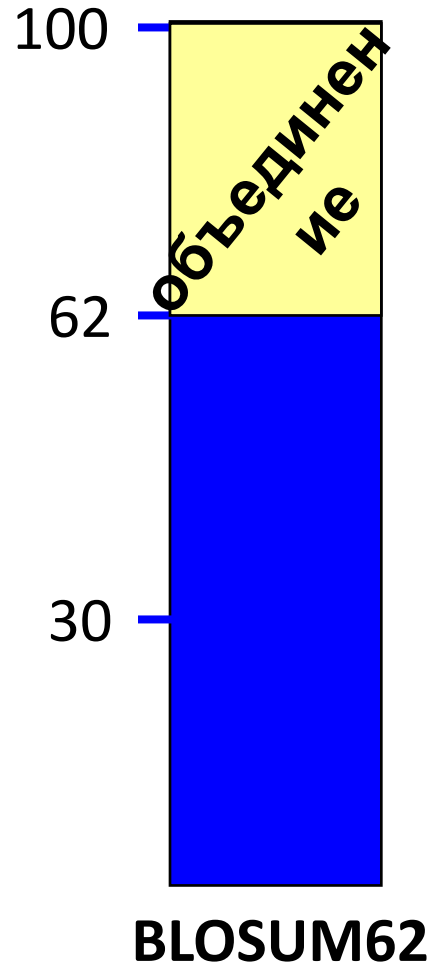
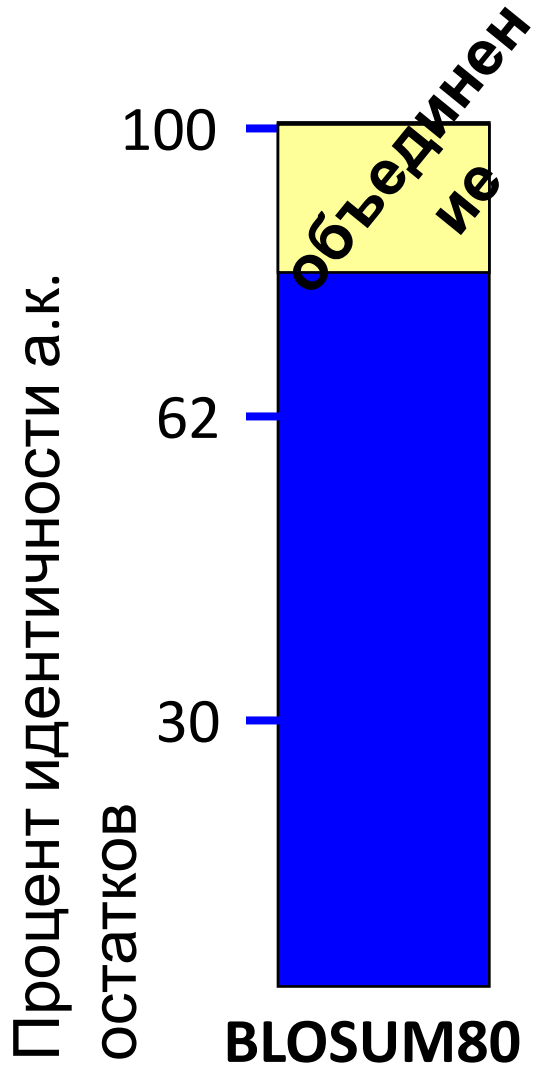
Менее консервативный
Глобин крысы и бактерии



Матрица BLOSUM (Block substitution matrix)

- Основана на локальном выравнивании
- Основана на рассмотрении только консервативных участков (блоков) не близкородственных последовательностей
- BLOSUM62 - матрица вычисленная из сравнения последовательностей с не менее чем 62% -ым расхождением

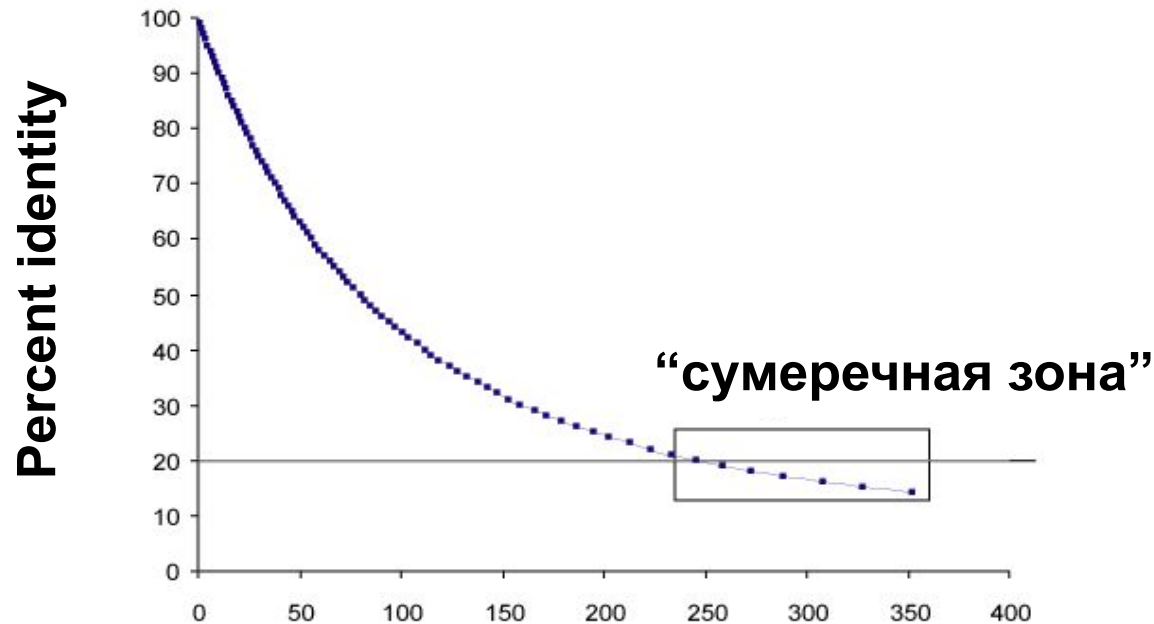
BLOSUM



BLOSUM62

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

Две случайно расходящиеся последовательности белка изменяются обратно экспоненциально



Эволюционное расстояние PAMs

Алгоритмы выравнивания: Ниделмана-Вунша (Needleman-Wunsch) и Смита-Уотермана (Smith-Waterman)

- Алгоритм глобального выравнивания Ниделмана-Вунша (1970)
- Алгоритм локального выравнивания Смита-Уотермана (1981)
- BLAST (Basic Local Alignment Search Tool), эвристическая версия Смита-Уотермана

Алгоритм глобального выравнивания Ниделмана-Вунша

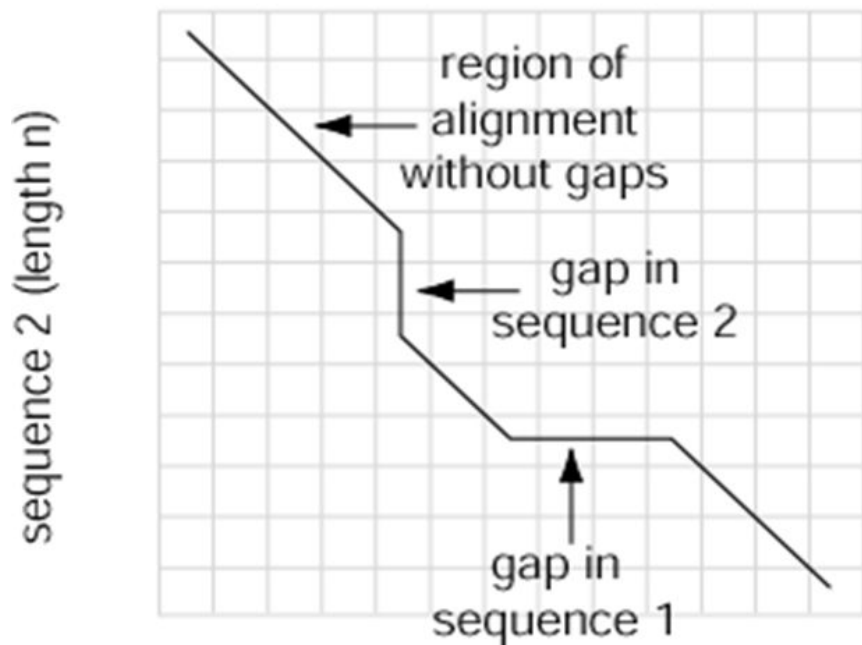
- Две последовательности сравниваются в матрице с осями X и Y (каждая из осей является соответствующей последовательностью)
- Если остатки в позиции одинаковые, то путь в этой ячейке рисуется в виде диагонали
- Поиск оптимальных подпутей, и их добавление для достижения лучшего результата. Включает:
 - Добавление если нужно пробелов
 - Разрешение консервативных замен
 - Изменение системы оценки (скоринга)
- Гарантирует нахождение оптимального выравнивания

Три шага в глобальном выравнивании алгоритмом Ниделмана-Вунша

- Построить матрицу
- Оценка матрицы
- Выбрать оптимальное выравнивание

Четыре возможных исхода при выравнивании двух последовательностей

sequence 1 (length m)



[1] идентичность

(оставаться вдоль диагонали)

[2] несовпадение

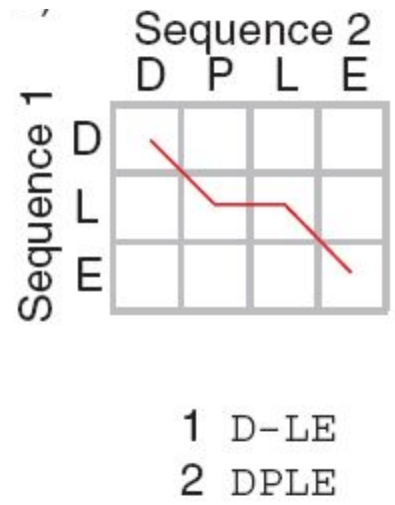
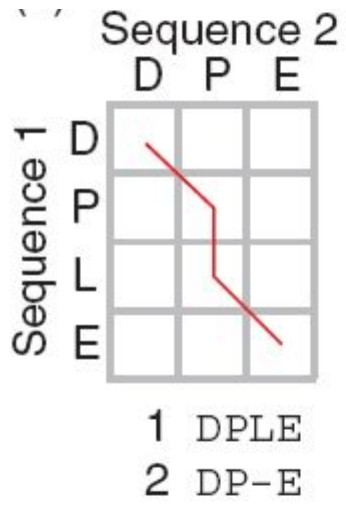
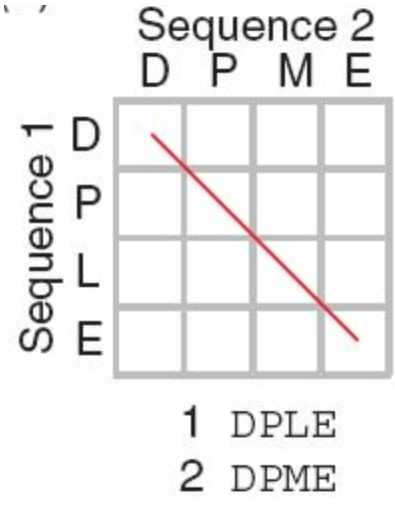
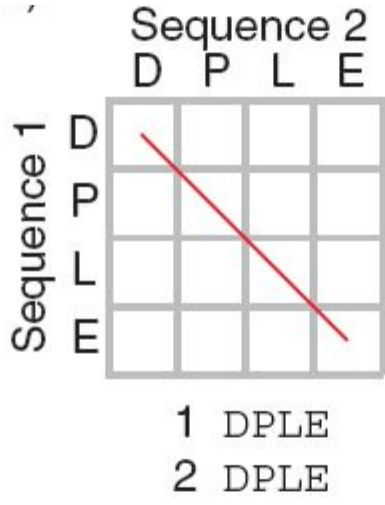
(оставаться вдоль диагонали)

**[3] пробел в одной
последовательности**

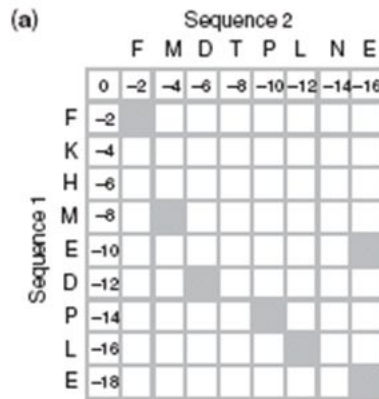
(передвижение по вертикали!)

**[4] пробел в другой
последовательности**

(передвижение по горизонтали!)



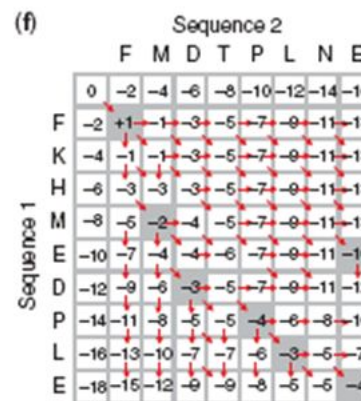
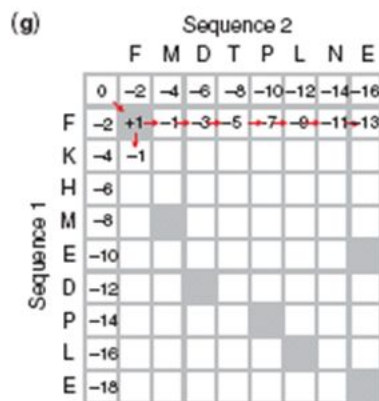
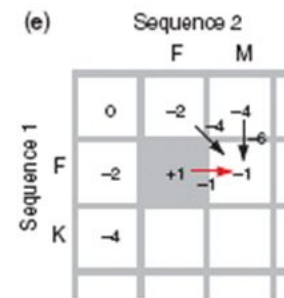
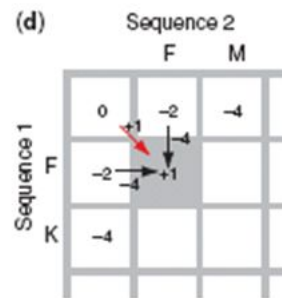
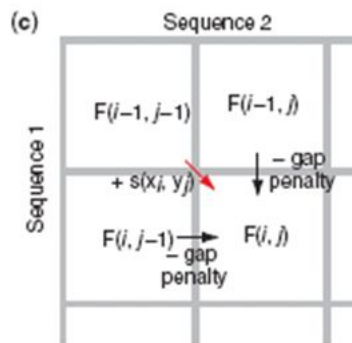
Заполнение матрицы с использованием «динамического программирования»



(b)

$$\text{Score} = \text{Max} \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - \text{gap penalty} \\ F(i, j-1) - \text{gap penalty} \end{cases}$$

Score (this example) = +1 (match)
 -2 (mismatch)
 -2 (gap penalty)



Заполнение матрицы с использованием «динамического программирования»

(a)

Sequence 2

	F	M	D	T	P	L	N	E	
Sequence 1	0	-2	-4	-6	-8	-10	-12	-14	-16
F	-2								
K	-4								
H	-6								
M	-8								
E	-10								
D	-12								
P	-14								
L	-16								
E	-18								

Алгоритм начинается с построения матрицы идентичности

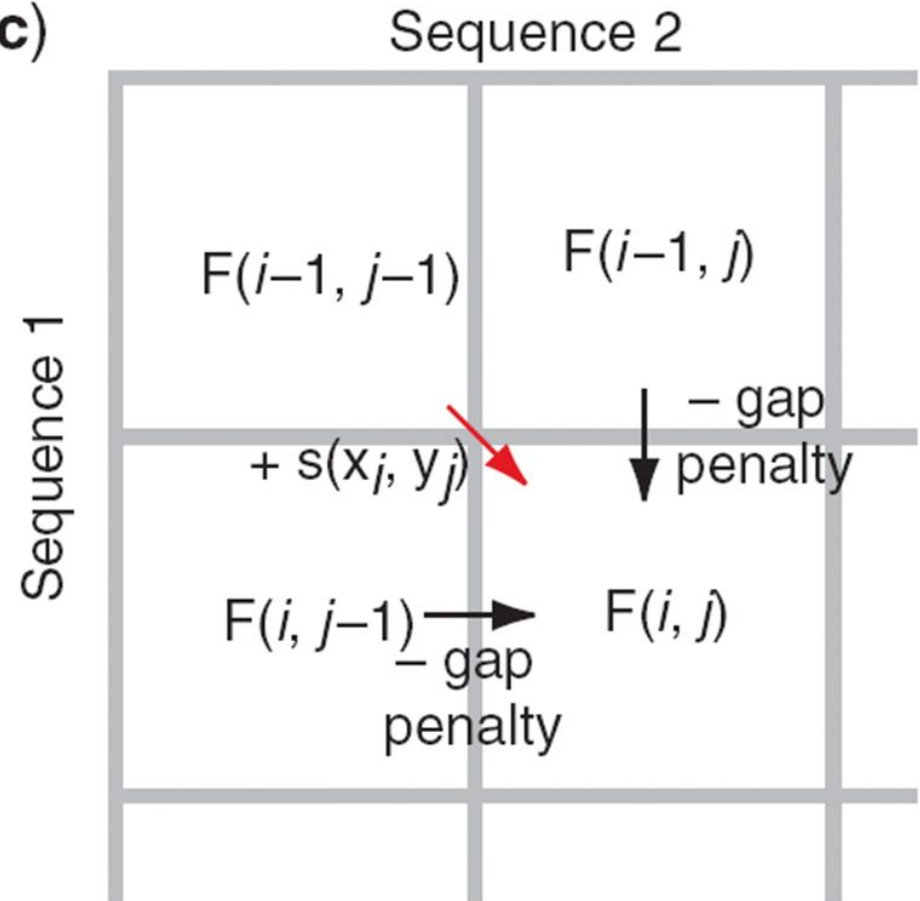
Заполнение матрицы с использованием «динамического программирования»

(b)

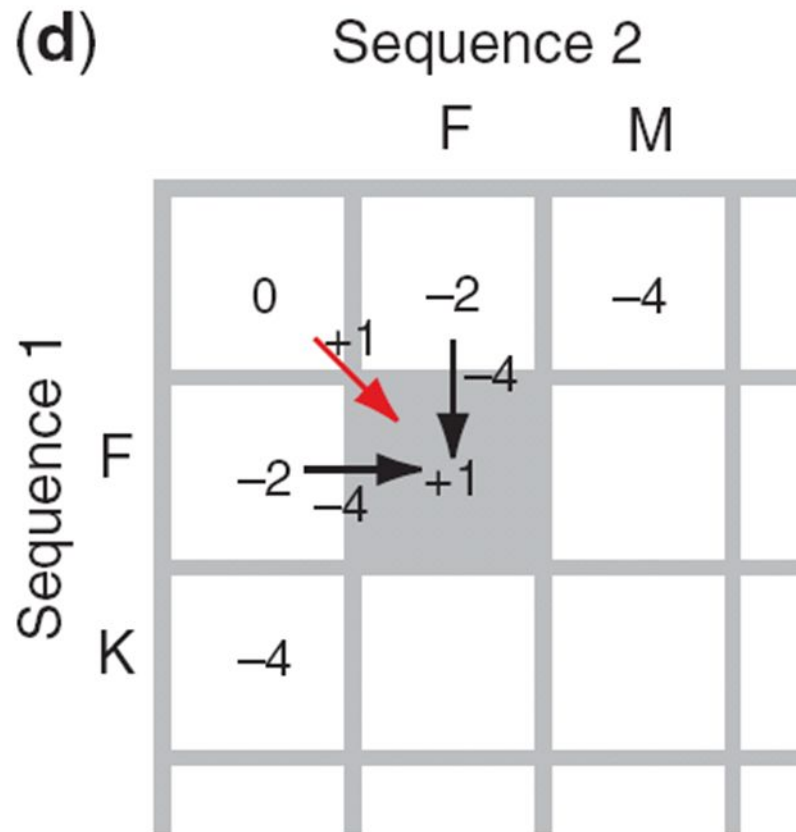
$$\text{Score} = \text{Max} \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - \text{gap penalty} \\ F(i, j-1) - \text{gap penalty} \end{cases}$$

Score (this example) = +1 (match)
 -2 (mismatch)
 -2 (gap penalty)

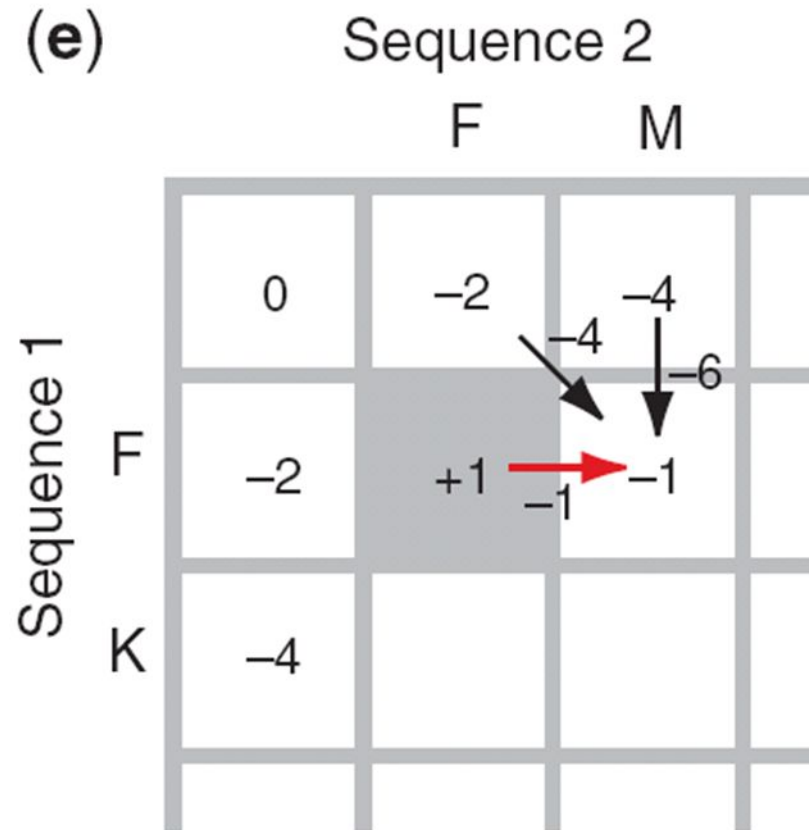
(c)



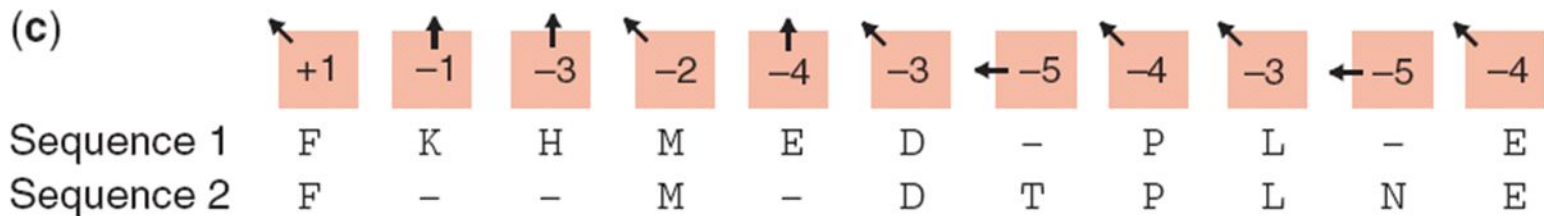
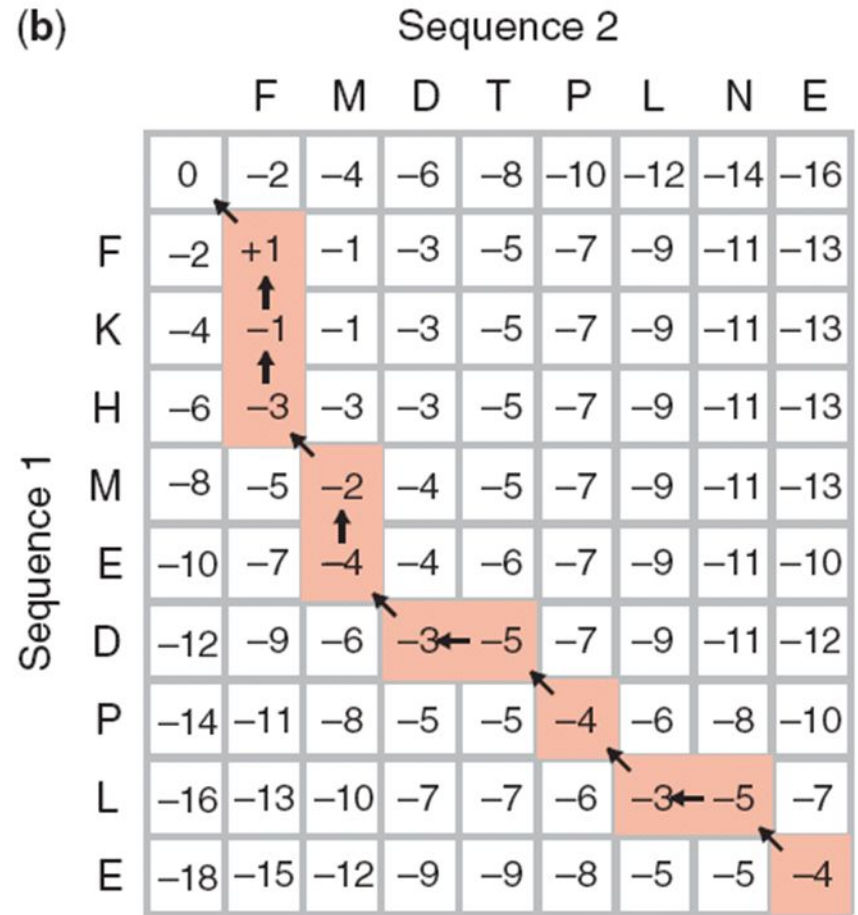
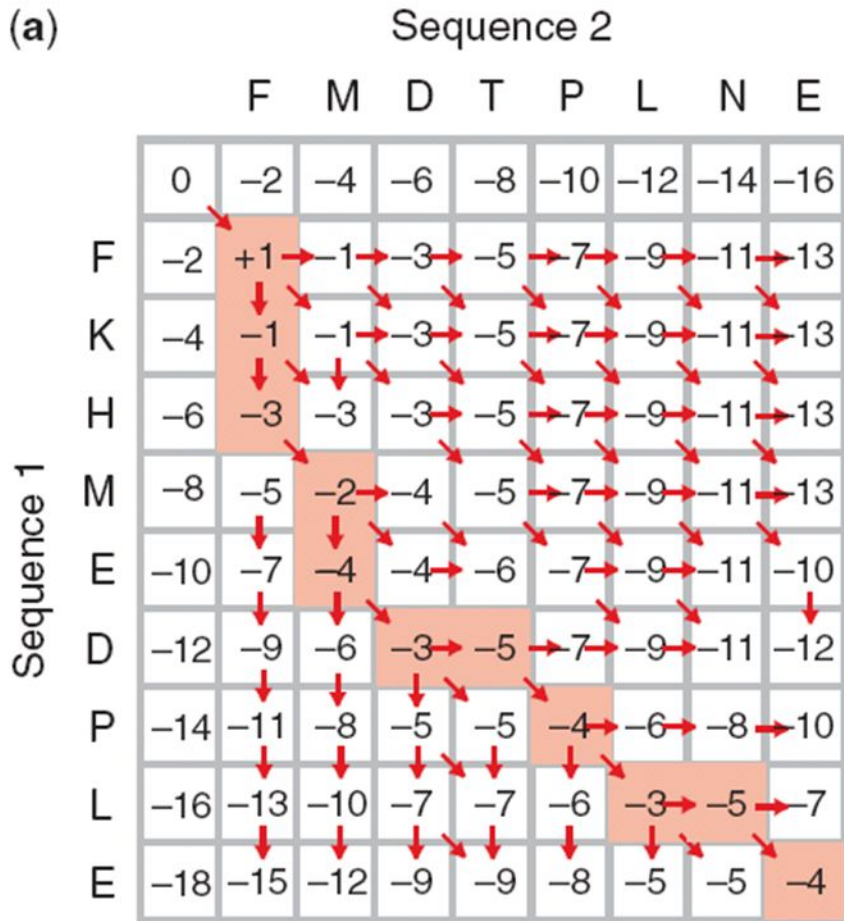
Заполнение матрицы с использованием «динамического программирования»



Заполнение матрицы с использованием «динамического программирования»



Нахождение оптимального (лучшего) попарного выравнивания



EMBOSS Pairwise Alignment Algorithms

This tool is used to compare 2 sequences. When you want an alignment that covers the whole length of both sequences, use needle. When you are trying to find the best region of similarity between two sequences, use water.

Method		Gap Open
EMBOSS::needle (global)		10.0
Gap Extend	Molecule	Matrix
0.5	Protein	Blosum62

Sequence 1: paste Sequence in any format OR upload a file:

Help

```
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLK
AFSDGLAHL DNLKGTFA TLSELHCDKLHVDPENFRLLGNVLCVLAHFFC
ALAHKYH
```

Seq. 1 Upload a file:

Browse...

Sequence 2: paste Sequence in any format OR upload a file:

Help

```
MVLSPADKTNVKA AWGKVG AHAGEYGA EALERMFLSFPTTKTYFPHFDLS
VAHVDDMPNALS ALSDLH AHKLRVDPVNF KLLSHCLLVTLAAHLPAEFTI
YR
```

Queries:
beta globin
(NP_000509)
alpha globin
(NP_000549)

```
#####
# Program: needle
# Rundate: Tue Aug 22 16:29:58 2006
# Align_format: srspair
# Report_file: /ebi/extserv/old-work/needle-20060822-16295743003385.output
#####
```

```
#-----
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:   90/149 (60.4%)
# Gaps:         9/149 ( 6.0%)
# Score: 292.5
#
#
#-----
```

```
EMBOSS_001      1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD      48
      || |:|:|:|.|.||| | :.|.|.|||.|:|:|.|.|.|.|.|.
EMBOSS_001      1 MV-LSPADKTNVKAAWGKVGAGHAGEYGAEALERMFLSFPTTKTYFPHF-D      48

EMBOSS_001     49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL DNLKGT FATLSELHCDKLH      98
      ||      .|:|:|.|.|||||.|.|.|.|.|.|.|.|.|.|.|.|.|.
EMBOSS_001     49 LS-----HGSAQVKGHGKKVADAL TNAVAHVDDMPNALSALSDLHAHKLR      93

EMBOSS_001     99 VDPENFRL LGNVLVCVLAH HFGKEFTPPVQAAYQKV VAGVANALAHKYH      147
      |||.||:|:|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
EMBOSS_001     94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLT SKYR      142
```

Глобальное vs. локальное выравнивания

- Глобальное выравнивание (Ниделмана-Вунша) проходит от одного конца каждой последовательности к другому концу.
- Локальное выравнивание находит регионы с оптимальным соответствием в двух последовательностях ("подпоследовательности").
- Локальное выравнивание почти всегда используется для поиска в базах данных, таких как BLAST. Оно полезно для поиска доменов (или ограниченных областей гомологии) внутри последовательностей.
- Смит и Уотерман (1981) решили проблему выполнения оптимального локального выравнивания последовательностей. Другие методы (BLAST, FASTA) быстрее, но менее тщательны.

Глобальные выравнивания (Верх) включает совпадения, игнорируемые локальным

выравниванием (ниже)

```

NP_824492.1      1  MCGDMTVHTVEYIRYRIPEQQSAEFLAAAYTRAAAQLAAAPQCVDYELARC
NP_337032.1      1
NP_824492.1     51  EEDFEHFVLRITWTSTEDHIEGFRKSELFPDFLAEIRPYISSIEEMRHYK
NP_337032.1      1
NP_824492.1    101  PTTVRGTGAAVFTLYAWAGGAEAFARLTEVFYKVLKDDVLA PVFEGMAP
      :.|.....|. |...|||:|. |.....||.:|.:|:|...|:  |
NP_337032.1      1  MEGMDQMPKSFYDAVGGAKTFDAIVSRFYAQVAEDEVLR RVY----P
NP_824492.1    151  EH-----AAHVALWLGEVFGGPAAYSETQGGHGHMVAKHLGKNITEVQRR
      |.      .....:|. |.:|:|...|||  |. |||.....:|...|:|:|. |.
NP_337032.1     44  EDDLGAEEERLRMFLEQYWGGPRTYSE-QRGHPRLRMRHAPFRISLIERD
NP_824492.1    196  RWVNLLQDAADDAGLPT-DAEFRSAFLAYAEWGT RLAVYFSGPDAVPPAE
      .|:.....|.....|  |.|.|...|. |.|.....|  :.|.
NP_337032.1     93  AWLRCMHTAVASIDSETLDDEHRRELLDYLEMAAHS LV--NSPF
NP_824492.1    245  QVPVQWSWGAMPPYQP      260
NP_337032.1    135      134
    
```

15% identity

```

NP_824492.1    113  TLYAWAGGAEAFARLTEVFYKVLKDDVLA PVFEGMAPEH-----AAHVA
      :.|...|||:|. |.....||.:|.:|:|...|:  ||.      .....
NP_337032.1     10  SFYDAVGGAKTFDAIVSRFYAQVAEDEVLR RVY----PEDDLGAEEERLR
NP_824492.1    158  LWLGEVFGGPAAYSETQGGHGHMVAKHLGKNITEVQRRRWVNLLQDAADD
      :|:.....:|...|||  |. |||.....:|...|:|:|. |.:.....|...
NP_337032.1     56  MFLEQYWGGPRTYSE-QRGHPRLRMRHAPFRISLIERDAWLRCMHTAVAS
NP_824492.1    208  AGLPT-DAEFRSAFLAYAE      225
      ....|  |.|.|...|. |.|
NP_337032.1    105  IDSETLDDEHRRELLDYLE      123
    
```

30% identity

NP_824492, NP_337032

Алгоритм локального выравнивания Смита-Уотермана

- Создание матрицы между двумя белками (размер $m + 1, n + 1$)
- Нет отрицательных значений в скоринговой матрице! $S > 0$
- Счет в каждой клетке максимальный из четырех значений:
 - [1] $s(i-1, j-1)$ + новая оценка $[i,j]$ (совпадение или несовпадение)
 - [2] $s(i,j-1)$ – gap penalty
 - [3] $s(i-1,j)$ – gap penalty
 - [4] 0

Алгоритм Смита-Уотермана позволяет выравнивать подпоследовательности

Sequence 1 (length m)

C A G C C U C G C U U A G

Sequence 2 (length n)

		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A		0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
A		0.0	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
U		0.0	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0
G		0.0	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7
C		0.0	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3
C		0.0	1.0	0.7	0.0	1.0	3.0	1.7	1.3	1.0	1.3	1.7	0.3	0.0
A		0.0	0.0	2.0	0.7	0.3	1.7	2.7	1.3	1.0	0.7	1.0	1.3	1.3
U		0.0	0.0	0.7	1.7	0.3	1.3	2.7	2.3	1.0	0.7	1.7	2.0	1.0
U		0.0	0.0	0.3	0.3	1.3	1.0	2.3	2.3	2.0	0.7	1.7	2.7	1.7
G		0.0	0.0	0.0	1.3	0.0	1.0	1.0	2.0	3.3	2.0	1.7	1.3	2.3
A		0.0	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	3.0	1.7	1.3	2.3
C		0.0	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	3.0	2.7	1.3	1.0
G		0.0	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0
G		0.0	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0

Compare Two Sequences:

<http://fasta.bioch.virginia.edu/>

[Statistical Significance from Shuffles](#)

[Find Internal Duplications](#)

Choose (A) program and (B, C) sequences to compare:

Queries:

beta globin (NP_000509)

alpha globin (NP_000549)

(A) Program:

Query post-trans modifications
"*@?#^~+=*" included for annotation

(B) Enter first (query) sequence: Subset range:

```
>gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]
MVHLTPEEKSAVTALWGKVVNDEVGGEALGRLLVYYPWTQRRFESFGDLSTPDVAVMGNPKVK.
AFSDGLAHLNDNLKGTFFATLSELHCDKLHVDPENFRLLGNVLCVLAHFGKEFTPPVQAAAYQI
ALAHKYH
```

[Entrez protein sequence browser](#)

[Entrez DNA sequence browser](#)

Protein DNA (both-strands) DNA (forward only) DNA (rev-comp only)

(C) Enter second (library) sequence:

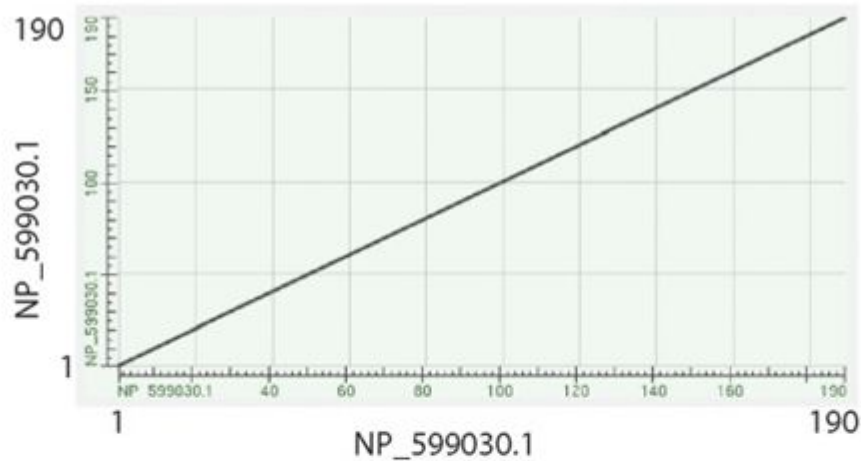
```
MVLSPADKTNVKAAMGKVGAGAGEYGAEALERMFSLFPTTKTYFPHFDLSHGSAQVKGHGKKV
VAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASV
YR
```

Other options:

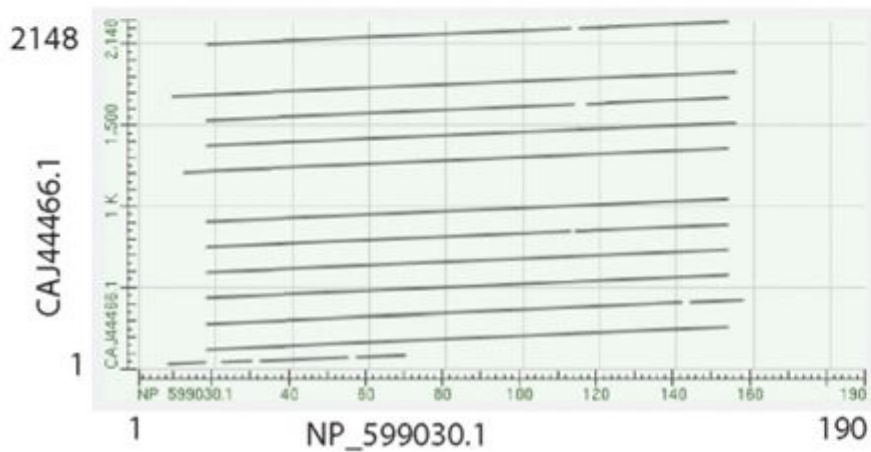
Ktup: Scoring matrix:

Dot matrix (Точечная матрица)

(a) Human cytoglobin compared to itself



(b) Cytoglobin compared to a snail globin (BLOSUM62)



(c) Cytoglobin compared to a snail globin (PAM250)

