

Выравнивание последовательностей

17 сентября 2025

Так выглядит выравнивание

DNAK_THEAC	82	KFKVFDKEFTPQQISAFILQKIKKDA-EAFLGEPVNEAVITVPAYFNDNQ	131
DNAK_PICTO	82	KYKIFGKEYTPQQISAFILQKIKRDA-EAFLGEPVTDAVITVPAYFNDNQ	131
HSCA_ACIF2	116	RLRTVAGEKSPVEVSAEILRVLKERAVETLGGEPEGAVITVPAYFDEAQR	165
HSCA_ACIF5	116	RLRTVAGEKSPVEVSAEILRVLKERAVETLGGEPEGAVITVPAYFDEAQR	165
DNAK_THEAC	132	QATKDAGTIAGFDVKRIINEPTAAALAYGVDKSGKSEKILVFDLGGGTLDV	182
DNAK_PICTO	132	QATKDAGAIAGLNVRRRIINEPTAACLAYGIDKLNQTLKIVIIYDLGGGTLDV	182
HSCA_ACIF2	166	QATKDAARLAGLNVLRLLAEPTAAAVAYGLDKGSEGI-FAIYDLGGGTFDI	215
HSCA_ACIF5	166	QATKDAARLAGLNVLRLLAEPTAAAVAYGLDKGSEGI-FAIYDLGGGTFDI	215
DNAK_THEAC	183	TIMDFGDGVFQVLSTSGDTRLGGTDMDEAIVNYIADD FQKKEGIDL RKDRS	233
DNAK_PICTO	183	TIMDFGQGVFQVLSTSGDTHLGGTDMDEAIVNFLADNFQRENGIDL RKDH	233
HSCA_ACIF2	216	SILRLQAGVFEVLATAGDSALGGDDMDHALAEWLMQE-EGGDASDPLW	262
HSCA_ACIF5	216	SILRLQAGVFEVLATAGDSALGGDDMDHALAEWLMQE-EGGDASDPLW	262
DNAK_THEAC	234	AYIRLRDAAEKAKIELSTTLSTDIDL PYITVTNSGPKHIKMTLTRAKLEEL	284
DNAK_PICTO	234	AYIRLRDAAEKAKIELSTVLETEINL PYITATQDGPKHLQYTLTRAKFEEL	284
HSCA_ACIF2	263	RRQVLQQ-ARTAKEALSAVAET-MIVLTPSGRAAREIKLSRGRLESL	307
HSCA_ACIF5	263	RRQVLQQ-ARTAKEALSAVAET-MIVLTPSGRAAREIKLSRGRLESL	307
DNAK_THEAC	285	ISPIVERVKGPIDKALEGAKLKKTEITKLLFVGGPTRIPYVRKYVEDYLG	335
DNAK_PICTO	285	IAPIVDRSKVPLDTALEGAKLKKGDIDKII LIGGPTRIPYVRKYVEDYFGR	335
HSCA_ACIF2	308	IQPVIQRS LPACRRALRDAGLKLDEIEGVVLVGGATRVPAVRAMVEEFFRQ	358
HSCA_ACIF5	308	IQPVIQRS LPACRRALRDAGLKLDEIEGVVLVGGATRVPAVRAMVEEFFRQ	358

Секвенирование миллиардов последовательностей делается главным образом ради выравниваний

Пример из интернета: выравнивание в жизни (спецслужб?)

Text1: The caller ident--ified the bomber

Text2: The caller ---n-am---ed the bomber

Text1: as ----- Yussef Attala, 20,

Text2: as 20-year old Yussef Attala- ---

Text1: from the Balata refugee camp near

Text2: from the Balata refugee camp near

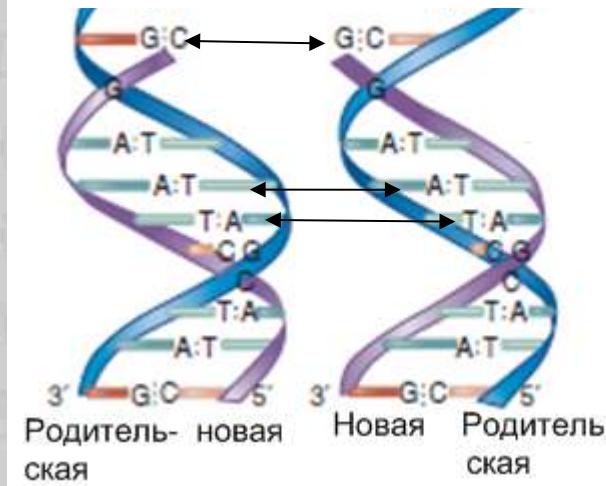
Text1: Nablus

Text2: Nablus

Бессмысленное
выравнивание

В выравнивании есть содержательные колонки и
бессмысленные

Гомология нуклеотидов ДНК



При репликации ДНК большинство нуклеотидов потомка “знают” своего предка в геноме родителя

Иногда из-за ошибок репликации ДНК (или после деления клетки) в ДНК новой клетки появляются небольшие локальные изменения по сравнению с ДНК родительской клетки.

При замене нуклеотида на другой можно указать какой именно нуклеотид родителя изменился!

Гомологичные нуклеотиды двух потомков – те, которые произошли от того же самого нуклеотида общего предка

Гомологичные нуклеотиды разных последовательностей
располагают в одной колонке выравнивания

**ЛОКАЛЬНУЮ ЭВОЛЮЦИЮ ПОСЛЕДОВАТЕЛЬНОСТИ
МОЖНО ОТОБРАЗИТЬ ВЫРАВНИВАНИЕМ**

Правильное выравнивание последовательностей ДНК живущих сегодня организмов

Гомологичные нуклеотиды ставим друг под другом

ПРЕДОК

сын

внук

правнук

праправнук-1

праправнук-2

.....

живет сейчас

живет сейчас

живет сейчас

1. TAT - - GCGAAT - GCCCTGAA

2. TAT - - GC**A**AAT - GCCCTGAA замена

3. TAT - - GCAAAT - GCT**T**CTGAA замена

4. TAT - - GCAAAT**C**GCTC**G**GAA вставка и замена

5. TAT - - GCAAA**G**CGCTCGGAA замена

6. TAT - - GCAAA - CGCTCGGAA делеция

.....

a. TAT - - GCAT**A** - CGC - - - GAA дел. 3, зам.1.

b. TAT**A**TGCAAAGCGCTCGGAA вставка 2 п.н.

c. TAT - - GCAAA - - GC**G**CT**T**GAA дел. 1, зам. 2

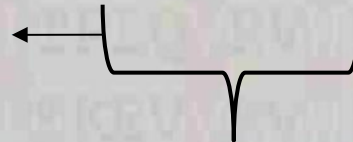
Задача построения правильного выравнивания сложна и неоднозначно решается

Вот правильное выравнивание с пред. слайда

а. TAT - - GCATA - CGC - - - GAA

б. TATATGCAAAGCGCTCGGAA

с. TAT - - GCAAA - - GCCTGAA



Программа выравнивания ориентируется на сходство.
Сдвиг даст больше совпадений б. и с.

ПРОГРАММЫ ВЫРАВНИВАНИЯ МОГУТ ОШИБАТЬСЯ!

Эволюция последовательности белка —
следствие эволюции кодирующей
последовательности

Последовательности большинства белков
находится под стабилизирующим отбором —
против изменений

Выравнивание белков

- Теория: аминокислотные остатки белков гомологичны, если их кодоны гомологичны. В выравнивании гомологичные остатки располагают друг под другом
- Практика: гомологичность белков, участков и отдельных остатков выводят из выравнивания, построенного программой на основе сходства букв

- **Выравнивать последовательности генов или белков?**

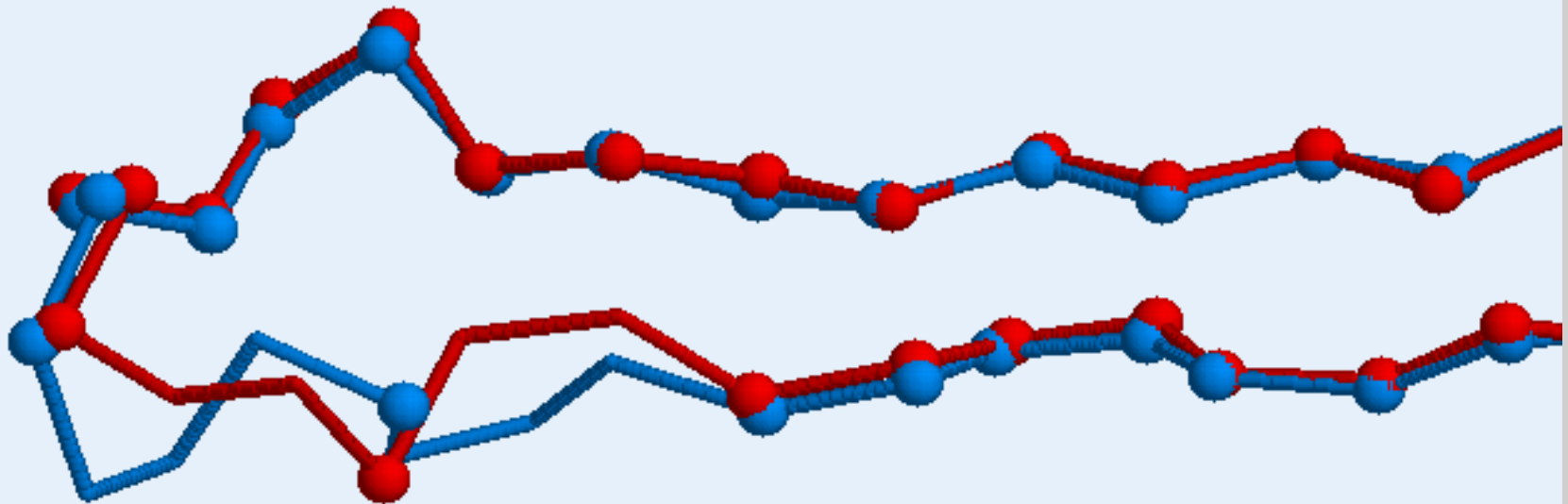
БЕЛКОВ:

- в нуклеотидном выравнивании программа может нарушить кодонную структуру
- в нуклеотидном выравнивании больше мутаций, чем в белковом, из-за синонимичных замен
- белковое выравнивание учитывает сходство свойств аминокислот

Можно ли проверить правильность выравнивания?

- Программа может построить неправильное выравнивание. Есть ли способы независимой проверки?
- ДНК: нет способов (не считая генноинженерного мутагенеза в лаборатории)
- Белки: можно! – если известна 3D структура белка.
 - Если при совмещении полипептидных цепей хорошо совмещаются C α атомы, то такие остатки можно считать гомологичными
 - Ограничение: Структуры известны для 80 000 белков, а последовательности – для 250 000 000
 - Есть базы эталонных выравниваний, построенных по совмещению структур (Valibase и др.). Их используют для сравнения программ выравнивания
- РНК: частично да. Тоже можно учитывать 3D структуру

Белки: совмещение структур и выравнивание последовательностей



		*	20	*	
Seq_A	:	LTGYGRWEAEFagnkae	--sdt	aqgKTrlAFAGLK	: 33
Seq_B	:	LTGYGQWEYNFqgnnsegadaqtgn	KTrlAFAGLK	:	35
Aligned	:	AAAAAAAAAAAAAAAAAAAA	-----	AAAAAAAAAAAA	: 27

Все программы ищут сходство,
а сходство — не то же, что гомология!

ПРОГРАММЫ ВЫРАВНИВАНИЯ

Что нужно знать об алгоритме выравнивания **двух** последовательностей?

- Матрица весов замен
- Штраф за открытие “гэпа”
- Штраф за продолжение “гэпа”
- Вес выравнивания
- Оптимальное выравнивание – выравнивание с максимальным весом

Программа успешно выровняет любые две последовательности, даже не гомологичные!

Упражнение: вычислите вес выравнивания

PDB:1osm : **KIDGLHYFS**D--Dkd : 28
 PDB:1hxx : **KAVGLHYFS**KgnGen : 30

A R N D C Q E G H I ...

A 4 -1 -2 -2 0 -1 -1 0 -2 -1 -1

gap open -6

R -1 5 0 -2 -3 1 0 -2 0 -3 -2

gap extension -2

N -2 0 6 1 -3 0 0 0 1 -3 -3

D -2 -2 1 6 -3 0 2 -1 -1 -3 -4

C 0 -3 -3 -3 9 -3 -4 -3 -3 -1 -1

Q -1 1 0 0 -3 5 2 -2 0 -3 -2

E -1 0 0 2 -4 2 5 -2 0 -3 -3

G 0 -2 0 -1 -3 -2 -2 6 -2 -4 -4

H -2 0 1 -1 -3 0 0 -2 8 -3 -3

I 1 3 3 3 1 3 3 4 3 4 2

Штраф за продолжение
 маленький, т.к. делеция
 нескольких кодонов
 может произойти как
 одно событие

Матрица BLOSUM62

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

Парное выравнивание

- Локальное (алгоритм *Smith – Waterman*) – находит наиболее сходные участки двух последовательностей.
- Глобальное (*Needleman – Wunsch*) – по всей длине последовательностей: **пригодно только для последовательностей, гомологичных по всей длине!**

Выравнивание последовательностей

		Номер столбца выравнивания																																
		* 20																																
MTA1_YEAST	:	----	K	S	S	I	S	P	Q	A	R	A	F	L	E	Q	V	F	R	R	K	---	Q	S	L	N	S	:	24					
MAT2_YEAST	:	K	P	Y	R	G	H	R	F	T	K	E	N	V	R	I	L	E	S	W	F	A	K	N	I	E	N	P	Y	L	D	T	:	31
		3 2 LE F 4 L13																																
		40 * 60																																
MTA1_YEAST	:	K	E	K	E	V	A	K	K	C	G	I	T	P	L	Q	V	R	V	W	F	I	N	K	R	M	R	S	K	-	:	53		
MAT2_YEAST	:	K	G	L	E	N	L	M	K	N	T	S	L	S	R	I	Q	I	K	N	W	V	S	N	R	R	R	K	E	K	T	:	61	
		K E 6 K 63 6Q64 W N4R 4 K																																

Название последовательности

Консервативный остаток

Функционально консервативная позиция

Номер последнего в строке остатка ИЗ ЭТОЙ ПОСЛЕДОВАТЕЛЬНОСТИ

Множественное выравнивание

- Как правило, глобальное; программы
 - Muscle
 - MAFFT
 - Clustal Omega
 - T-coffee
 -
- Локальное – находит лучшие т.н. “мотивы” – короткие похожие участки
 - MEME
 -

Примеры выравниваний

РНК-зависимые РНК полимеразы пикорнавирусов

	20	*	340	*	360	*	
POL1_BAYMG :	LIYDEI	LNTT	ICLANGMVIRKNV	GNN	SGQPST	TVVD	NTLVLMTAFLYAYIHKTG : 331
POLN_SOUV3 :	VVAQDLLAP	SEMDVG	DYVIRVKEGLP	SGFPC	TSQVNS	INH	WLITLCALSEVTG : 321
POLN_LORDV :	IVAEDLLSP	SVMDVG	DFKISINEGLP	SGVPC	TSQWNS	IAHWLL	TLCALSEVTN : 321
POLN_SMSV1 :	SACATLKSN	PIGIFNGVAFKVAGGLP	SGMPL	TSI	INS	LNHCL	LMVGSAAVKALE : 318
POLG_HPAV2 :	TALINT	IIYSK	HLLYNCCYHVC	GSMPS	SGSPCT	TALLNS	IINNVLNLYYVFSKIFG : 329
POLG_TMEVB :	AEYLRSLAV	SRHAYEDRRVLIRGGLP	SGCAAT	SM	LN	TI	MNNVIRAALYLTYS : 319
POLG_CXA9/ :	TNYIDYLC	NSHHLYRDKHYFVRGGMP	SGCSGT	TS	IFNS	MINN	IIIRTLMLKVYK : 315
POLG_CXA21 :	VDYIDYLN	SHHLYKNKTYCVKGGMP	SGCSGT	TS	IFNS	MINN	LIIRTLMLLRITYK : 314
POL1_GFLV/ :	KNLLLA	ICGRLSICGNQVYATEAGIP	SGCAL	T	V	LN	SIFKELLMRYCFKKIVP : 346

g pSG T N36

Фрагменты геномов двух видов бруцелл

	*	20	*	40	
AE008917_1899844_1899896 :	tttag	aaaattccagagcgggttcctggttaaa	aacg	gaatcggttg	: 42
AE008917_686288_686337 :	tttag	---ttctggagcgggttcctggttt	aac	gaatcggttg	: 39

	*	
AE008917_1899844_1899896 :	gagccgctcta	: 53
AE008917_686288_686337 :	gaaccgctcta	: 50

Пример: выравнивание ROU-белков. Блоки достоверного выравнивания. Домены.

```

*          420          *          440          *          460          *          480          *          500          *          520          *
BRN1_RAT   : VHPGLVRGDTPELAENHHHHHHHNAHPHPHPHHAQCPPHHGCGG--AGPGLNSHDPHSDEDTPSTDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--VFSQTTICRFEAQLSFKNNCKLPLLNKWL : 382
ZP12_BRARE : VHPGLMRGESPEMDHHHHHHHHHQQHPPHHHHQHHA-----GVNSHDSHSDDEDTPSTDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--VFSQTTICRFEAQLSFKNNCKLPLLNKWL : 312
OC3N_MOUSE : GLHHHGLRDAHDEPH-----HADHHHPHPS-HPHQPPPPPPPPQPGPPCHPGAHHDPHSDDEDTPSTDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--VFSQTTICRFEAQLSFKNNCKLPLLNKWL : 335
PO3B_XENLA  : GMINPGSCQGMHHHC-----LRSDHDDHHGDHGHQPPSQTQQQQQHSQLQCGHHHSDDEDTPSTDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--VFSQTTICRFEAQLSFKNNCKLPLLNKWL : 275
BRN4_MOUSE : NVYSQPGFTVSCMLE-----HGCLTPPPAAASTQSLHPVLRPPDPHGLCGSHHCQDHSDEETPTSDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--VFSQTTICRFEAQLSFKNNCKLPLLNKWL : 257
POU2_XENLA  : ~~~~~~MLD-----HGELTPPLPAGTTQSLHPVLRPNHDVLDGSHHCQDHSDEETPTSDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--VFSQTTICRFEAQLSFKNNCKLPLLNKWL : 114
POU1_XENLA  : HQPLIYSQSSYTNLNGMLGPQASSLHSHMRDPLHDDPGVHDTPQVSPQHLGHHQDHSDEAPSSDDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--VFSQTTICRFEAQLSFKNNCKLPLLNKWL : 246
ZP50_BRARE : HQSLVYSQTGYT-----MLSPQPS-LHHGLRDLHDDAGSHDNQMESPPQPFSSHQDHSDEAPSSDDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--VFSQTTICRFEAQLSFKNNCKLPLLNKWL : 258
OCT6_RAT    : HHALHEDGHEAQLPSPPPHLGAHGHAHGAHAGGLHAAAAHLHPGAGCGGSSVGEHSDAPSSDDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--VFSQTTICRFEAQLSFKNNCKLPLLNKWL : 318
CF1A_DROME  : PLQYHAMNMGMLHHPAHAVAAAHQSVAPLHHTLRGESPOLHHHHMGCGDRDAISGCEEDTPSTDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--VFSQTTICRFEAQLSFKNNCKLPLLNKWL : 322
SGF3_BOMMO  : LQHAAGHWAPVVSPHYGAARPSHCMEDTQCPCTNTICSETSSPRDPLHHHMERDQPEEDTPSTDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--VFSQTTICRFEAQLSFKNNCKLPLLNKWL : 220
POU1_DUGJA  : TKPTHNLNMDIN-----LCRFPQNBIIIMPDLINAPSPTLTNN-ILFYSCGDHRSCKTEDLPSSDDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--VFSQTTICRFEAQLSFKNNCKLPLLNKWL : 330
HM06_CABEL  : TSVVTSTPSCRIQEVEPPIIQLRMPWPYPAYQFSCDSGSVSGAGGPHQLPLSDSDSEQTCDDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--VFSQTTICRFEAQLSFKNNCKLPLLNKWL : 203
OCT1_HUMAN  : -QLPQQSQANLQSQ-PSITLTSQPATPTRTIAAT-PIQTLPQ--SQSTPKRIDTPS-LEPSDLLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--DFSQTTISRFEANLSFKNNCKLPLLNKWL : 351
OCT2_MOUSE  : GLLPTPNLFLPQQT-QGALLTSQPRACLTQ-----P--PKCLEPPSHPEEPSDLLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--DFSQTTISRFEANLSFKNNCKLPLLNKWL : 249
SK1A_RAT    : NLLSFPQQSTLLLP-QTGPCLTSQAVGRPCLSGSSLEPHLEASQHLPGPKHLPGPCGND EPTDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--DFSQTTISRFEANLSFKNNCKLPLLNKWL : 247
OC11_MOUSE  : NLLSFPQQSTLLLP-QTGPCLTSQAVGRPCLSGSSLEPHLEASQHLPGPKHLPGPCGND EPTDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--DFSQTTISRFEANLSFKNNCKLPLLNKWL : 247
SK1I_RAT    : NLLSFPQQSTLLLP-QTGPCLTSQAVGRPCLSGSSLEPHLEASQHLPGPKHLPGPCGND EPTDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--DFSQTTISRFEANLSFKNNCKLPLLNKWL : 165
PDM2_DROME  : TANSLGMSAVMTPN-TPSMQQQPQ-----LQSTPKP-TSCLTVASAMA-KLEQSPETDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--DFSQTTISRFEANLSFKNNCKLPLLNKWL : 357
PDM1_DROME  : HPLQITPSSAASLR-LSGMLTPTSTPSC-TQMSQGTTPQPKTVASAAAAAAGPSPETDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--DFSQTTISRFEANLSFKNNCKLPLLNKWL : 492
PIT1_MOUSE  : FPDHT-----LSHGFPLHQLPAAEDPAASEFK-----QELARKSKLVEPIDMDSPEI--RELEQAKQKQKQRIKLCFTQADVCLADGTLYGN--EFSQTTICRFEANLSFKNNCKLPLLNKWL : 195
PIT1_CHICK  : FPEHALSAS-SCALGHSFTPMHQTLLSDDPTASDFK-----QEFRRKSKSVEEPVDMDSPEI--RELEQAKQKQKQRIKLCFTQADVCLADGTLYGN--EFSQTTICRFEANLSFKNNCKLPLLNKWL : 266
PIT1_MELGA  : FPEHALSAS-SCALGHSFTPMHQTLLSDDPTAADF--QEFRRKSKSVEEPVDMDSPEI--RELEQAKQKQKQRIKLCFTQADVCLADGTLYGN--EFSQTTICRFEANLSFKNNCKLPLLNKWL : 273
PO61_MOUSE  : PIQTPQAVPQPAVILTSPTPALKPSAATPIITCSETP---TVSQLVSKPHTPSLDEDGINL--EETREAKMKIRLRLSGLTQTQVQQAATATEGP---AYSQSAICRFEADITPKSAQKLPVLEKWL : 210
POUC_BRARE  : KVAPSKVIIAPQPSVVKPVTSLTAAC----VIACGEMP---TVGLVNKP-SAVKDEEAINL--EETREAKMKIRLRLSGLTQTQVQQAATATEGP---AYSQSAICRFEADITPKSAQKLPVLEKWL : 519
BR3A_HUMAN  : N--MPSC-LPHPLVAAAAH-HCAAAAAAQAQVAAAAAAGVVG--AAGLASICDSDTDP--RELEQAKQKQKQRIKLCFTQADVCLADGTLYGN--EFSQTTICRFEANLSFKNNCKLPLLNKWL : 338
BR3A_MOUSE  : N--MPSC-LPHPLVAAAAH-HCAAAAAAQAQVAAAAAAGVVG--AAGLASICDSDTDP--RELEQAKQKQKQRIKLCFTQADVCLADGTLYGN--EFSQTTICRFEANLSFKNNCKLPLLNKWL : 336
BRN3_CHICK  : NSALPLGCVPCPDVCGSTPSHPHSHMSAINHMAHNSQPMNMMSHPHCLA--SHAVISGPETETDP--RELEQAKQKQKQRIKLCFTQADVCLADGTLYGN--EFSQTTICRFEANLSFKNNCKLPLLNKWL : 256
BR3C_MOUSE  : L--SVSG-LCAPEHSVMPAQIHPHHLGAMCHLHQAAGMSHPHVAAPH--SAMPACLSDVESDP--RELEQAKQKQKQRIKLCFTQADVCLADGTLYGN--EFSQTTICRFEANLSFKNNCKLPLLNKWL : 253
BR3C_HUMAN  : L--SVSG-LCAPEHSVMPAQIHPHHLGAMCHLHQAAGMSHPHVAAPH--SAMPACLSDVESDP--RELEQAKQKQKQRIKLCFTQADVCLADGTLYGN--EFSQTTICRFEANLSFKNNCKLPLLNKWL : 253
BR3B_MOUSE  : LEHLSPLGALGAMAGPDGTVVSTPAHAPHMATMNMHQAALSMAHAGCLPSHMGCMDSVDADP--RDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--EFSQTTICRFEANLSFKNNCKLPLLNKWL : 326
BR3B_HUMAN  : LEHLSPLGALGAMAGPDGTVVSTPAHAPHMATMNMHQAALSMAHAGCLPSHMGCMDSVDADP--RDLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--EFSQTTICRFEANLSFKNNCKLPLLNKWL : 325
IPOU_DROME  : PLTPHQLHLCYSYHSMNHHHHPCTLSGHTCGHHGHSVAVHPVITAAVAAAGLHPDPTDP--RELEQAKQKQKQRIKLCFTQADVCLADGTLYGN--EFSQTTICRFEANLSFKNNCKLPLLNKWL : 296
UN86_CABEL  : VPTSSDDQMTPLQVQMAMQSYGAPPPFYQYMMTHFPSTTSSIASSNNLARYPIAPPTSDMDTDP--RQLEQAKQKQKQRIKLCFTQADVCLADGTLYGN--EFSQTTICRFEANLSFKNNCKLPLLNKWL : 339
OC3A_HUMAN  : QVCGVLVPCQGLLEQPEQEGACVGVESNSDGASPECTVTPGAVKLEKRLQNPNESQIKALQKLEQAKLKKQRITLCYTTQADVCLTGVFGK---VFSQTTICRFEANLSFKNNCKLPLLNKWL : 209
OCT3_MOUSE  : QVGLCVLPQGVETLQPEQEGACVGVESNSG-TSEPCADRPNAVAKLEKVEPT-PRESQDMKALQKLEQAKLKKQRITLCYTTQADVCLTGVFGK---VFSQTTICRFEANLSFKNNCKLPLLNKWL : 202
OC3B_HUMAN  : ~~~~~~MHFYRLFGLATRRFLNPEWKGEIDNWCYVYVLTSLLPFKIQSQDIKALQKLEQAKLKKQRITLCYTTQADVCLTGVFGK---VFSQTTICRFEANLSFKNNCKLPLLNKWL : 114
SPM1_RAT    : CPGPVEWGVPLGPSYFYGCIAPYCAIYETRTWQSNSEDYTPGYIALRYMNLALPEDVSAIQKMEQLAKLKRQRMTCYTTQADVCLTGVFGK---VFSQTTICRFEANLSFKNNCKLPLLNKWL : 184
HM18_CABEL  : IAPSPILLQSLMHTPTQPTASLTPKKAENRPVVSQTLKASKRRFLDFTSRIEAASMSGDDRIMNLEQAKLKKQRITLCYTTQADVCLTGVFGK---DFSQTTISRFEANLSFKNNCKLPLLNKWL : 361
POU2_BRARE  : APPTPSASSPSLSPSPGNGFCSPGFFSGCTAQNIPSAQAQAPRSSGSSSGCCNSSEEBETLTEDLEQAKLKKQRITLCYTTQADVCLTGVFGK---EFSQTTICRFEANLSFKNNCKLPLLNKWL : 320

```

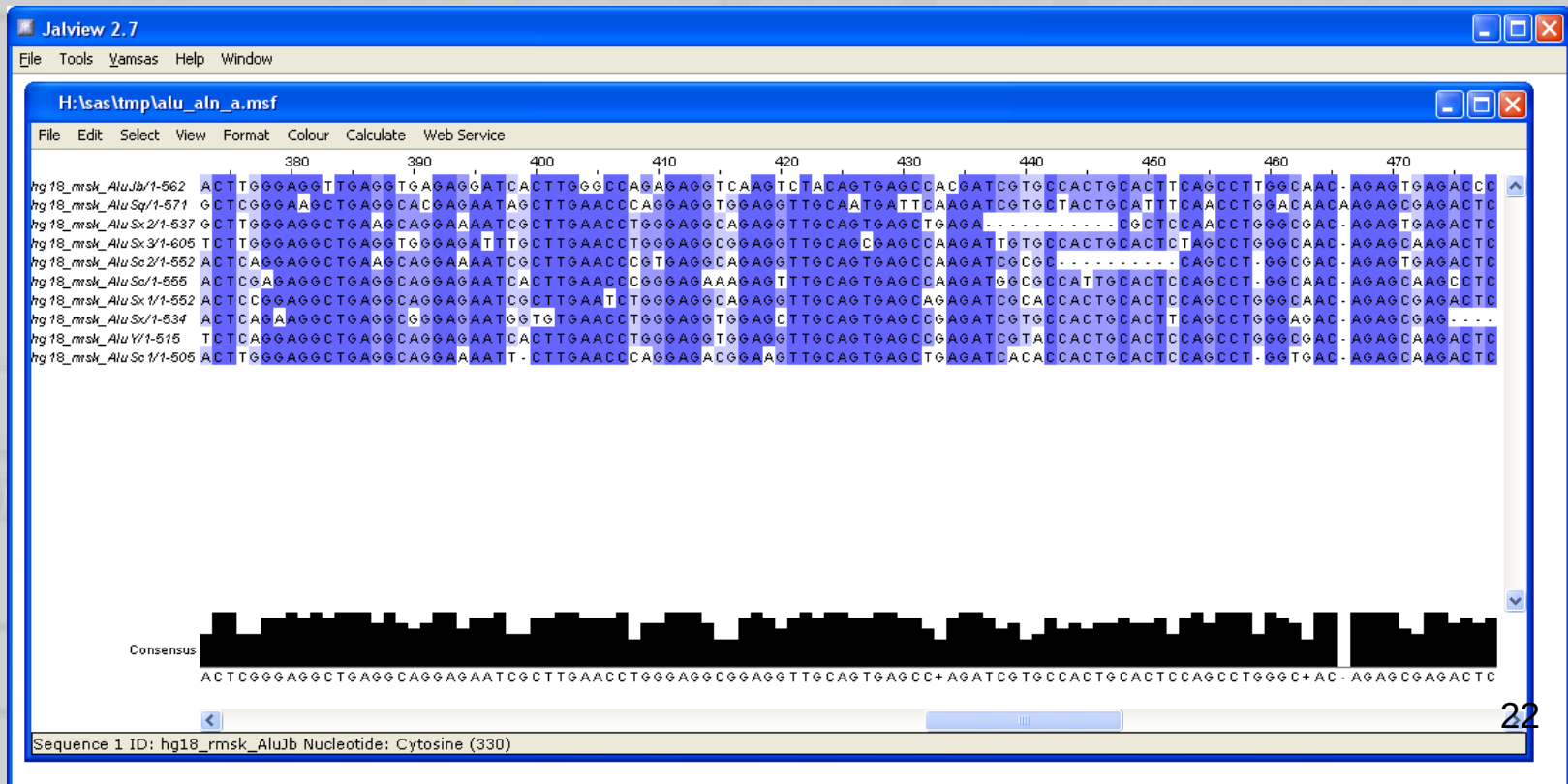
Продолжение

	*	540	*	560	*	580	*	600	12345*	10	616	20	6	30	640	8	0	*78	51	460
BRN1_RAT	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGALESHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGIQ	:	467							
ZP12_BRARE	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGALESHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGVP	:	397							
OC3N_MOUSE	:	PLLNKQLEEDDSSSGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGALESHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGGT	:	420							
PO3B_XENLA	:	PLLNKQLEEDDSSSGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGALESHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGGT	:	360							
BRN4_MOUSE	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDQ	:	342							
POU2_XENLA	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
POU1_XENLA	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
ZP50_BRARE	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
OCT6_RAT	:	PLLNKQLEEDDSSSGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
CF1A_DROME	:	PLLNKQLEEDDSSSGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
SGF3_BOMMO	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
POU1_DUGJA	:	PLLNKQLEEDDSSSGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
HM06_CAEL	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
OCT1_HUMAN	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
OCT2_MOUSE	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
SK1A_RAT	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
OCT1_MOUSE	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
SK1I_RAT	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
PDM2_DROME	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
PDM1_DROME	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
PIT1_MOUSE	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
PIT1_CHICK	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
PIT1_MELCA	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
PO61_MOUSE	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
POUC_BRARE	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
BR3A_HUMAN	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
BR3A_MOUSE	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
BRN3_CHICK	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
BR3C_MOUSE	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
BR3C_HUMAN	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
BR3B_MOUSE	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
BR3B_HUMAN	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
IPOU_DROME	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
UN86_CAEL	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
OC3A_HUMAN	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
OCT3_MOUSE	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
OC3B_HUMAN	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
SPM1_RAT	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
HM18_CAEL	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							
POU2_BRARE	:	PLLNKQLEEDDSTGSP-TSIDKIAAQ							CRKKRKRTSIEVSVRGVLETHHLKCPKPSAQEITSLADSLQ	EREVVRVVF	CNRQKQKRMTPPGDP	:	199							

Биологически осмысленное выравнивание может быть в одной части
выданного программой выравнивания и не быть в другой!

Множественное выравнивание

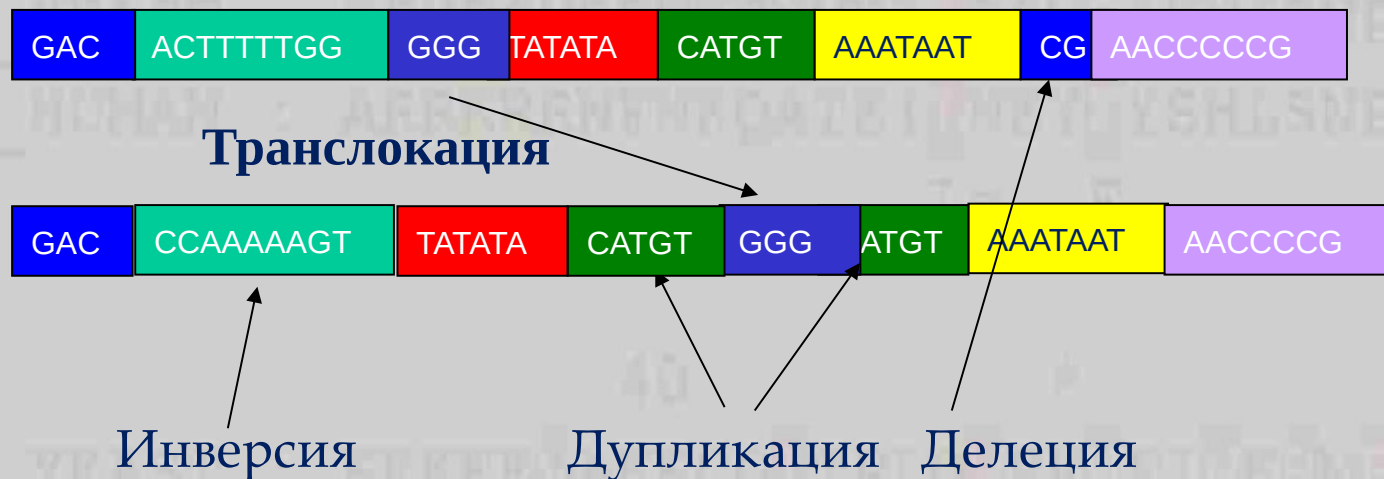
- Программа JalView <https://www.jalview.org/> позволяет визуализировать выравнивание. Она же может послать ваши последовательности на один из серверов, делающих выравнивания.



Кроме точечных изменений редко, но происходят крупные перестройки ДНК (или РНК – для РНК вирусов)

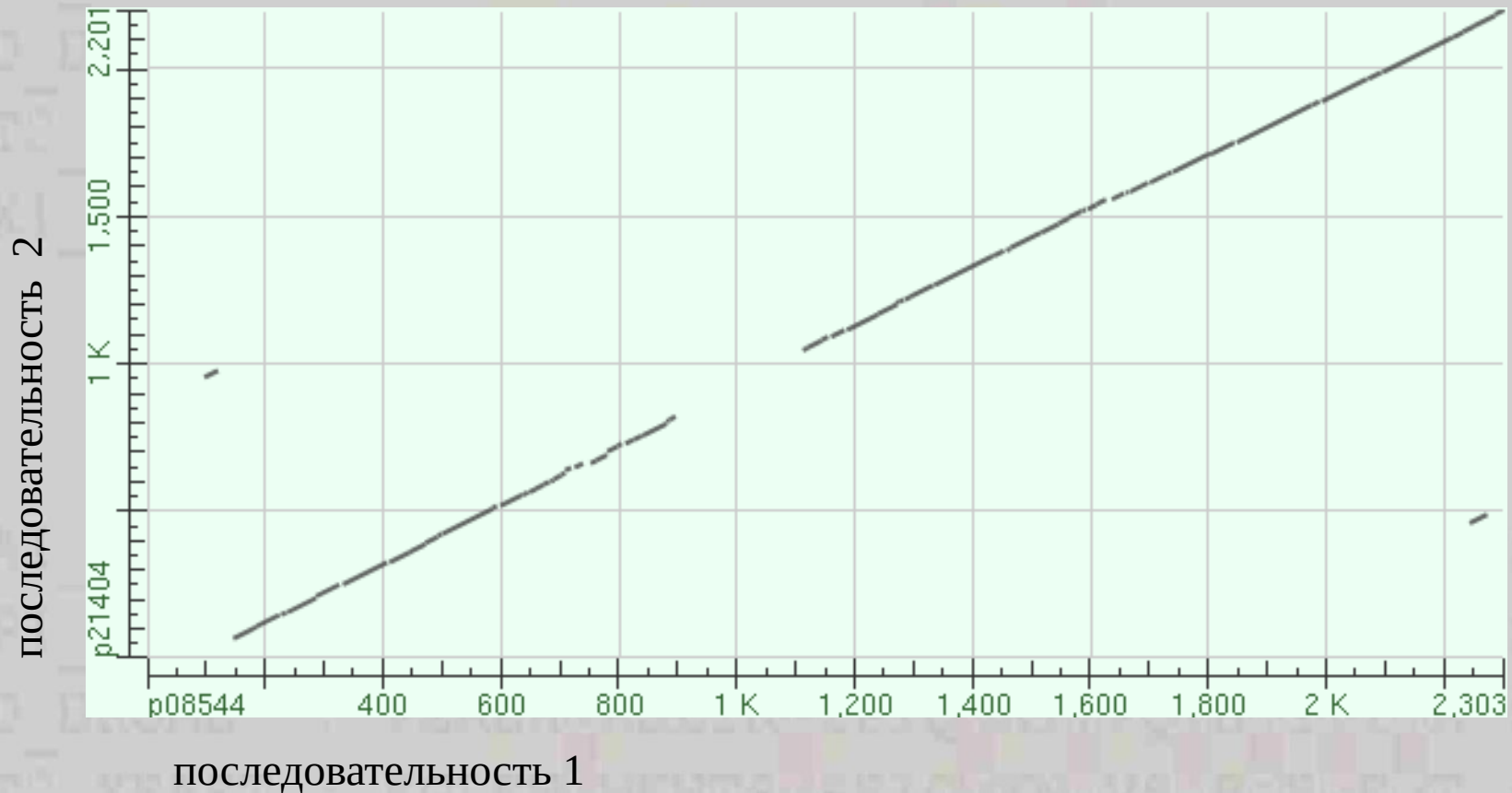
ГЛОБАЛЬНЫЕ ПЕРЕСТРОЙКИ ГЕНОМОВ

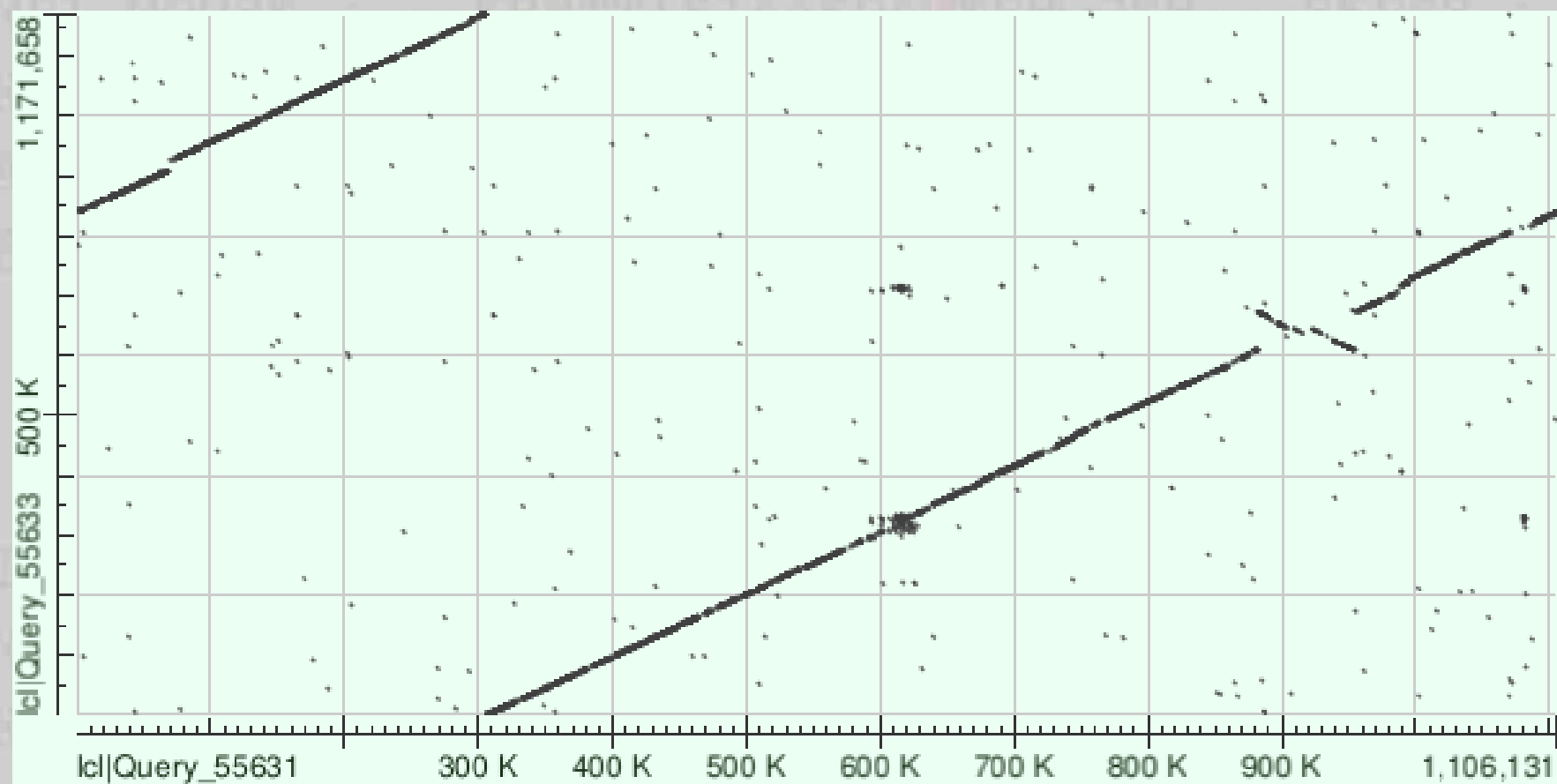
Крупные перестройки ДНК



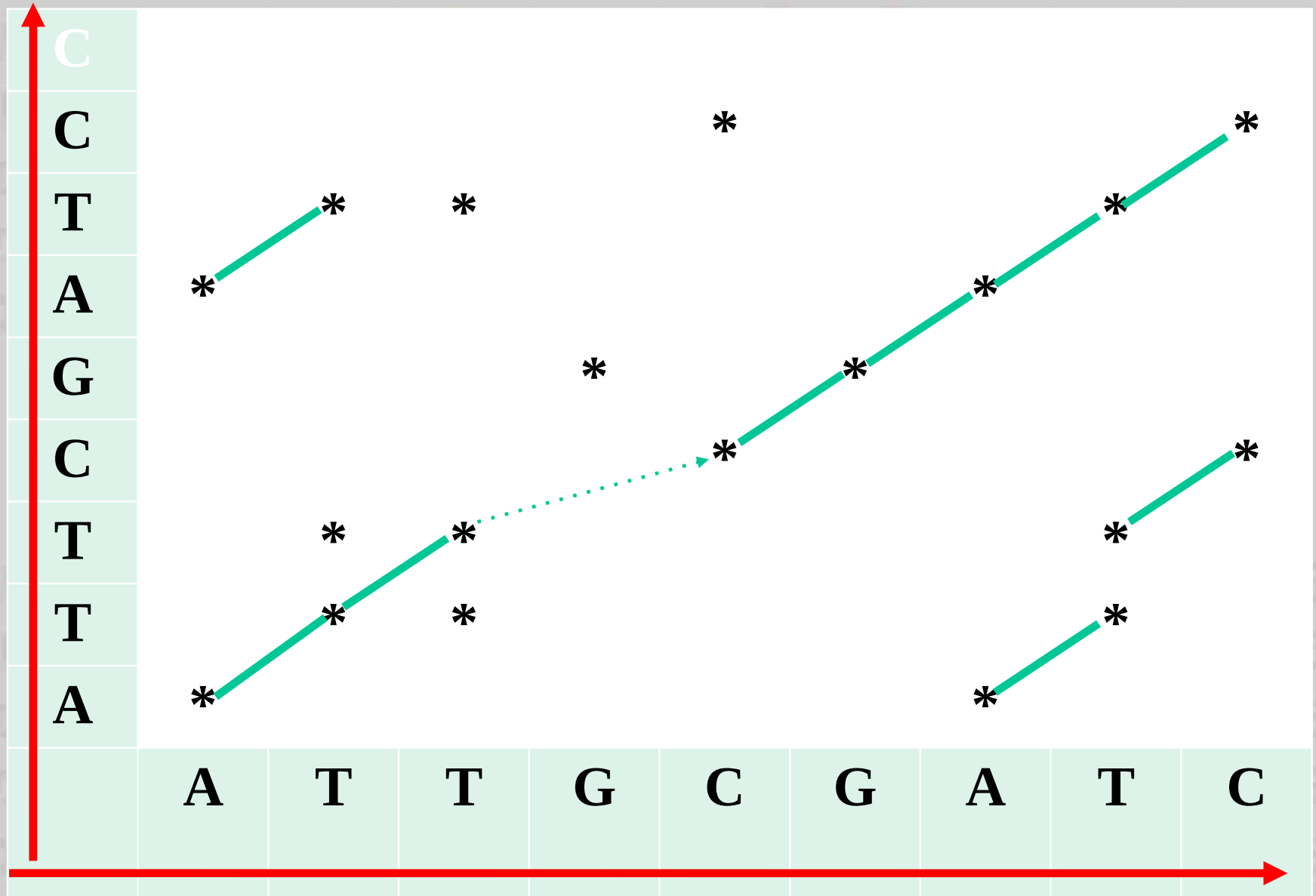
Участки могут состоять из сотен, тысяч и миллионов пар нуклеотидов

Карта локального сходства позволяет описывать крупные перестройки геномов

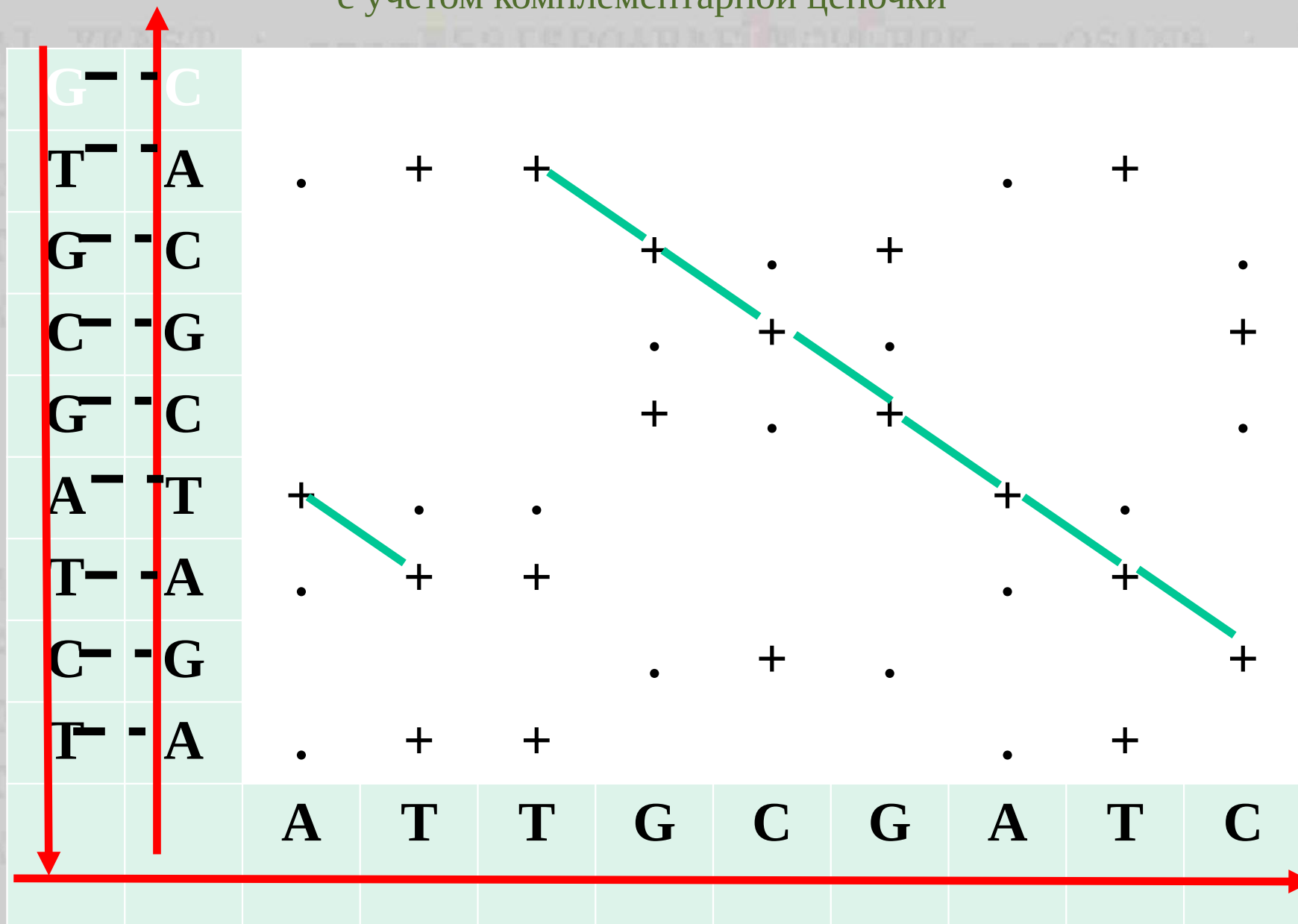




Карта сходства двух последовательностей

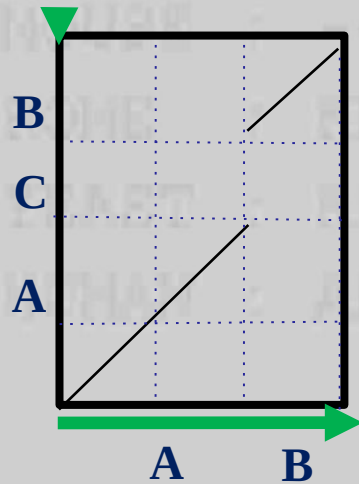


Карта сходства двух последовательностей с учетом комплементарной цепочки

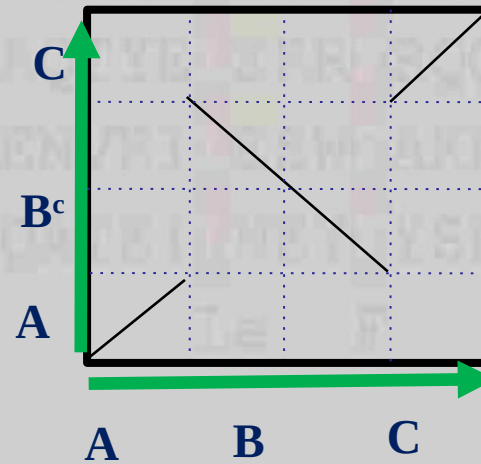


Крупные эволюционные события на карте локального сходства.

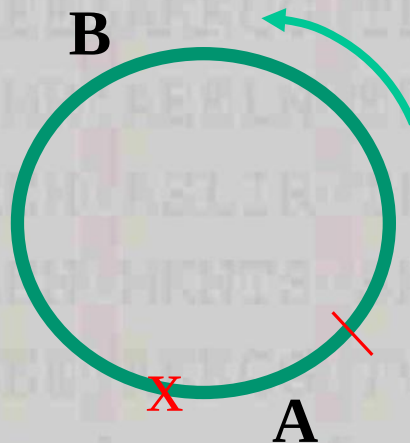
На примере двух бактериальных хромосом



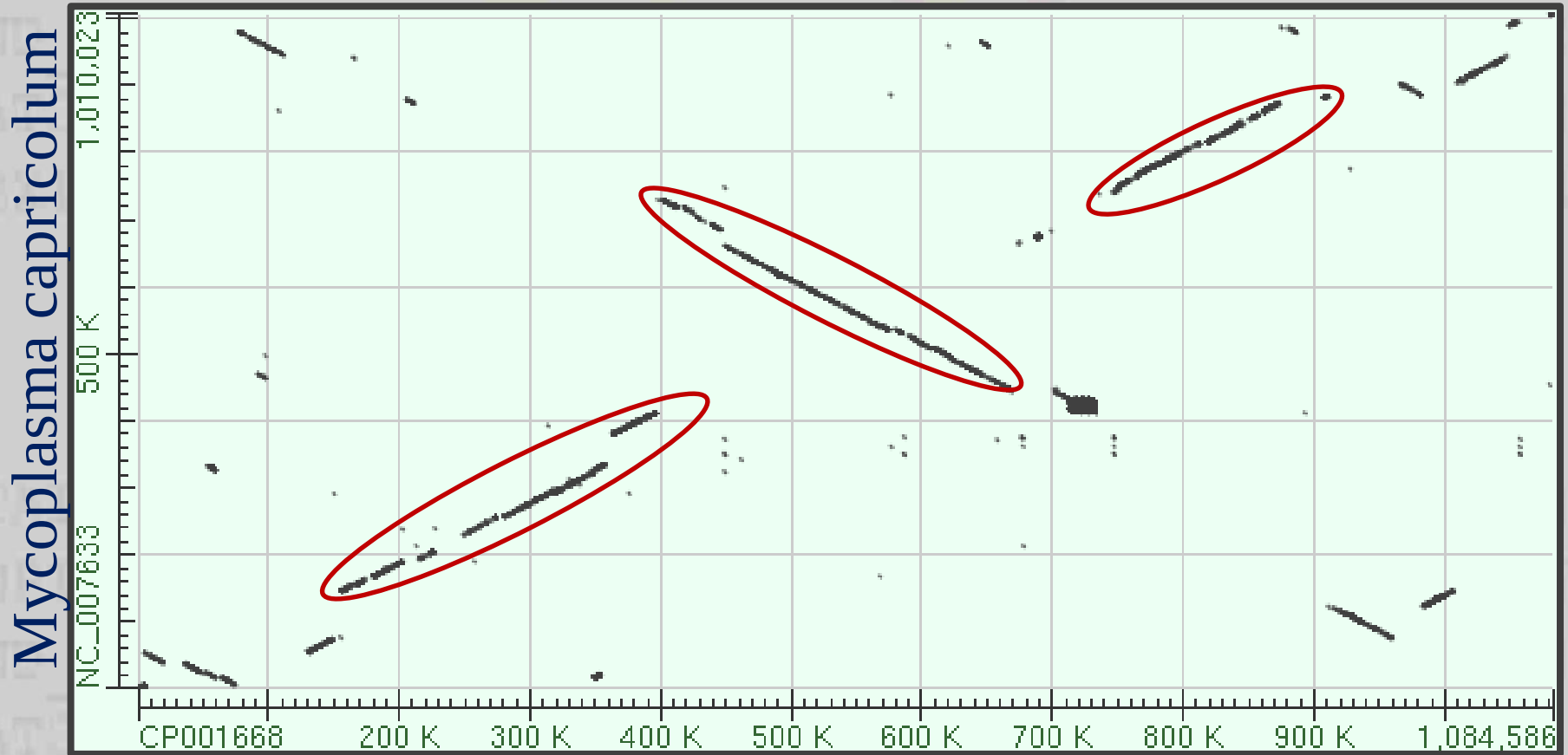
Делеция



Инверсия



Карта локального сходства геномов *M.capricolum* и *M.mycoides*



Pfam –одна из популярных БД семейств доменов белков и их выравниваний

GeneBee Motifs' Map - Adv... aba@belozersky.msu.ru - DB... Pfam: Family: Pico_P2A (...)

wellcome trust **sanger** institute

HOME | SEARCH | BROWSE ABOUT | FTP | HELP

Pfam

keyword search Go

19 1135 1 245 3
architectur sequences interaction species structures

Family: *Pico_P2A* (PF00947)

Summary
Domain organisation
Alignments
HMM logo
Trees
Curation & models
Species
Interactions
Structures

Summary

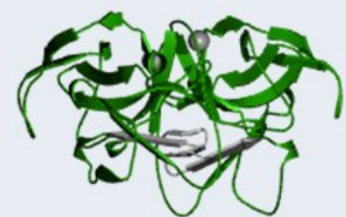
Picornavirus core protein 2A [Add annotation](#)

This protein is a protease, involved in cleavage of the polyprotein.

Literature references

1. Petersen JF, Cherney MM, Liebig HD, Skern T, Kuechler E, James MN; , EMBO J 1999;18:5463-5475.: The structure of the 2A proteinase from a common cold virus: a proteinase responsible for the shut-off of host-cell protein synthesis.
[PUBMED:10523291](#)

InterPro entry [IPR000081](#)



Example structure

PDB entry 2hrv: 2A CYSTEINE PROTEINASE FROM HUMAN RHINOVIRUS 2

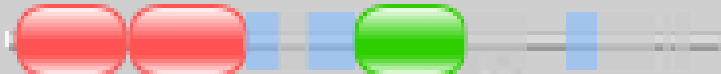
View a different structure:

2hrv

Крупные перестройки встречаются и в последовательностях белков

Примеры доменных архитектур с гомеодоменом (зеленый)

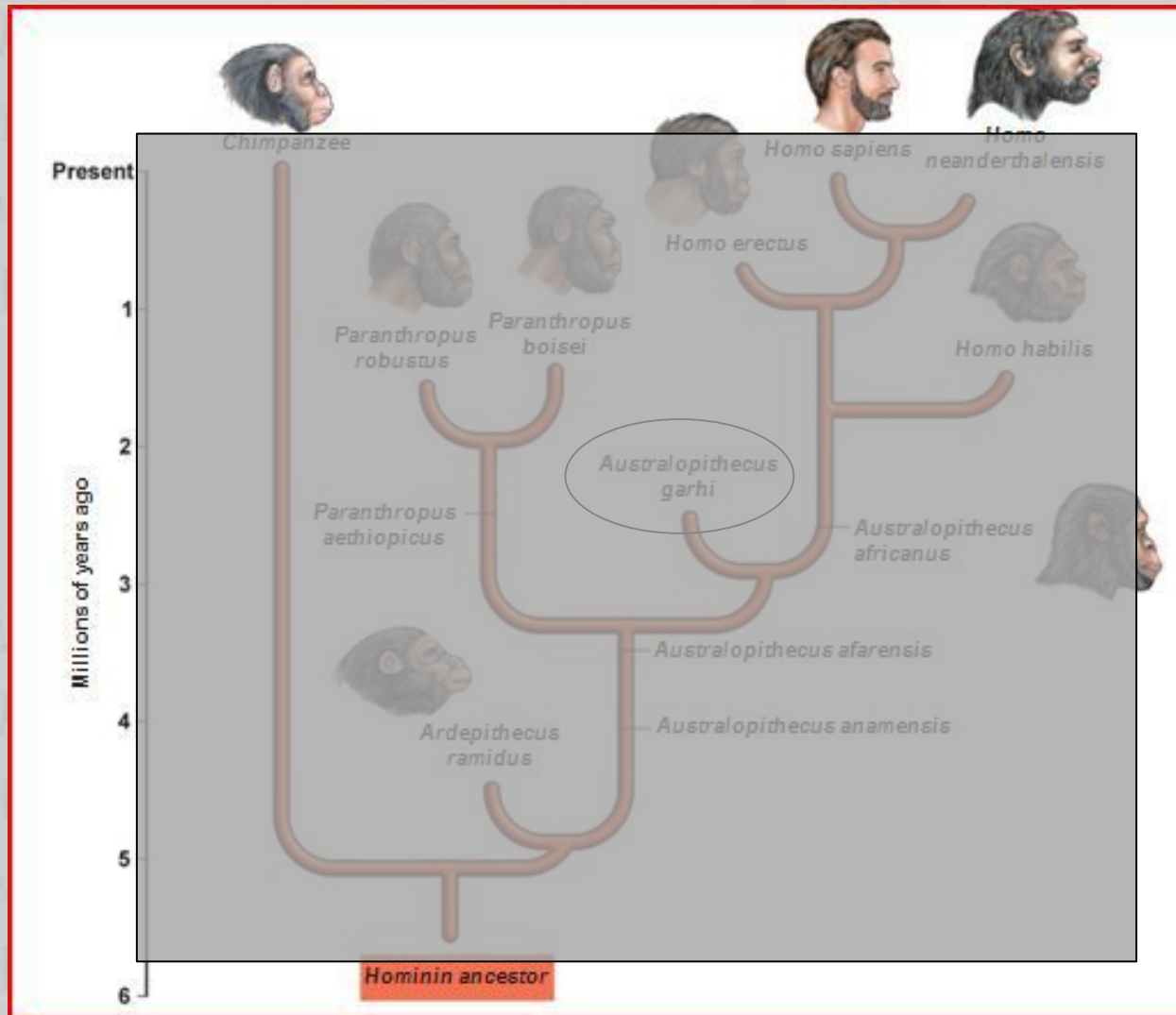
 2005 белков

 5068 белков

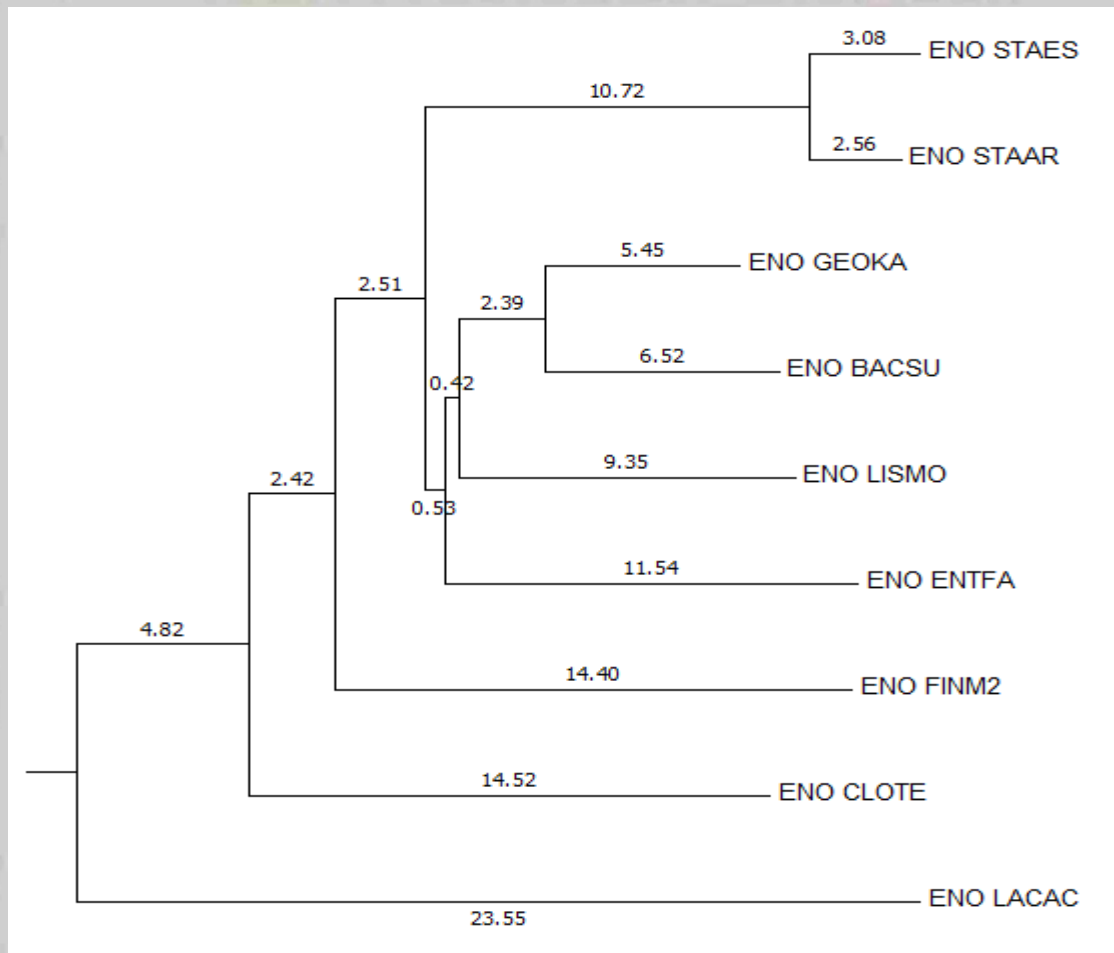
 2370 белков

Молекулярная филогения

Филогенетическое дерево видов

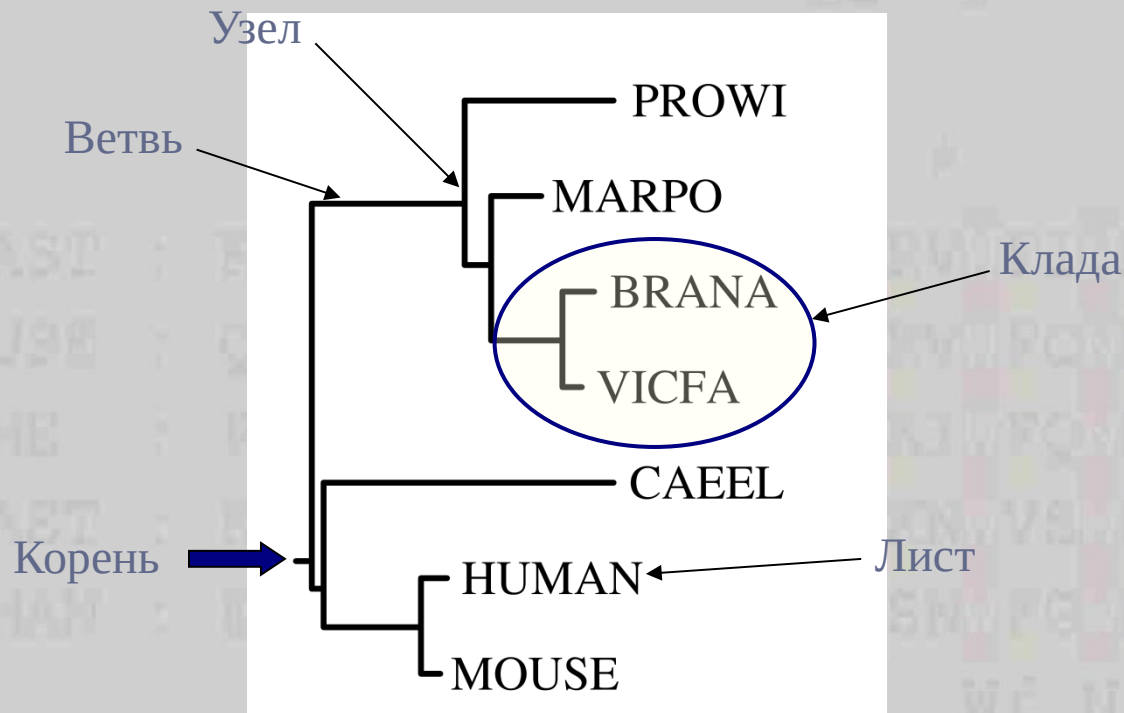


Филогенетическое дерево последовательностей энолаз из фирмикут



Описание структуры дерева (терминология)

- **Узел (node)** — точка разделения предковой последовательности (вида, популяции) на две независимо эволюционирующие. Соответствует внутренней вершине графа, изображающего эволюцию.
- **Лист (leaf)** — реальный (современный) объект; внешняя вершина графа.
- **Ветвь (branch)** — связь между узлами или между узлом и листом; ребро графа.
- **Корень (root)** — гипотетический общий предок.
- **Клада** — группа организмов, которые являются потомками единственного общего предка и всех потомков этого предка.



ENO_CLOTE 1 MKNYIEI IDVSARQVLDSSRNPTEVEVDVILEGGTTGRAAVPSGASTGIFEAVELRDGDKSVYNGKGV67
 ENO_FINM2 1 --- MSSII DVYAREVLDSRGNPTVEVEVYTEAGAMGSAIVPSGASTGVHEAVELRDNDKRFYLGKGV64
 ENO_ENTFA 1 --- MSIITDIYAREVLDSRGNPTIEVEVYTESGAFGRGMVPSGASTGEYEAVELRDGDKARYLGKGV64
 ENO_LACAC 1 -MLKSVIENVHLEIFDSRGNPTVEVHVHTLSNGVVGAKEVPSGASTGENEAVELRDGG-SRLGGKGV65
 ENO_LISMO 1 --- MSIITEVYAREVLDSRGNPTVEVEVYTEAGAFGRALVPSGASTGEYEAVELRDGDKARYLGKGV64
 ENO_BACSU 1 --- MPYIVDVYAREVLDSRGNPTVEVEVYTEDGAFGRALVPSGASTGEYEAVELRDGDKARYLGKGV64
 ENO_GEOKA 1 --- MSAIIDVYAREVLDSRGNPTVEVEVYTEDGGFGRALVPSGASTGEYEAVELRDGDKARYLGKGV64
 ENO_STAAR 1 --- MPIITDVYAREVLDSRGNPTVEVEVLTESGAFGRALVPSGASTGEHEAVELRDGDKARYLGKGV64
 ENO_STAES 1 --- MPIITDVYAREVLDSRGNPTVEVEVLTESGAFGRALVPSGASTGEHEAVELRDGDKARYLGKGV64

ENO_CLOTE 68 LKAIENVTNTIAEELV--GMNVLDQIAIDKTMIELDGTDNKKGKLGANAMLVSLACAKAAAYLGLS132
 ENO_FINM2 65 ETAVDNNVLEIADELV--GWDVFDQVIGIDNYLIELDGTENKSRRLGANAILGVSLAVAKAAADENGQR129
 ENO_ENTFA 65 TAVDNNVNTIIAEAIL--GYDVRDQMAIDKAMIDLDGTENKKGKLGANAILGVSIAVARAADYLEV129
 ENO_LACAC 66 SKAVNNVNTIEDALK--GMDPFDQPKIDGTIMIDLDTENKKGKLGANAILGVSMATAVAANANRQP130
 ENO_LISMO 65 LKAVENNVNDIADKII--GFDVTDDIGIDKAMIELDGTENKKGKLGANAILGVSLAAARAAADLGVH129
 ENO_BACSU 65 LTAVNNVNEIIAPELL--GFDVTEQNAIDQLLIELDGTENKKGKLGANAILGVSMACARAAADFLQIP129
 ENO_GEOKA 65 LKAVENNVNEIIAPEII--GLEVADQVAIDRKLIELDGTENKSGKLGANAILGVSLAVARAADDELGLP129
 ENO_STAAR 66 TAVENNVNEIIAPEIEGGEFVSVDQVSDKMMIALDGTENKKGKLGANAILGVSIAVARAADLLGQP131
 ENO_STAES 65 TAVENNVNEMIAPEIVEGEFVSVDQVSDKMMIQLDGTENKKGKLGANAILGVSIAVARAADLLGQP131

ENO_CLOTE 133 LYQYIGGVNAKVLPPVPMNIIINGGEHADNNVDLQEFMIMPAGASSFSEALRMCSVEYVHSLKSLKAK199
 ENO_FINM2 130 LFEYIGGVNGKTLPPVPMNIIINGGGHADNNVDIQEFMIMPVGGENFKESLRIGTEVFHNLKNNVLKSR196
 ENO_ENTFA 130 LYHYLGGFNKVLPTPMNIIINGGSHADNSIDFQEFMIMPVGAPTFKEALRMGAIEVFHALASILKGR196
 ENO_LACAC 131 LYRYLGGIDLE-PPQTFHNVIINGGEHADNGIDIEFEMITPIAKTSFRDGFKEIINVYHTLKKVLEDM196
 ENO_LISMO 130 LYQYLGGVNGKVLPPVPMNIIINGGEHADNNVDVQEFMIMPVGAPNFKEALRMGAELHALAVLKGK196
 ENO_BACSU 130 LYQYLGGVNSKTLPPVPMNIIINGGEHADNNVDIQEFMIMPVGAPNFREALRMGAELFHSLKSVLSAK196
 ENO_GEOKA 130 LYQYLGGVNAKTLPPVPMNIIINGGEHADNNVDIQEFMIMPVGAEFREALRMGAELFHSLKAVLKA196
 ENO_STAAR 132 LYKYLGGFNGKQLPVPMMNIIINGGSHSDAPIAFQEFMILPVGATTFKESLRWGTEIFHNLKSLSKR198
 ENO_STAES 132 LYKYLGGFNGKQLPVPMMNIIINGGSHSDAPIAFQEFMILPVGAEFKESELRWGAEIFHNLKSLSER198

ENO_CLOTE 200 GYDTGVGDEGGFAPNLSNEEAQIVIEAIEKAGYKPGVEIFIALDPASSEIF--EDGKYNL---KG261
 ENO_FINM2 197 KLNATAVGDEGGFAPNLSNEEAQIVIEAIEKAGYEPGKDVMIADAASEFY--EDGKYNM---AG258
 ENO_ENTFA 197 GLATSVGDEGGFAPNLSNEEGFEVIEAIEKAGYVPGKDVVLAMDAASSEFYDKEKGVVVLAD-SG262
 ENO_LACAC 197 GYETGLGDEGGFAPNLMKNEEALKALHESIIEKAGYKPGEDIGIACDCAASYFYNKEDGKYHL----258
 ENO_LISMO 197 GLNTGVGDEGGFAPNLSNEEALETIMQAIKDAGYKPGEEVKLAMDAASSEFYNRETGKYEL---KG260
 ENO_BACSU 197 GLNTAVGDEGGFAPNLSNEEAQIVIEAIEKAGFKPGEEVKLAMDAASSEFYNKEDGKYHL---SG260
 ENO_GEOKA 197 GYNTAVGDEGGFAPNLSNEEAQIVIEAIEKAGYKPGQVMLAMDVASSELYNKEDGKYHL---EG260
 ENO_STAAR 199 GLEATAVGDEGGFAPKFEGETDAVETIIQAIIEAGYKPGEEVFLGFDCASEFY--ENGVDYSKFEG263
 ENO_STAES 199 GLEATAVGDEGGFAPRFEGETDAVETIIKAIIEKAGYKPGEDVFLGFDCASEFY--ENGVDYTKFEG263

ENO_CLOTE 262 E-GKVLTPQEMVDYANLVAKYP-IISIEDGMAEEDWEGWRLITERLGTKIQLVGGDDLFLVTNTDRLT326
 ENO_FINM2 259 D-NAVRTSDEMIEYYKNLVKYP-IISIEDGLAEDDWDGWWKLTQLGSKIKQLVGGDDLFLVTNVKRLQ323
 ENO_ENTFA 263 E-GE-KTTEEMIAFYEELVSKYP-IISIEDGLDENDDWDGFKKLTVELGDKVQLVGGDDLFLVTNTTKLA326
 ENO_LACAC 259 E-GKVLTDDELAAYYDKLLDEFPPELMSMEDPYDENVEGMVKFTATHKDRIQIVLDDFICTNPALLK324
 ENO_LISMO 261 E-GVTRTSEEMVWYWEEMITKYP-IISIEDGLDENDDWGFKLLTERIGDRVQLVGGDDLFLVTNTTKL325
 ENO_BACSU 261 E-GVVKTSAMVDWYEELVSKYP-IISIEDGLDENDDWEGHKKLTERLGKVVQLVGGDDLFLVTNTTKLS325
 ENO_GEOKA 261 E-GVVKTSSEEMVWYWEELVSKYP-IISIEDGLDENDDWEGHKKLTERLGKVVQLVGGDDLFLVTNTTKLA325
 ENO_STAAR 264 EHGAKRTAAEQVDYLEQLVQKYP-IITIEDGMDENDDWDGWWKLTTERIGDRVQLVGGDDLFLVTNTEILA329
 ENO_STAES 264 EHGAKRSAAEQVDYLEELIGKYP-IITIEDGMDENDDWEGWKKLTDRIGDKVQLVGGDDLFLVTNTEILS329

ENO_CLOTE 327 MGIERGVANSILIKLNQIGTLTETLNAIEMAERAGYTAVVSHRSGETEDTTIADLVVAVNAGQIKTG393
 ENO_FINM2 324 KGIEMGVANSILIKLNQIGTLTETLDAIELAKRHGYTAVISHRSGETEDTTIADVAVATNAGQIKTG390
 ENO_ENTFA 327 EGIEKGIANSLIKVNVQIGTLTETFEAIEMAKAGYTAVVSHRSGETEDSTISDIAVATNAGQIKTG393
 ENO_LACAC 325 KAIKEGAGNASILIKLNQIGTVTETLETIRMSRKNGYNTMISHRSGETEDTFIADLAVAINGGQLKTG391
 ENO_LISMO 326 EGIEKGIANSLIKVNVQIGTLTETLDAIEMAKRAGYTAVISHRSGETEDSTIADIAVATNAGQIKTG392
 ENO_BACSU 326 EGIKNGVANSILIKVNVQIGTLTETFDAIEMAKRAGYTAVISHRSGETEDSTIADIAVATNAGQIKTG392
 ENO_GEOKA 326 EGIEKGVANSILIKVNVQIGTLTETFDAIEMAKRAGYTAVVSHRSGETEDSTIADIAVATNAGQIKTG392
 ENO_STAAR 330 KGIENGIGNSILIKVNVQIGTLTETFDAIEMAQKAGYTAVVSHRSGETEDTTIADIAVATNAGQIKTG396
 ENO_STAES 330 KGIENGIGNSILIKVNVQIGTLTETFDAIEMAQKAGYTAVVSHRSGETEDTTIADIAVATNAGQIKTG396

ENO_CLOTE 394 APARSERVEVAKYNQLLRIEDELGEMGEYRGLKAFY--NIIKK-431
 ENO_FINM2 391 STSRTRDRICKYNQLLRIEDMLSETAVYDGIKTFY--NLKK-428
 ENO_ENTFA 394 SLSTRDRIAKYNQLLRIEDQLGDVAEYKGLKSFY--NLKNK-432
 ENO_LACAC 392 APARSERVEVAKYNQLLRIEEELGKGERL---AFFPDDVDHD-428
 ENO_LISMO 393 APTRTRDRVAKYNQLLRIEDNLADLAHYHGNDFY--NLKK-430
 ENO_BACSU 393 APSRTDRVAKYNQLLRIEDQLAETAAQYHGNISFY--NLNK-430
 ENO_GEOKA 393 APSRTDRVAKYNQLLRIEDELGHTAIYQGRSFI--NLKK-430
 ENO_STAAR 397 SLSTRDRIAKYNQLLRIEDELFEAKYDGIKSFY--NLDK-434
 ENO_STAES 397 SLSTRDRIAKYNQLLRIEDELYETAKFEGIKSFY--NLDK-434

Входные данные для
 построения
 филогенетического
 дерева –
 выравнивание
 последовательностей

Схема алгоритмов построения деревьев

Выравнивание

Матрица расстояний между
последовательностями

символьно-
ориентированные
методы

Neighbor-Joining,
Fitch, FastME

Дерево

Матрица расстояний

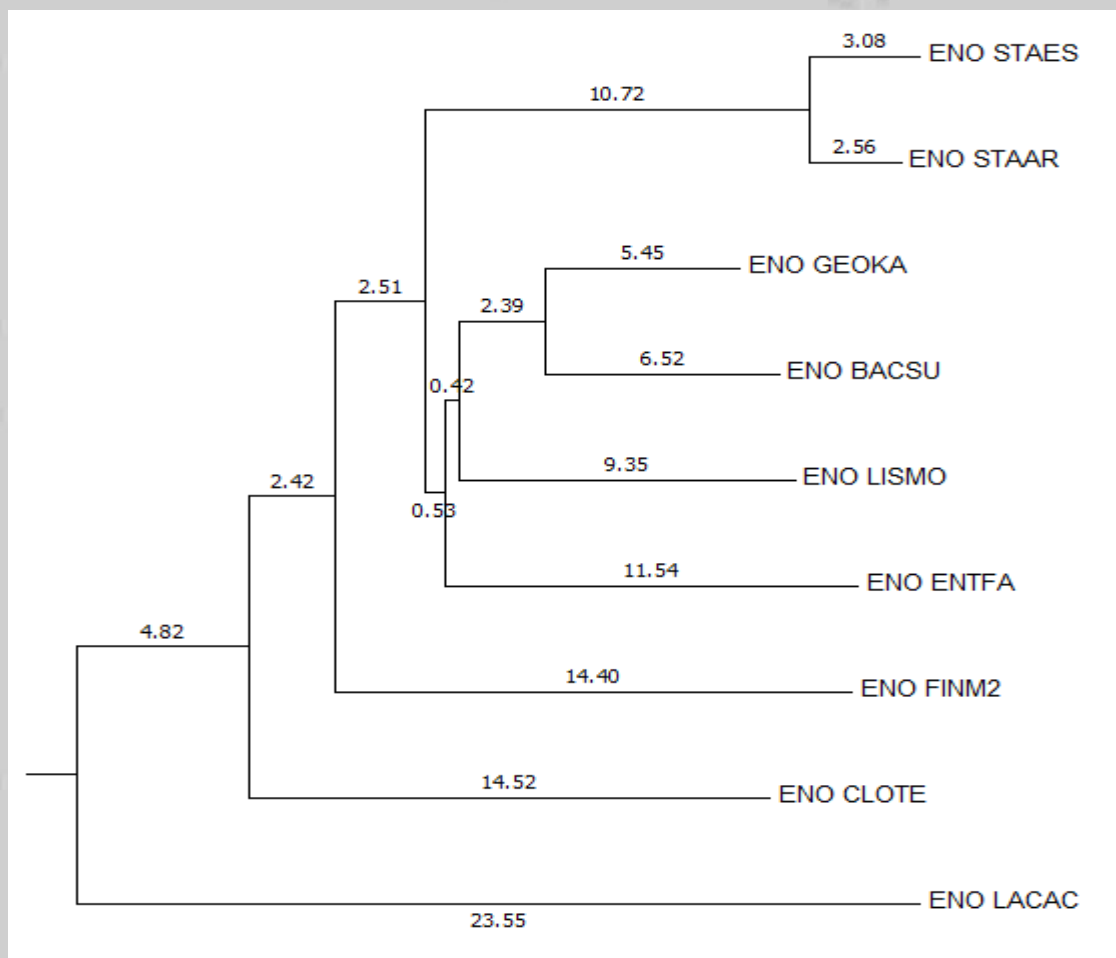
	MUSDO	CHICK	BOVIN	HUMAN
MUSDO	0	9.5	8.9	9.2
CHICK	9.5	0	3.4	2.8
BOVIN	8.9	3.4	0	1.7
HUMAN	9.2	2.8	1.7	0

Матрица расстояний

6

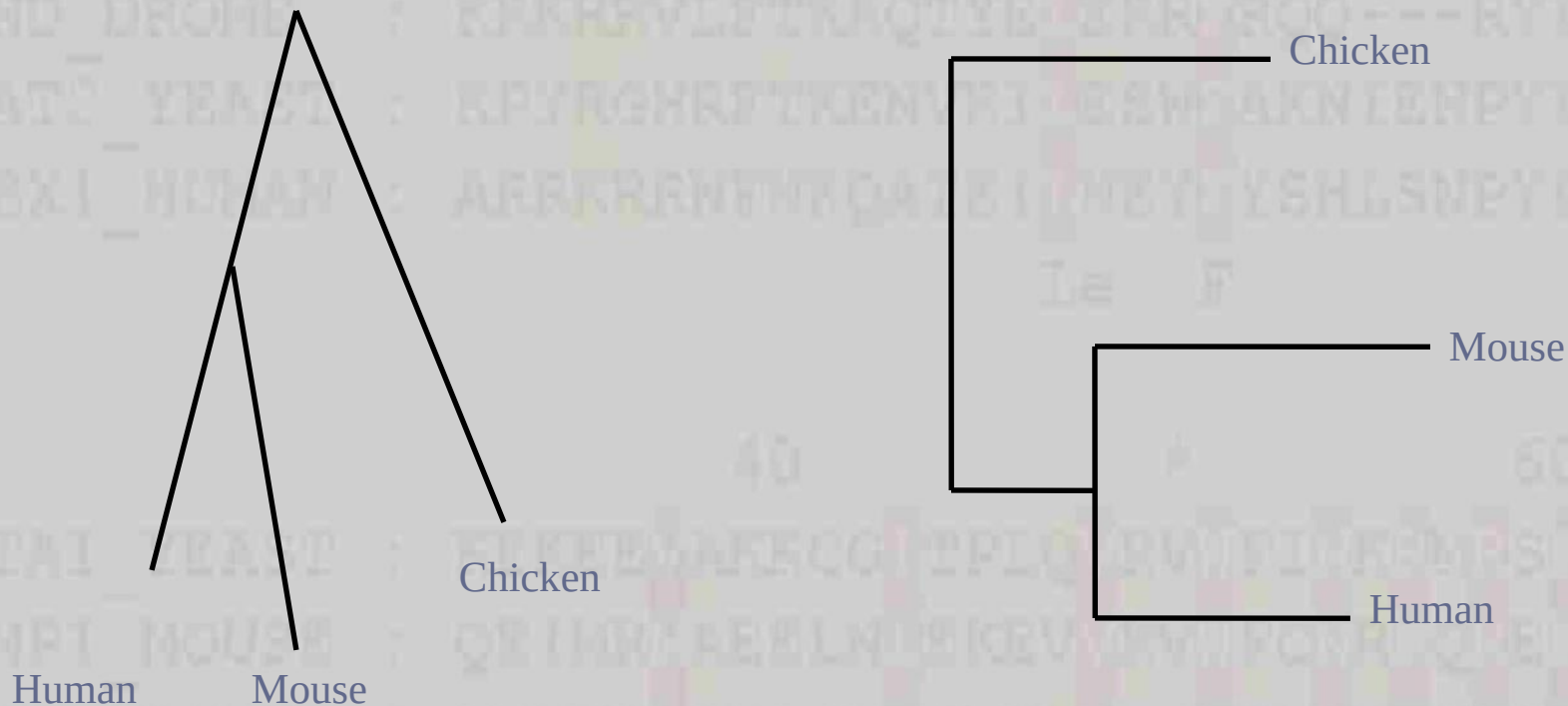
ENO_CLOTE/	0.000000	0.422314	0.412416	0.703889	0.374250	0.356587
ENO_FINM2/	0.422314	0.000000	0.397118	0.748527	0.345432	0.328530
ENO_ENTFA/	0.412416	0.397118	0.000000	0.756866	0.240851	0.271009
ENO_LACAC/	0.703889	0.748527	0.756866	0.000000	0.732893	0.710658
ENO_LISMO/	0.374250	0.345432	0.240851	0.732893	0.000000	0.212414
ENO_BACSU/	0.356587	0.328530	0.271009	0.710658	0.212414	0.000000

Расстояние между листьями на дереве



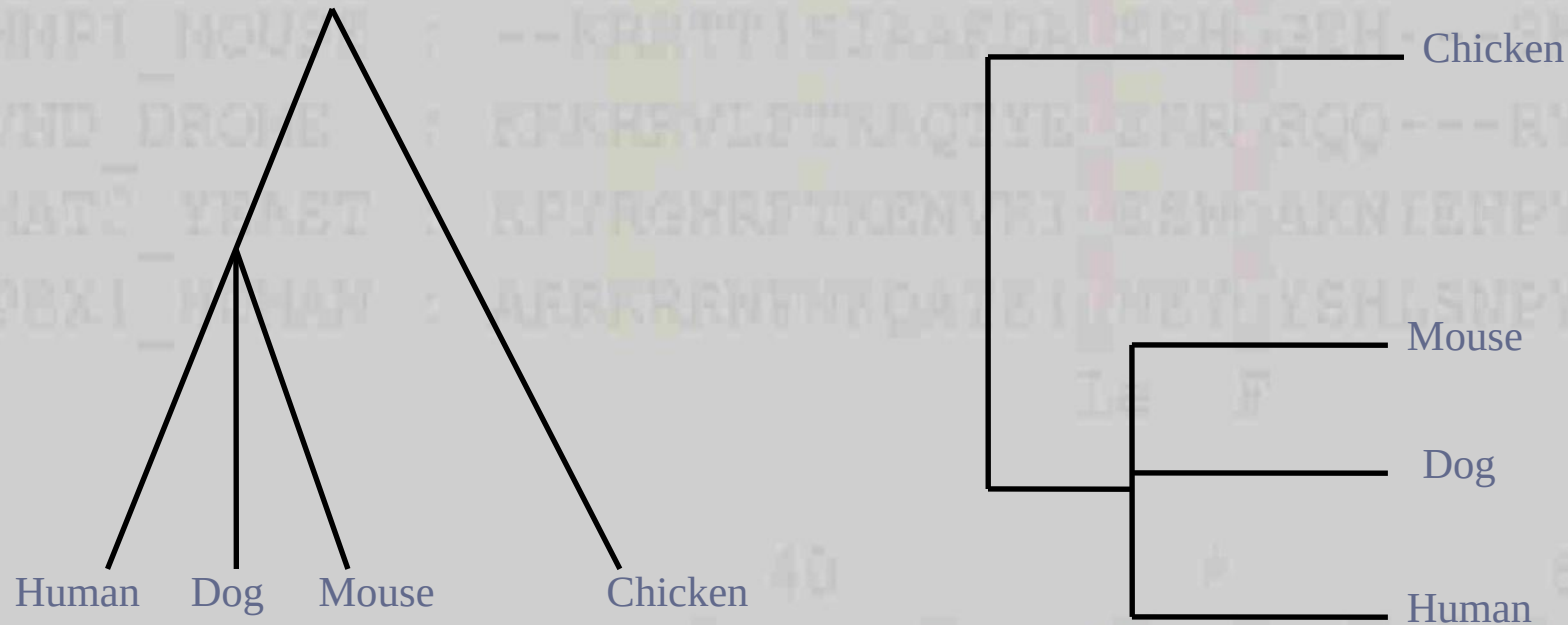
Задача алгоритма – построить такое дерево, чтобы расстояния между листьями на дереве было примерно таким же, как в матрице расстояний

«Молекулярные часы»: всегда идут, но иногда неточно

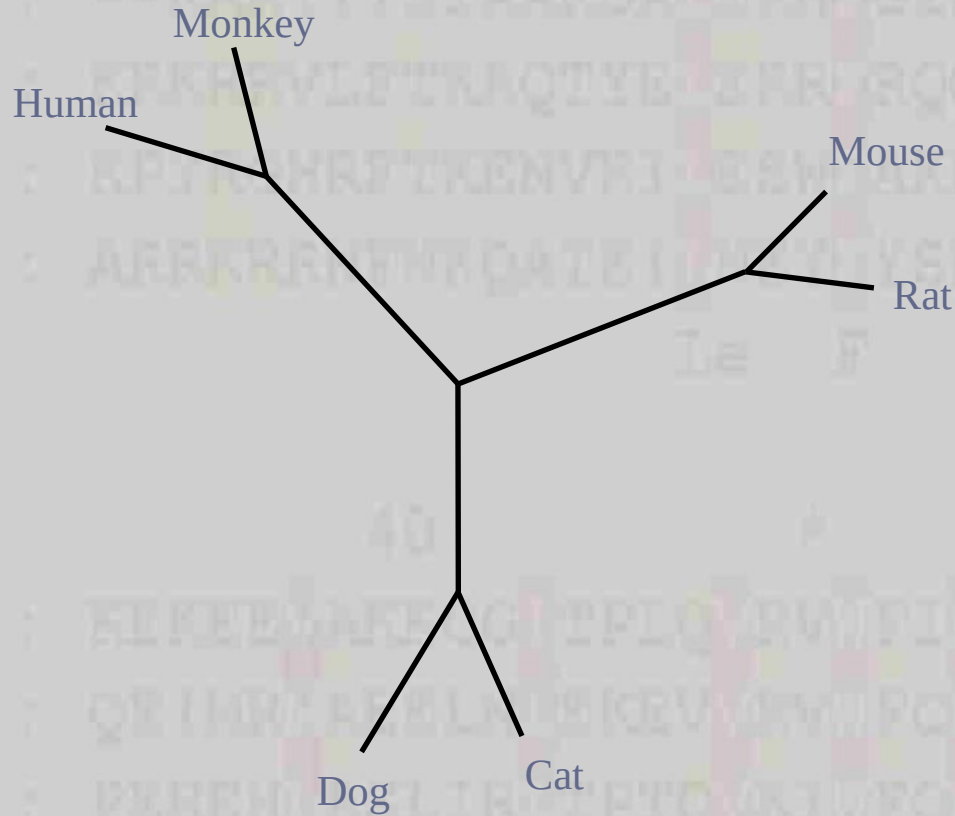


Когда хотят отразить разное число мутаций, произошедших на пути от общего предка, получается что-то вроде такого.

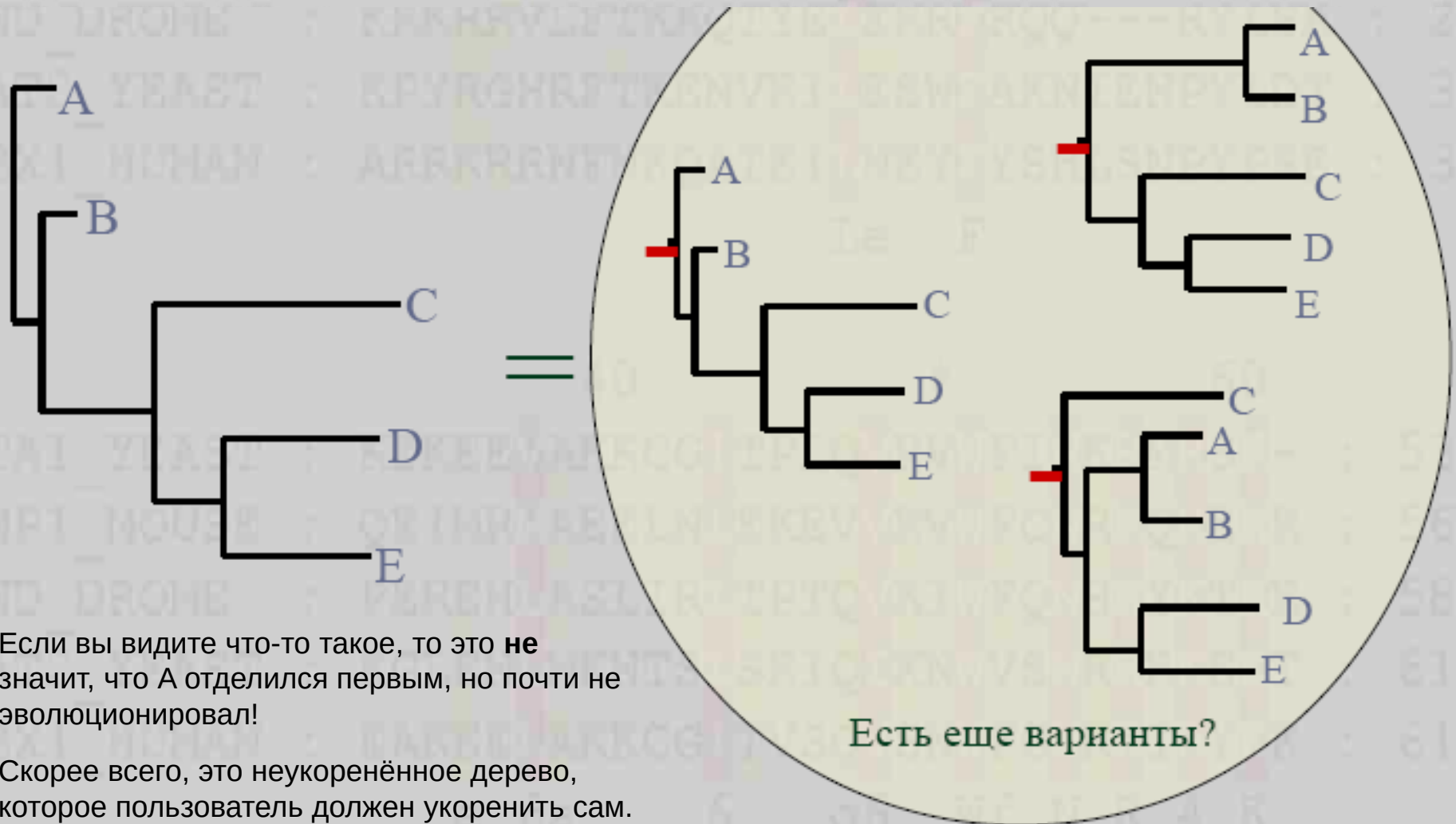
Небинарное дерево



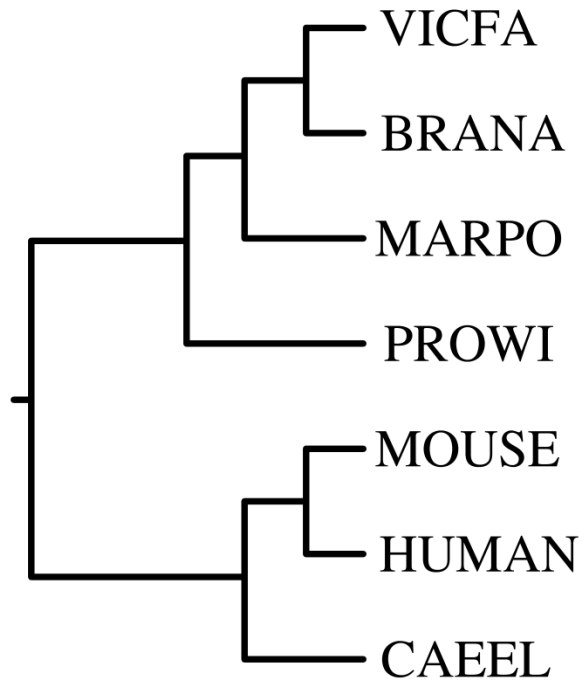
Неукоренённое дерево



Может быть укоренено многими способами



Скобочная формула



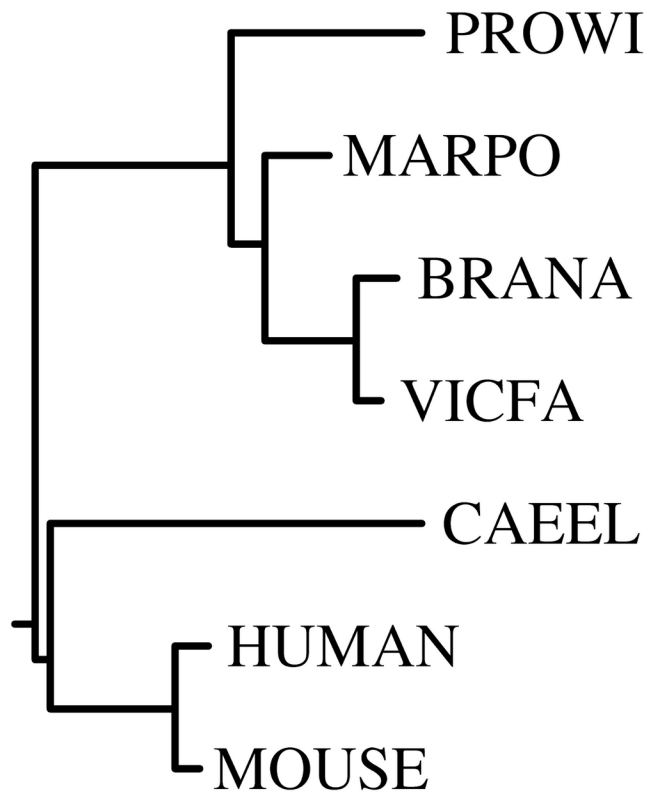
Newick Standard:

(((VICFA:3, BRANA:3):3, MARPO:6):2, PROWI:8):7, ((MOUSE:3, HUMAN:3):3, CAEEL:6):9);

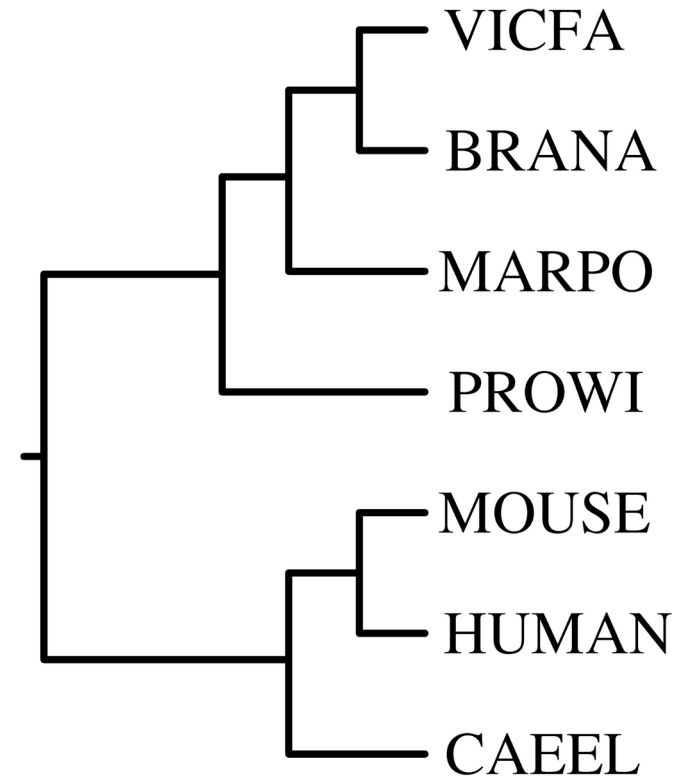
«The reason for the name is that the second and final session of the committee met at Newick's restaurant in Dover, and we enjoyed the meal of lobsters.»

Joseph Felsenstein, <http://evolution.genetics.washington.edu/phylip/newicktree.htm>

Топология дерева



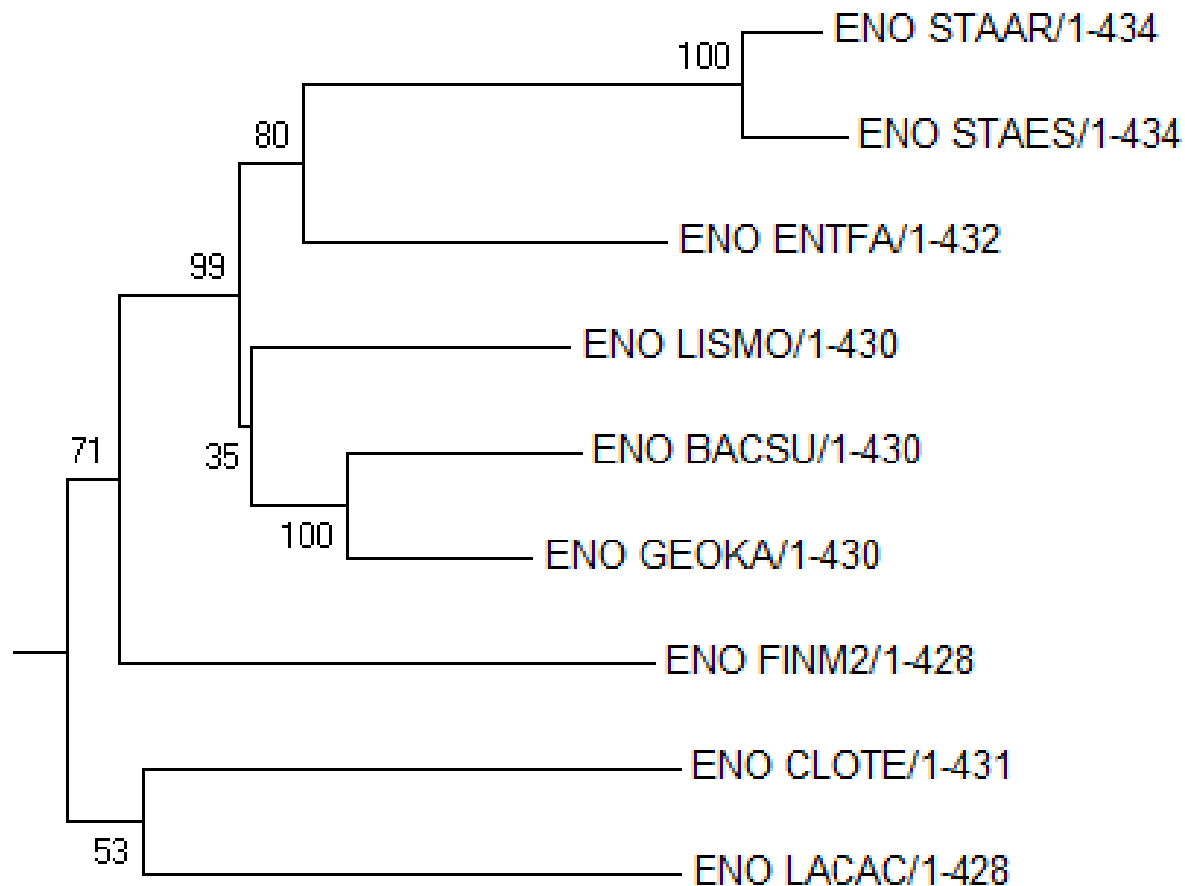
=



Можно ли проверить достоверность дерева?

Бутстрэп-анализ

(пример результата)



Эволюция видов и эволюция белков

Когда виды разделяются, то разделяются пути эволюции всех их белков...

В результате большинству белков одного вида соответствует **ортолог** в другом виде.

Но:

1) Бывают дубликации белков без разделения видов: два родственных белка существуют в одном геноме и эволюционируют (почти) независимо — такие белки называются **паралогами**.

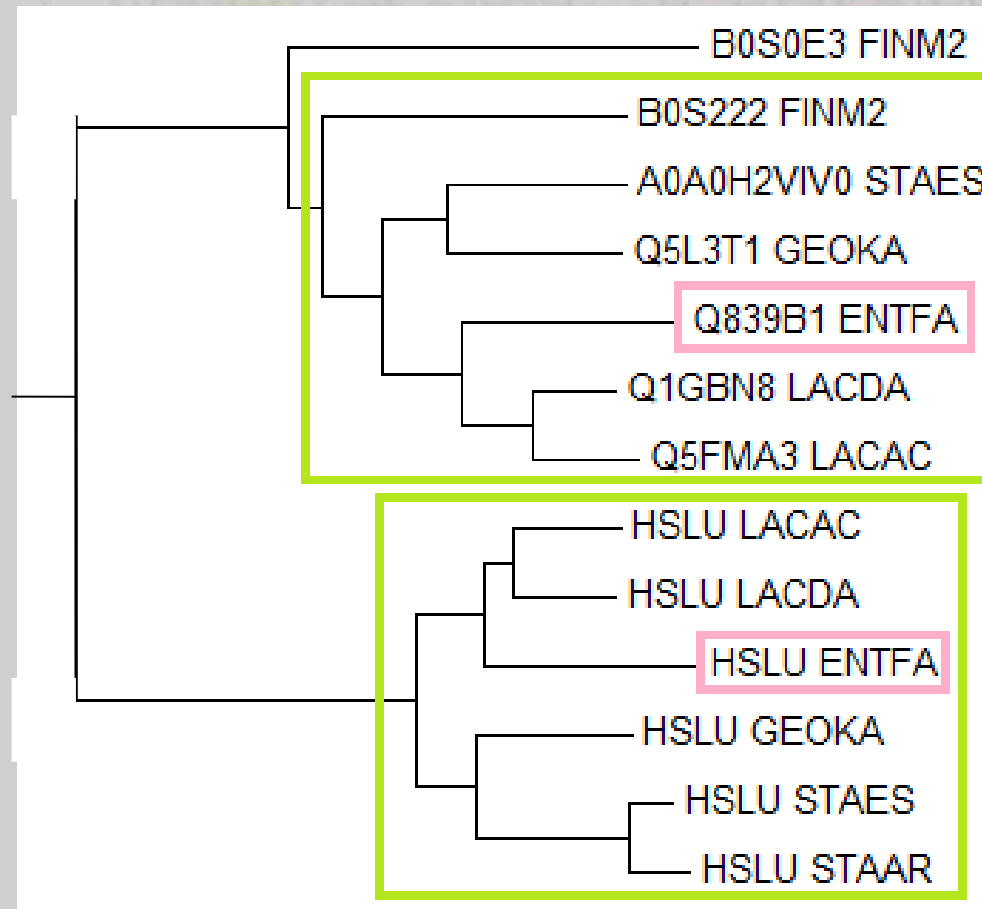
2) Бывают потери генов.

Если в двух видах потерялись по одному белку из пары паралогов, то может получиться, что общий предок белков, которые выглядят как ортологи, «жил» существенно раньше, чем общий предок видов.

3) Бывает, что два белка объединяются в один многодоменный, и наоборот.

Поэтому правильнее говорить об эволюции белковых доменов.

Ортологи и паралоги. Пример.



Программы

Веб-интерфейс к нескольким программам:

<https://ngphylogeny.fr/>

Сервис iTOL <https://itol.embl.de/> можно использовать для изображения деревьев