

Гомология

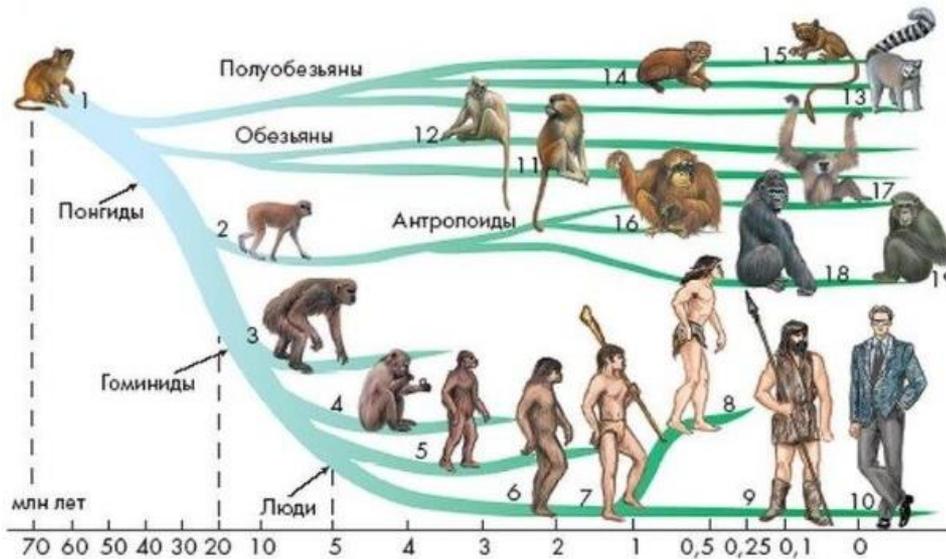
и

выравнивание

Множественное выравнивание последовательностей
гомологических белков

I. Гомология и сходство

1. Общность происхождения. Эволюция



В словарь:

- * Последний общий предок (LCA)
- * Гомология

Последний общий предок ныне живущих обезьян.

Гомоло́гия в биологии

сопоставимость частей сравниваемых биологических объектов, обусловленная общностью происхождения

wiki

Для целых организмов термин «гомология» не употребляют

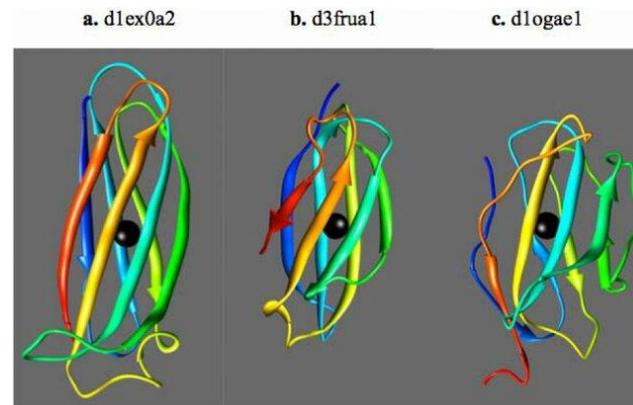
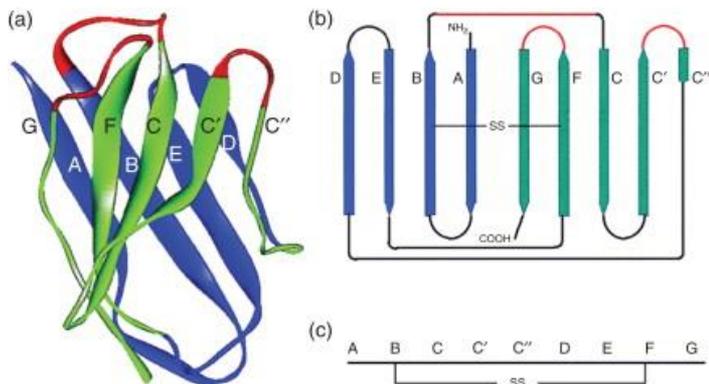
Гомология в молекулярной биологии

- Гомология того, что закодировано в геномах – генов **белков** и РНК, и просто участков ДНК
- **Белки гомологичны**, если их гены произошли из гена их общего предка.
- При каждом делении клетки ген дочерней клетки – копия гена материнской клетки.
- Ошибки – мутации – в ДНК случаются. Самая распространённая мутация: из пары mC-G получается пара T-G (mC метилированный цитозин). Такие неправильно спаренные основания часто возникают в процессе репликации
- Пары G-T репарируются системой репарации неправильно спаренных оснований

Гомология в молекулярной биологии 5

- О гомологии белков судят по сходству их последовательностей
- ГОМОЛОГИЧНЫЕ белки, давно разошедшиеся от общего предка, МОГУТ иметь отличающиеся функции
- ГОМОЛОГИЯ \neq СХОДСТВО последовательностей
 - последовательности могли сильно измениться в эволюции от общего предка
 - есть системы быстрой эволюции белковых последовательностей. Иммуноглобулины у животных и diversity-generating retroelement (DGR) у вирусов и бактерий
 - дольше в эволюции сохраняется сходство 3D структур (след. Слайд)
- СХОДСТВО \neq ГОМОЛОГИЯ Бывает случайное совпадение (аналогия). Чем длиннее фрагмент последовательности, тем менее вероятно случайное совпадение

Иммуноглобулиновая укладка (ImFold)



Kim C, Basner J, Lee B. Detecting internally symmetric protein structures. BMC Bioinformatics. 2010

ImFold встречается во многих белках одного организма, связанных с иммунитетом и не только, и у всех животных

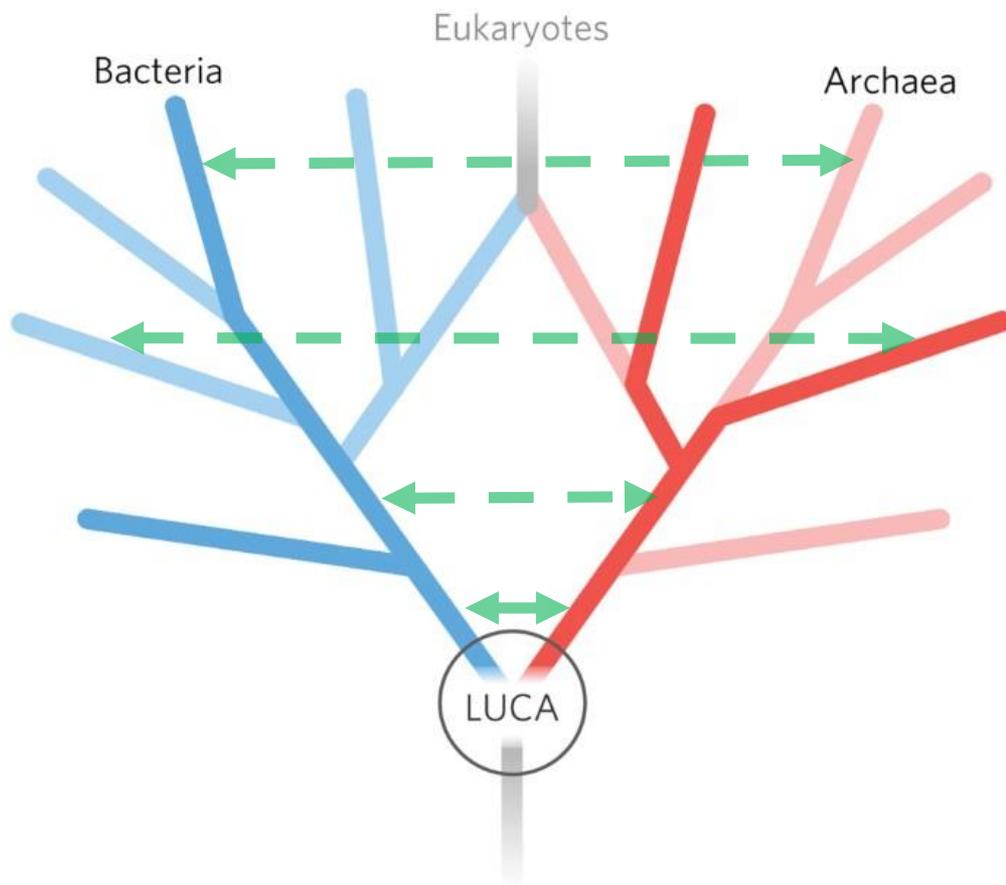
Последовательности ImFold могут сильно отличаться, детектируемое сходство только на коротких участках (увидите, когда будете делать упражнение)

Ход полипептидной цепи и элементов вторичной структуры очень похожи

Совмещать полипептидные цепи в пространстве научат в последующих курсах 3D

LUCA – Last Universal Common Ancestor.

7



О нём эволюционисты задумывались давно

Обоснование на молекулярном уровне было сформулировано в начале 2000х (2002 ref)

Были собраны все белки, гомологи которых встречаются во всех таксонах высокого порядка

Возраст исчисляется миллиардами лет.

Про горизонтальные переносы авторы рисунка забыли. ИСПРАВЛЕНО!

НЕ СЛЕДУЕТ ДУМАТЬ, что LUCA БЫЛ чем-то одним.

Всё выжившее – сложно устроено и обязательно разнообразно

Descendants of Queen Victoria



Много гомологичных частей у всех потомков.

Руки, ноги, лицо – общее можно найти, *кроме принца Уэльского.*

Очевидна дихотомия – мужчины, женщины.

Половое размножение вносит своё в анализ гомологичности.

Видимы и другие признаки, разделяющие потомков. Клетчатая юбка. Тоже не всё однозначно)))

The Prince and Princess of Wales with their children Left to right, standing: Prince George; Alexandra, Princess of Wales; the Prince of Wales; Princess Victoria of Wales. Seated: Princess Maud, with small dog on her lap; Prince Albert Victor; Princess Louise. Highland dress.



Aravind L et al., Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA. Proteins. 2002

2. Эволюция геномов бактерий

Половое размножение отсутствует



Мутации. Небольшие, локальные изменения от поколения к поколению.

Замены нуклеотидов, короткие делеции и вставки

Изменения локальные, но их может быть много в эволюции

Мутации. Происходят случайно. С разной частотой^{*)}

Контролируются отбором – носители вредных и слабо вредных мутации удаляются из популяции.

Накапливаются от поколения к поколению. Пытаются по их числу измерять время от потомков до последнего общего предка

Тем более, под отбором **Крупные единовременные изменения генома!** Сохраняются только не летальные. *Но из-за огромного количества организмов, мы видим те, которые нашли как приносить пользу*

^{*)} У *Deinococcus radiodurans* частота повыше. Почему?

Эволюция белков

Локальная - небольшие изменения в гене
(Замены а.к. Делеции Вставки)

Большие изменения:

- 1) Накопленные небольшие изменения
- 2) Небольшие изменения гена ведущие к большим изменениям белка
 - 1) Мутация стоп кодона => удлинение последовательности белка
 - 2) Мутация кодона на стоп кодон
 - 1) гибель белка = псевдогенизация или
 - 2) Укорочение последовательности белка
 - 3) Программируемый сдвиг рамки считывания
 - 3) Мутация в сайте инициации (начала) трансляции
- 3) Крупные перестройки генома, затрагивающие гены!
- 4) Закодированные в геноме перестройки для быстрого изменения последовательности белка (Img, DGR)

Гомологию белков выводят из сходства их последовательностей

Белки, как молекулы, определяются (почти ^{*)} однозначно) своей последовательностью

Поэтому их гомология определяется похожестью последовательностей

Как говорить можно, и как нельзя

Высокая ~~ГОМОЛОГИЯ~~ последовательностей
– **НЕТ** У гомологии **НЕТ СТЕПЕНЕЙ**

СТЕПЕНИ ЕСТЬ У СХОДСТВА

+Высокое **СХОДСТВО** последовательностей

^{*)} почему почти?



II. Эволюционное выравнивание

Выравнивание последовательностей потомков относительно предка

14

предок	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17	
предок	TATGCGAATGCCCTGAA	
сын	TATG A GAATGCCCTGAA	замена
внук	TATG C GAATG C TCTGAA	замены
правнук	TATG C GAAT C G C TCTGAA	вставка 1 п.н.
праправнуку	TATG A GA A A C G C TCTGAA	замены
прапраправнук	T G A GA A A C G C TCTGAA	делеция 2 п.н.
потомок	1 4 5 6 7 8 9 9a 10 11 12 13 14 15 16 17	

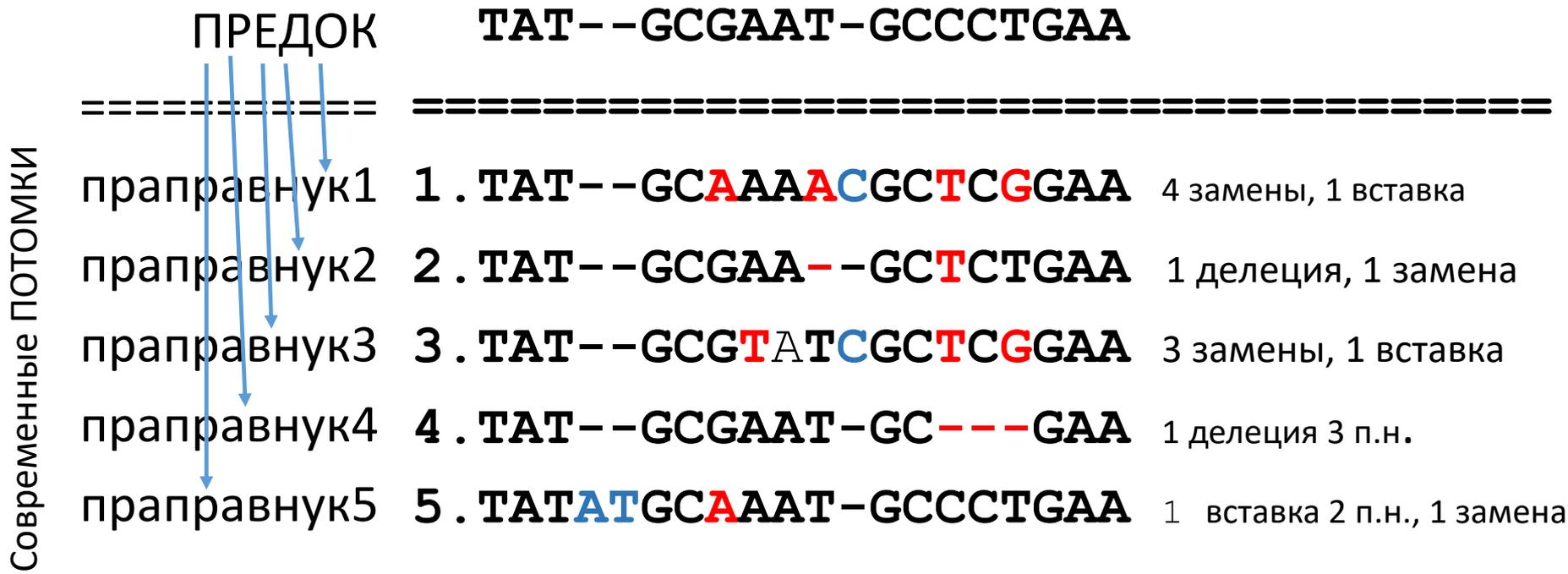
Нукл-ы потомка с номерами как у предка являются гомологами нукл-в предка

предок	TATGCGAAT-GCCCTGAA
сын	TATG A GAAT-GCCCTGAA
внук	TATG C GAAT-G C TCTGAA
правнук	TATG C GAAT C G C TCTGAA
праправнуку	TATG A GA A A C G C TCTGAA
прапраправнук	T--G A GA A A C G C TCTGAA

Выравнивание: гомологичные нуклеотиды - друг под другом

Идеальное выравнивание потомков относительно общего предка

Синий – вставка
 Красный – замена
 - Делеция
 относительно ПРЕДКА



Такое выравнивание бывает известно только в экспериментах по изучению эволюции. E.coli (Ленский), шизофилум (А.Кондрашов), др.

Единовременные крупные изменения в последовательности белка. Длина кратна 3 или нет.

- Делеция или вставка в гене
 - кратна 3
 - НЕ кратна 3
- Инверсия – кусочек вырезан и вставлен, но прямая и обратная цепочка перепутаны.
- Транслокация – перенесение фрагмента в другое место. С сохранением ориентации или нет.
- Потеря стоп-кодона
- Приобретение нового старт кодона.

Сокращения

* а.к.о. – аминокислотный остаток

* aa – amino acid residue

4. Множественное выравнивание последовательностей гомологичных белков

(анализ результата выравнивания, построенного программой)

Эволюционное выравнивание (редко достижимый идеал) :
в каждой колонке стоят гомологичные аминокислотные
остатки (или символы гэпа “-” на месте делеций и вставок)

17

Смысл выравнивания последовательностей гомологичных белков

- Некоторые кодоны а.к.о. гомологичных белков потомков *произошли из одного кодона последнего общего предка* этих белков, или были делятированы в эволюции, или появились в результате вставки новых кодонов
- Цель программ множественного выравнивания *последовательностей гомологичных белков* воспроизвести эволюционное выравнивание
- Это не всегда хорошо получается. Есть проблемы.
- *Программы выравнивания основываются на сходстве последовательностей*, так как последовательности белков обычно подвержены стабилизирующему отбору и потому их последовательности изменяются медленно
- Сходство может появиться случайно (теор. вер.)

Вывод. Нужно учиться чему верить в выравнивании, а чему нет!

Для этого нужно набираться опыта на примерах.

Биологический смысл выравнивания

10 20 30 40 50 60

EJL77459.1 GVDLVF GGPPCQGFSSQIGMRR-LDDER-NE LYQQYTRIVAKLKPRVFLMENVPNLALMNKGH
RXK67093.1 DLDVVF GGPPCQGYSSQIGTRR-LDDER-NE LYLQYARIVEKQRPRMFLMENVPNMVL LNKGH
OJY44288.1 NVDLVF GGPPCQGYSSQIGTRD-LHDPR-NR LFEEFARVVATLKPFLMENVPNL LLLNKGH
TRU90449.1 NPEMIV GSPPCQDFSSAGKRNEGLGR--ANLTLTFAEIVTRVSPQWFVMENVD---RIEKSK
OXI46696.1 GTDLVF GGPPCQGFSSQIGMRR-LDDER-NE LYKQYTRVVSTLRPRVFLMENVPNLALMNKGH
AVZ30243.1 EIDVVF GGPPCQGFSLIGKRS-FEDPR-NS LVFHYIRLVLELSPKFFVIENVKGMTAGNHQA
AFZ12381.1 DIIGFI GGAPCPDFSVGGKNRGSEGDK-GKLSASYIELICQQKPDFFLFENVKGLYKTKKHR
HCQ21462.1 HIIGFI GGPPCPDFSVGGKKNKGLGDN-GKLSASYIELICQNLPDFFLFENVKGLWRTTKHR
EDN77159.1 SLIGFI GGPPCPDFS IAGKNKGKDGDN-GKLSLSYTNLI IEMKPDFFLFENVKGLWRTARHR
SOD91684.1 EVSLVV GGAPCQPF SNIGKKLKGNDRNGDLFLEFVRMVKGIQPEAFIFENVVGI TQNKHSD
QCS48280.1 NVVGF I GGPPCPDFS IGGKNRGRQGDH-GKLSSESYIDLIIQHQPDFFI FENVKGLYRTKKHR
SMB95934.1 GLFGII GGPPCPDFSVGGKNRGENGEQ-GRLSKVFDKIDLQPVFFLYENVPGLIRTAKHR
RUO38876.1 SPVGF I GGPPCPDFSVGGKNRGHEGEN-GR LTRTYVDGIIKYAPDFFI FENVKGLWRTKRHR
OIP70538.1 TIDLIC GGPPCQGFSTIGTND-KKDHR-NLFFEF LRMVETFKPNFI ILENTGLLAKKNES
AFY60915.1 NLVGFV GGPPCPDFS IGGKKNKQYGDN-GKLTKVYVDII IENQPDFFI FENVKGLWRTRSRHR
CUR30340.1 DLIGFI IAGPPCPDFSVGGKNRGKNGDQ-GKLTACYVELICQQRPDFFI FENVKGLWSTKKHR
TAK03971.1 QAALVV GGAPCQPF SNLGSKRGTADSR-GTLFQDFIRIVKGV RPKGFI FENVEGLTQDKHKG
AEE51071.1 KVALVV GGAPCQPF SNIGKKEGENDA KNGDLFLEFVRMVKGIQPEAFIFENVAGIIQSKHSG
RTR31666.1 RLVGFV GGPPCPDFSVGGKKNKGS EGEN-GK LTRTYIDLIVKDNPDYFI FENVKGLWRTTRHR
PTU64472.1 NIDLVF GGPPCQGFSSQIGTRR-LDDER-NE LYKQYTRIVKTLKPRVFLMENVPNLAMMNKGH

DNA (cytosine-5-)-methyltransferases

Продолжение того же выравнивания

Не всегда так хорошо как на предыдущем слайде

20

```
200      210      220      230      240      250      260      270      280
YG-VPQDRKRVFIVGYREDLNLK-----FEFPKPLNKKVTLRD-----AIGDLPE-F
YG-VAQDRERVFYVGFVKDLNLSN-----FE-FYPPISEKERKYLKD-----SIWDLKDNA
YG-VAQERKRVFYIGFRKDLEIKF-----SFPKGSTVEDKDKITLKD-----VIWDLQDTA
YG-VAQDRKRVFYIGFRKELNIN-----YLPPIPHLIKPTFKD-----VIWDLKDNF
YG-IPQQRDRLLVFAAKQG-----VIKIIPPTHTPENYR-----TVRDVIGSLATNY
YG-VPQSRQRVFFIGLKSDRPLNQQ-----ILTLP-----PSKVI ESEYTSLEEAI SDLPVIE-----AGEGGEVQDYPVAE
CG-VPQLRKRTFVIGHRHGS IAD-----LANVLQQRLAKQSL-----TVRDYFG-
CG-VPQSRTRFSLIGKLNSEHNF-----LIPTLSRKLSDKPM-----TVRDYLG-
YG-VPQRRHRI IIVGIRKDQD-----VAFRVPEPTHKEKYR-----TASEALADIPEDA
IG-AHHQRHRWFCLAIRKDYEP EE-----IIVSVNATKFDWENNEPPCQVDNK-----SYENSTLVRLAGYS
FGNIPQNRERIYIVGFRN-----IEHYKNFNFPMPQP-----LTLTIKDMINLS
FN-VPQNRERLYIIGIREDLIKNEE-----WSLDFKRKDI LQKGKQRLVELDIKSFNFRWTAQ-----SAATKRLKDLLEEY
FG-IPQNRERVFCSILN-----PNEDFTFPQKQ-----NLTL SMNDLLEEM
FG-SSQARRRVFMISTLNEF-----VELPKGDKKPKS-----IKKVLNKIVSE
FG-IPQNRERIYLVGF-----LNHDVDFRFPQP-----IGQATAVGDI LEA
FG-LPQNRERIYIVGFDRKS-----ISNYSDFQMPTP-----LQEKTRVGNILES
FG-VPQNRERIYIVGFNKEK-----VRNHEHFTFPTP-----LKT KTRVGDILEK
FG-VPQNRERIYIVGFHKS-----TGVNSFSYPEP-----LDKIVTFADIREEK
FQ-VPQNRRLVYIVGLDQSQPELT-----ITSHIGATDSHKFKQLSNQASLFD-----TNKIMLVRDILED
FG-VPQNRVRIYILGILGSKPKLT-----LTSNVGAADSHKYK-NEQISLFD-----ES-YATVKDILED
FG-IPQKRKR FYLVAF LNQN-----IHFEF PKP-----PMISKDIGEVLES
FG-LPQRRERIVIVGFHPDLG-----INDFSFPKGN-----PDNKVPINAI LEH
YG-IPQKRERIYMICFRNDLN-----IQNFQFPKP-----FELNTFVKDLLLPD
YG-NAQRRRRVFI FGYKQDLNYSKAME-----ESPLDKI IYHNGLFAEAFP I EDYANKNR-----VNRTHITHDIVDISDNF
YG-TPQRRKRAI IRLNKKGT IWN-----LPLKQNI VSVEQ-----AIGNLPSIESGK
GG-TPQVRERVFITATLVPERMRDER I PR TETGE I DAEAIGPKPVATMNDRFP I KKGTEL FHPGDRKSGWNLLTSGI I REGDPEF
YG-VAQNRDRVFI IGIQQKLGVPD-----FSFPEYSESEQRLYDILDNLQTPSII-----PESLPIQRNLFGEF
FG-VAQNRDRVFI VGIQQKLDLNG-----FSFPEYAESDQRLYHILDNLEAPETK-----LESIP IQRNLFGEF
YD-VAQKRERIVIIGIREDLVK-----EQYPPFRFPLAQ-----VYKPVLKDV LKDY
YG-VSQLRPRVLFVALKNEYTN-----FFKWPEPNSEQPK-----TVGELLFDLMSE
```

Почему такая неоднородность качества колонок в выравнивании?

- Программа построила выравнивание неправильно?
- Неравномерная скорость мутаций в разных местах белка. Почему?
 - Мутации в геноме происходят неравномерно – НЕТ. Мутации происходят случайно, им всё равно – где (в первом приближении).
 - Отбор решает какие мутации и даже крупные перестройки оставить, а какие запретить.
- Вернёмся к слайдам. Где выравнивание правильное (имеет шанс соответствовать эволюционному), а где - НЕТ

Биологический смысл выравнивания

хорошее выравнивание

	10	20	30	40	50	60
EJL77459.1	GVDLVF	GGPPCQGF	SQIGMRR-LDDER-NEL	YQQYTRIVAKLKP	RVFLMENVPNL	LALMNKKGH
RXK67093.1	DLDVVF	GGPPCQGY	SQIGTRR-LDDER-NEL	YLQYARIVEKQRP	RMFLMENVPNM	VLLNKGH
OJY44288.1	NVDLVF	GGPPCQGY	SQIGTRD-LH DPR-NRL	FEEFARVVATLKP	KFLMENVPNL	LLL NKGH
TRU90449.1	NPEMIV	GSPPCQDF	SSAGKRNEGLGR--	ANLTLTFAEIVTRV	SPQWFVMENV	D---RIEKSK
OXI46696.1	GTDLVF	GGPPCQGF	SQIGMRR-LDDER-NEL	YKQYTRVVSTLR	PRVFLMENVPNL	LALMNKKGH
AVZ30243.1	EIDVVF	GGPPCQGF	SLIGKRS-FEDPR-NS	LVFHYIRLVLEL	SPKFFVIENVK	GMTAGNHQA
AFZ12381.1	DIIGFI	GGAPCPDF	SVGGKNRGSEGDK-	GKLSASYIELIC	QQKPDFFLFEN	VKGLYKTKKHR
HCQ21462.1	HIIGFI	GGPPCPDF	SVGGKNKGHLGDN-	GKLSASYIELIC	QNLPDFFLFEN	VKGLWRTTKHR
EDN77159.1	SLIGFI	GGPPCPDF	SVAGKNKGKGDGN-	GKLSLSYTNLI	IEMKPDFFLFEN	VKGLWRTARHR
SOD91684.1	EVSLVV	GGAPCQPF	SNIGKKLGKNDER	NGDLFLFVRMV	KGIQPEAFIFEN	VGITQNKHSD
QCS48280.1	NVVGFI	GGPPCPDF	SVGGKNRGRQGDH-	GKLSSEYIDLII	QHQPDFFLFEN	VKGLYRTKKHR
SMB95934.1	GLFGII	GGPPCPDF	SVGGKNRGENGEQ-	GRLSKVFVDKI	LDLQPVFFLYEN	VPGLIRTAKHR
RUO38876.1	SPVGFI	GGPPCPDF	SVGGKNRHEGEN-	GRLTRTYVDGI	IKYAPDFFLFEN	VKGLWRTKRHR
OIP70538.1	TIDLIC	GGPPCQGF	STIGTND-KK DHR-	NFLFFFLRMVET	FKPNFIILEN	VTGLLAKKNE
AFY60915.1	NLVGFI	GGPPCPDF	SVGGKNKGQYGDN-	GKLTKVYVDII	IENQPDFFVFEN	VKGLWRTSRHR
CUR30340.1	DLIGFI	AGPPCPDF	SVGGKNRGKNGDQ-	GKLTACYVELIC	QQRPDFFVFEN	VKGLWSTTKKHR
TAK03971.1	QAALVV	GGAPCQPF	SNLGSKRGTADSR-	GTLFQDFIRIV	KGVRPKGFI	FENVEGLTQDKHKG
AEE51071.1	KVALVV	GGAPCQPF	SNIGKKEGENDA	KNGDLFLFVRMV	KGIQPEAFIFEN	VAGIIQSKHSK
RTR31666.1	RLVGFI	GGPPCPDF	SVGGKNKGSEGEN-	GKLTRTYIDLIV	KDNPDYFIFEN	VKGLWRTTRHR
PTU64472.1	NIDLVF	GGPPCQGF	SQIGTRR-LDDER-NEL	YKQYTRIVKTLK	PRVFLMENVPNL	LAMMNKKGH

Учитывать

* Есть конс. позиции

* Нет гэпов

Зелёный – ОК.

жёлтый – есть выравнивания отдельных последовательностей (блоки);

между блоками подгонка программы, гомологии по а.к.о. по колонкам нет

13

200	210	220	230	240	250	260	270
YIG-VPQDRKRVFIVGYREDLNLK					FEFPPKPLNKKVTLRD		AIGDL
YIG-VAQDRERVFYVGFVKDLNLSN				FE	FPYPISEKERKYLKD		SIWDL
YIG-VAQERKRVFYIGFRKDLKIKF				SFPK	GSTVEDKDKITLKD		VIWDL
YIG-VAQDRKRVFYIGFRKELNIN				YLPP	IPHLIKPTFKD		VIWDL
YIG-IPQQRDLVLFVAAKQG				VIKI	IPPTHTPENYR		TVRDVI
YIG-VPQSRQRVFFIGLKS	DRPLNQQ		ILTP		PSKVI	ESEYTSLEEAI	SDLPVIE
CG-VPQLRKRTFVIGHRHGS	IAD			LANVL	QQRLAKQSL		TVRDY
CG-VPQSRTRFSLIGKLNSEHNF				LIPTL	SRKLSDKPM		TVRDY
YIG-VPQRRHRIIVGIRK	DQD			VAFRV	PEPTHKEKYR		TASEALADI
IIG-AHHQRHRWFCLAIRKDY	EPEE			IIVSV	NATKFDWENNE	PCQVDNK	SYENSTLVRL
IFGNI	PQNRERIYIVGFRN			I	EHYKNFNF	PMPQP	LTLTIKDM
IFN-VPQNRERLYIIGI	REDLIKNEE		WSLDF	KRKDILQK	GKQRLVELDI	KSFNFRWTAQ	SAATKRLKDL
IFG-IPQNRERVF	CISILN			PNED	FTFPQKQ		NLTL
IFG-SSQARRRVFMI	STLNEF			VELP	KGDKPKS		IKKVLNKI
IFG-IPQNRERIYLV	GVF			LNHD	VDFRFPQP		IGQATAVGD
IFG-LPQNRERIYIV	GFDRKS			ISNY	SDFQMPTP		LQEKTRVGN
IFG-VPQNRERIYIV	GFNKEK			VRN	HEHFTFPTP		LKTKTRVGD
IFG-VPQNRERIYIV	GFHKS			TGVNS	SFSYPEP		LDKIVTFADI
IFQ-VPQNRRLRVYIV	GLDQSQPELT			ITSH	IGATDSHKFKQL	SNQASLFD	TNKIMLVRDI
IFG-VPQNRVRIYIL	GLILGSKPKLT			LTSN	VGAADSHKYK	NEQISLFD	ESYATVKDI
IFG-IPQKRKR	FYLVAF	LNQN		I	HFEF	PKP	PMISKDIGEV
IFG-LPQRRERIVIV	GFHPDLG			INDF	SFPKGN		PDNKVPINA
YIG-IPQKRERIYMI	CFRNDLN			I	QNFQ	PKP	FELNTFVKDL
YIG-NAQRRRRVFIF	GYKQDLNYSKAME		ESPLDKI	IYH	NGLFAEAFP	IEDYANKNR	VNRTHITHDIVDI
YIG-TPQRRKRAIIR	LNKKGTIWN			L	PLKQNI	VSVEQ	AIGNLPSI
IGG-TPQVRERVFIT	ATLVPERMRDER	IPRTETGE	DAEA	IGPK	PVATMNDRFP	IKGGTEL	FHPGDRKSGWNLLTSGI
YIG-VAQNRDRVFI	IGIQKLGVPD			FSF	PEYSESEQRLYD	ILDNLQTPSII	PESLPIQRNL
IFG-VAQNRDRVFI	VGIIQKLDLNG			FSF	PEYAFSDQRI	YHII	DNI
IYD-VAQKRERIVI	IGIREDLVK			E	KYPFR	FPLAQ	VYKPVLKDV
YIG-VSQLRPRVLF	VALKNEYTN			FFK	WPEP	NSEQPK	TVGELLFDL



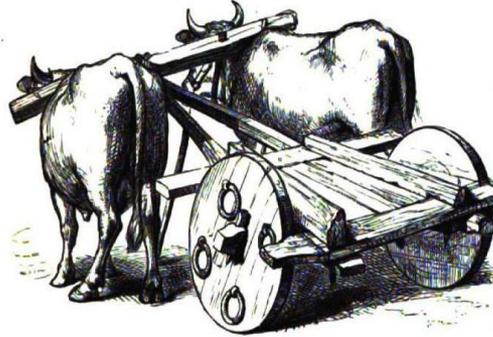
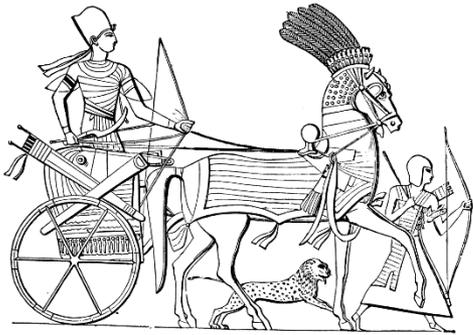
Выравнивание другой программой и раскраска по а.к.о.

	10	20	30	40	50	60	70	80
- GVPQDRKR VFI - - - - - VGY - - - - - REDL - - - - - NL - KFEFPKP - - - - LN - - KKVTLRDA I GDL								
- GVAQDRER VFY - - - - - VGF - - - - - RKDL - - - - - NISNFEFPYP - - ISEK - - ERKYLKDS I WDL								
- GVAQDRKR VFY - - - - - IGF - - - - - RKDL - - - - - EI - KFSFPKGSTVEDK - - DKITLKDVI WDL								
- GVAQDRKR VFY - - - - - IGF - - - - - RKEL - - - - - NI - NYLPP I PHLI - - - - - KPTFKDVI WDL								
- GI PQQRDR LVL - - - - - FAA - - - - - KQGV I - - - - - - - KI I PPTH - - - - - TPENYRTVRDVI GSL								
- GVPQSRQR VFF - - - - - IGL - - - - - KSDR - - - - - PLNQQIL T PPSK - - VIES - - EYTSLEEAI SDL								
- GVPQLRKRT FV - - - - - IGH - - - - - RHGSI - - - - - - - ADLANVLQ - - - - - QRLAKQSL T VRDY								
- GVPQSRTR FSL - - - - - IGK - - - - - LNSEH - - - - - - - NFL IPTLS - - - - - RKLSDKPMT VRDY								
- GVPQRRHR I I I - - - - - VGI - - - - - RKDQ - - - - - - - DV - AFRVPEP - - THKE - - KYRTASEALADI								
- GAHHQRHRWFC - - - - - LAI - - - - - RKDYEPEE I I - - - - - - - VSVNATKFDWENNE - PPCQVDNKSyenSTL VRL								
GNIPQNRERI YI - - - - - VGF - - - - - RNIE - - - - - - - HYKNFNFPMP - - - - - - QPLTLTIKDM								
- NVPQNRERLYI - - - - - IGI - - - - - REDL I KNE - - EWSLDFKRKDI LQKGKQRL VELD I KSFNFRWT - - - - - AQSAATKRLKDL								
- GI PQNRER VFC - - - - - ISI - - - - - LNPNE - - - - - - - DFTFPQK - - - - - - QNLTLSMNDL								
- GSSQARRR VFM - - - - - IST - - - - - - - - - - - - LNEFVELPKGD - - - - - - KKPKSIKKV								
- GI PQNRERI YL - - - - - VGF - - - - - LNHDV - - - - - - - DFRFPQP - - - - - - I GQATAVGD I								
- GLPQNRERI YI - - - - - VGF - - - - - DRKSI S - - - - - - - NYSDFQMPTP - - - - - - LQEKTRVGN I								
- GVPQNRERI YI - - - - - VGF - - - - - NKEKVR - - - - - - - NHEHFTFPTP - - - - - - LKTKTRVGD I								
- GVPQNRERI YI - - - - - VGF - - - - - HKST - - - - - - - GVNSFSYPE - - - - - - PLDKIVTFADI								
- QVPQNR LRVYI - - - - - VGLD - - QSQPELTITSHI - - GATDS - - - - - HKFKQ - - - - LSNQASL - - FD - - - - - TNKIMLVRDI								
- GVPQNRVRI YI - - - - - LGIL - - GSKPKLTLTSNV - - GAADS - - - - - HKYK - - - - - NEQISL - - FD - - - - - ESYATVKDI								
- GI PQKRKR FYL - - - - - VAF - - - - - LNQNI - - - - - - - HFEFPKP - - - - - - PMI SKDIGEV								
- GLPQRRERI VI - - - - - VGF - - - - - HPDL - - - - - - - GINDFSFPK - - - - - - GNPDNKVP I NA I								
- GI PQKRERI YM - - - - - ICF - - - - - RNDL - - - - - - - NIQNFQFPKP - - - - - - FELNTFVKDL								
- GNAQRRR R VFI - - - - - FGY - - - - - KQDLNYSKAMEESPLD - - KIIYHN - - - - - GLFAEAFPI EDYANKNRVNRTHI THDI VDI								
- GTPQRRKRAI IRLNKKGT - - - - - IWNL - - - - - - - PLKQNI V - - - - - - SVEQAIGNL								
- GTPQVRER VFI - - - - - TATLVPERMRDER I PR TET - - GEIDA - - EAIGPK - - - - - PVATMNDRFP I KK - - GGTELFHPGDRKSGWNL								
- GVAQNRDR VFI - - - - - IGI - - - - - QQKL - - - - - - - GVPDFSFP EYSESEQRLYDI LDNLQTPS I I								
- GVAQNRDR VFI - - - - - VGI - - - - - QQKL - - - - - - - DLNGFSFP EYAESEQRLYHI LDNLEAPETK								
- GVAQNRDR VFI - - - - - VGI - - - - - QQKL - - - - - - - DLNGFSFP EYTESEQRLYHI LDNLEVPETK								
- DVAQKRERI VI - - - - - IGI - - - - - REDLVKE - - - - - - - QKYPFRFP LA - - - - - - QVYKPV LKDV								
- GVSQLRPR VLF - - - - - VAL - - - - - KNEY - - - - - - - TNFFKWPEP - - - - - NSEQPKTVGELLFDL								

Консервативное значит важное

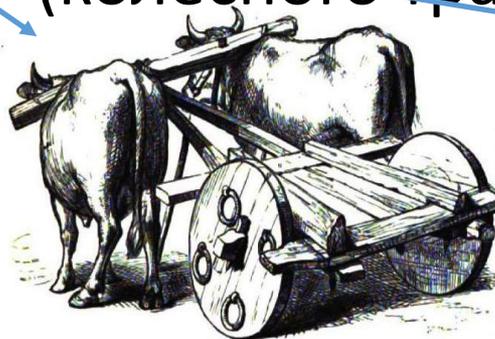
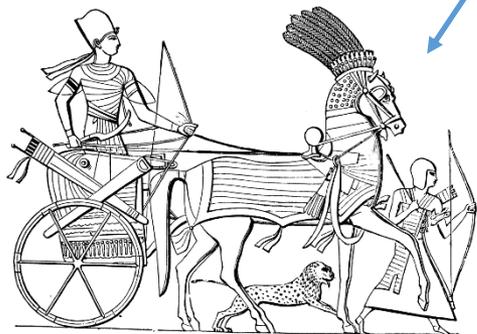
Консервативное - то, что длительно существует в эволюции, с несущественными изменениями

26



LUCA

(колесного транспорта)

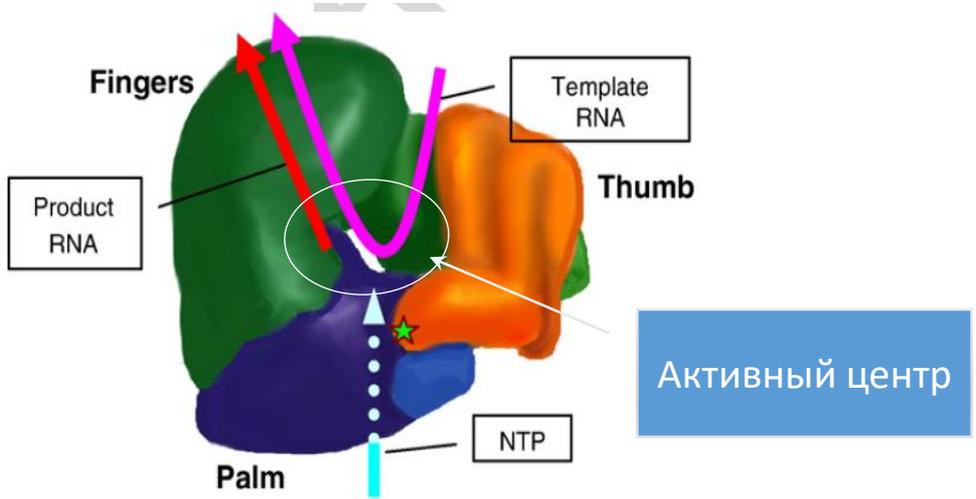


В выравниваниях белков – то же самое:
Сохраняющееся в эволюции (консервативное) – важно

РНК зависимая РНК полимераза (RdRP), консервативные участки

```

*      320      *      340      *      360      *      380      *      400      *      420      *      440      *      460
FKTMRIRFGDVGDLDDFSAFADASLSPFMIREA..GRIMSELS...GTPSHFGTALINTIIYSKHLIYNCCY.....HVCGSMPSGSPCTALNSTINNVLYYVFSKIFGKSPVFF.....CQALKILC.YGDDVIVFSDRV
EVAMQG.FERVYDVIDYSNEDSTHVSAMFRLL..A...EEFF.TPENGFDPLTREYLESLAISTHAFEEKRF.....LTGGLPSGCAATSMNTIMNNIIRAGLYLTYKNFEFDD.....VKVLS.YGDDLIVATNYQL
ETHFAQ.YKNVWDVLYSADANHCSDAMNMFEEVFERTEFG.....FHPNAEWILKTLVNTTEHAYENKRI.....VVEGCMPSGCSATSIINTILNNIYVLYALRRHYEGVELDT.....YTMIS.YGDDIVVASYDYL
.....WSLCVATIVSDHDTFWPGWLRDLICDELINMGYA.PWVVKLFETSLKLPVYVGAFAPEQGHLLGDPSNPDLVGLSSGQGATDLMGTLIMSTIYLVMLQDHTAPHLNSRIKMPSACRFLDSYWQGHEETROIS.KSDDAILGWTKGR
LRLRLE.NWVYCDADGSOEDSSSLTPYLINAV..LTLRSTYMEDWDVGLQMLRNLYTEIVYTPISTPDGTIV.....KKFRGNNSGQPSIVDNLSLMVVIAMHYALIKECFEVEEID.....STCVFFV.NGDDLIAVNPEK
HDKLNRPGLWLGSGDGRDSSIDPFFFDVV..KTKRKHEL..PSEHHRAIDLIIYDEILNPTTICLANGMVI.....KKNVGTQR.QPSTVDNTLVMITAFLYAYIHKTDGRELAL.....LNERFIFVC.NGDDNKFAISPQF
AISLASFSYPYGFNCFANEDGMFHPSSFSMV..SELANIFY...GNFLSTERDNLTRMLTNRFSLMKGAIL.....RVPGGPGSGFFMTVFNSEINLFYLQSAWIMLARFNGRQDISH.....PCNFPKYVRACV.YGDDNIVAIMEV
AARMKEKGNVDVLCODYSSEFDGLLSKQVMDVI..ASVINELC.GGEDQLKNARRNLMACCSRIAICKNTVW.....RVECGIPSGFFMTVFNSEINLFYLRHYHKIMREQQAPELMV.....QSFDKLIIGLVT.YGDDNLSVNAVV
YAEHAK.YKNHFDADYIANDSTQNRQIMTES..FSIMSRLT...ASPELAEVVAQDLLAPSEMDVGDYVI.....RVKEGLPSGFFPCTSQVNSINHWITLICALSEATGLSPDVV.....QMSYFYSFYGDDEIVSTDIDF
NNLTSKASDFLCLDYSKFDSTMSPCVVRIA..IDLADCC...EQTELTKSVVLTILKSHFMTILAMIV.....QTKRGLPSGMPFTSVINSICHWLLWSAAVYKSCAEIGLHCS.....NLYEDAPFYT.YGDDGVYAMTEMM
IQRIKS.AAKVYAVDYSKWDSTQSPRVSAAAS..IDLRYFS...DRSPIVDSAANTLKSPPIAIFNGVAV.....KVSSGLPSGMPFTSVINSINHCLYVGCAILQSLEARGVPVTW.....NLFSTFLMMT.YGDDGVYMFPMFM
TKRLERPKHDRYCVLYSKWDSTQPPKVTSSQS..IDILRHFT...DKSPIVDSACATLKSNIPIGIFNGVAF.....KVAGGLPSGMPFTSVINSINHCLMVGSAVVKALEDSGVRVTW.....NIFDSMDLFT.YGDDGVYIVPPLI
D      D      g      sg      T      n3      gDD
    
```



На каких участках выравнивание правильное – совпадает с эволюционным?

Множественное даёт аргументы, опровергающие оптимальное парное выравнивание. Пример.

```
      *           100           *           120           *           140           *           160
THEIE_LACLS : AGVSEFIVNDDVELARELNADGHIHGQTDSEVSKVREKVGQEMWLGLSVTKADELKTAQ-SSGADYLGIGPIYPTNSKND : 14
THEIE_MANSM : FQVBEFIVNDDVELALSIQADGHIHVGQKDTAVETILRNTRNKPIIIGLSINTLAQALANKDRQDIDYFEGVGPPIFPTNSKADH : 15
THEIE_STRA3 : YQVBEFIIDDDIDLVELIDADGHIHGQNDLPVDEARRRLPDKI-IGLSVSTMAEYQKSQ-LSVVDYIIGIGPFNPQSKADA : 14:
THEIE_LISIN : YQVBEFIINDDDVALALEIGADGHIHVGQNDDEEIRQVIASCAGKMKIIGLSVHVSVEAEEAERLGSVDYIIGVGPPIFPTISKADA : 14:
THEIE_ANOFW : YNIBEFIVNDDVDLALALQADGVHVGQDEVEAERVDRDRIGDKY-LGVSVHNLNEVKKAL-AACADYVGLGPIFPPTVSKEDA : 14:
THEIE_GEOTN : YGVBEFIVNDDVELAIAIDADGVHVGQDDEADARRVREKIGDKI-LGVSAHNVVEEARAAI-EAGADYIIGVGPPIYPTRSKDDA : 14
THEIE_BACSU : AGVBEFIVNDDVELALNLKADGHIHGQEDANAERVRAAIGDMI-LGVSAMTSEVVKQAE-EDGADYVGLGPIYPTETTKDDT : 14
THEIE_BACA2 : AGIBEFIVNDDVELALRLEADGVHIGQDDADAEEETRAAIGDMI-LGVSAMTSEVVKQAE-AAGADYVGMGPVYPTETTKDDT : 14
THEIE_OCEIH : FQIBEFIVNDDVDLAKQLDADGHIHGQDDQPVVVRKQFENKI-IGLSISTNNELNQSP-LDLVDYIIGVGPPIFDTNTKEDA : 14
THEIE_STAAB : YNVBEFIVNDDVSLAKEINADGHIHVGQDDAKVKEIAQYFTDKI-IGLSISDLGEYAKSD-LTHVDYIIGVGPPIYPTPSKHDA : 14:
THEIE_STACT : YNVBEFIVNDDVALAEEIDADGHIHVGQDDEAVDDFNRRFEGKI-IGLSIGNLEELNASD-LTYVDYIIGVGPPIFATPSKDDA : 14:
      6pFI61DD6 La 6 ADG6H6GQ D 6G6S 2 DY G6GP pT 3K Da
```

```
      *           180           *           200           *           220           *           240
THEIE_LACLS : AKETIGIKDLR-LMLLENQLPIVIGGGITQDSLTELSAIGLDGLAVISLLTEAENPKKVAQMIRQKITKNG~~~~~ : 21
THEIE_MANSM : SPIVGMNFIRQIRQLGIDKPCVAIGGITEESAAILRRLGADGVAVISAIHSHSVNIANTVKTLAQK~~~~~ : 22
THEIE_STRA3 : KPAVCNRTTKAVREINQDIPVVAIGGGITSDVFVDIIESEGADGLAVISAIISKANHIVDATRQLRYEVEKALVNRQKRSDVI : 22:
THEIE_LISIN : EPVSGTAILEEIRRAGIKLPIVIGGGITNETNSAEVLTAGADGVSVISAITRSEDCQSVIKQLKNPGSPS~~~~~ : 21
THEIE_ANOFW : KQACGLTMEHIRAEKRVPLVAIGGITEQTAQVIEAGADGLAVISAIKRAEHIYEQTKRLYEMVMRAKQKQKDR~~~~~ : 21
THEIE_GEOTN : NEAQQPGILRHLRREQGITIPVVAIGGGITADNTRAVIEAGADGVSVISAIASAPEPKAAAAALATAVREANL---R~~~~~ : 22
THEIE_BACSU : RAVQGVSLIEAVRRQGISIPVIGGGITIDNAAPVIEAGADGVSMISAISQAEDPESAARKFREEIQTYKTG--R~~~~~ : 22:
THEIE_BACA2 : EAVQGVTLIEEVRQGITIPVIGGGITADNAAPVIEAGADGVSMISAISQAEDPKAAARKFSEEIRRSKAGLSR~~~~~ : 22
THEIE_OCEIH : KTAVGLEWISLKKQHPSLPLVVAIGGGINTTNAQEIIEAGADGVSVISAITETDHIHQAVQRL~~~~~ : 20
THEIE_STAAB : HTPVGPMEIATFKEMNPQLPIVVAIGGGINTSNVAPIVEAGANGISVISAIKXSENIKTVNRFKDFFN~~~~~ : 21:
THEIE_STACT : SEPVGPKMIETLRKEVGDLEIPIVVAIGGITSLDNVQEVAKTSADGVSVISAIARSPhVTETVHKFLQYFK~~~~~ : 21:
```

```
      80           *           100           *           120           *           140           *
THEIE_LACLS : VSEFIVNDDVELARELNADGHIHGQTDSEVSKVREKVGQEMWLGLSV-TKADELKTAQSSGADYLGIGPIYPTNSKND :
THEIE_MANSM : VEFIVNDDVELALSIQADGHIHVGQKDTAVETILRNTRNKPIIIGLSINTLAQALANKDRQDIDYFEGVGPPIFPTNSKAD :
      V FIVNDDVELA 6 ADGIH6GQ D V 6 6GLS6 T A L DY G6GPI5PTNSK D
```

```
      160           *           180           *           200           *           220
THEIE_LACLS : AAKPTG---TKDLRLMLLENQLPIVIGGGITQDSLTELSAIGLDGLAVISLLTEAENPKKVAQMIRQKITKNG : 218
THEIE_MANSM : HSPVGMNFIRQIRQLGIDK--PCVAIGGITEESAAILRRLGADGVAVISAIHSHSVNIANTVKTLAQK----- : 220
      6G T4 6R 6 6 P V IGGI 2 S L 6G DG6AVIS 63 N 6 QK
```

В красном овале во множественном выравнивании – одна делеция между консервативными позициями.

В оптимальном парном выравнивании первых двух последовательностей в красном овале – четыре делеции. Участки те же.

III. Домены белков

ЭВОЛЮЦИОННЫЙ ДОМЕН – достаточно длинный (более многих десятков а.к.о.) участок предкового белка, который эволюционировал только по типу локальных мутаций. При этом белки-потомки могли претерпевать крупные перестройки, не затрагивающие домен. Домену дают название. Собирают представителей домена из всех белков в которых их удаётся найти и строят выравнивание.

В занятиях работаем с базой данных PFAM – protein families. Семейства – это семейства ДОМЕНОВ. Pfam в прошлом году была поглощена консорциумом InterPro, посвященном семействам белков

В белке может быть один домен, два или много.

Эволюционные домены в белке изображают так

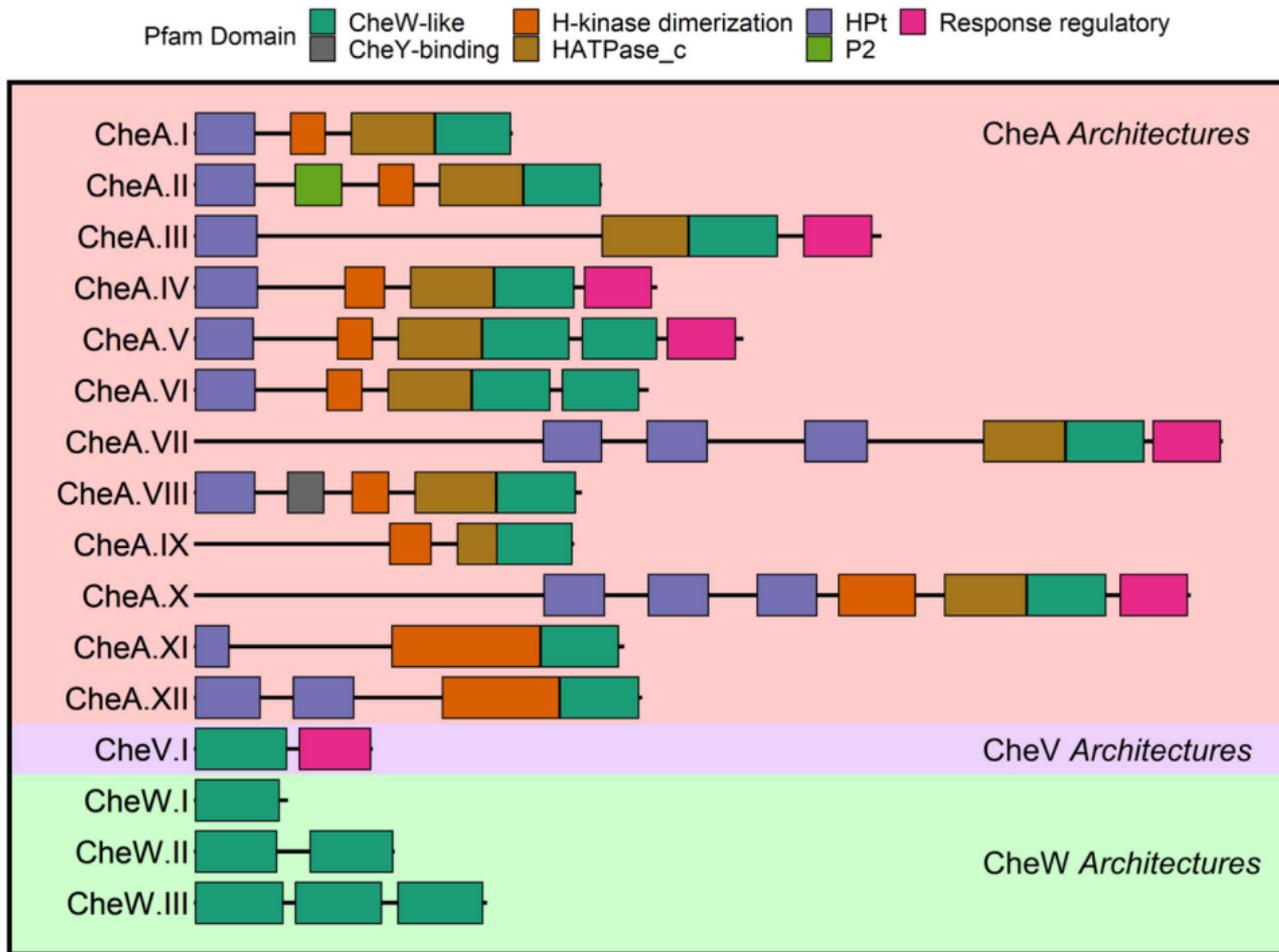
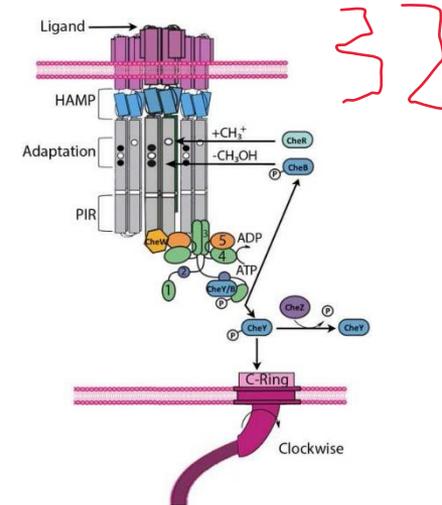


FIGURE 2. Schematic of the most abundant CheW-containing domain Architectures seen in nature.

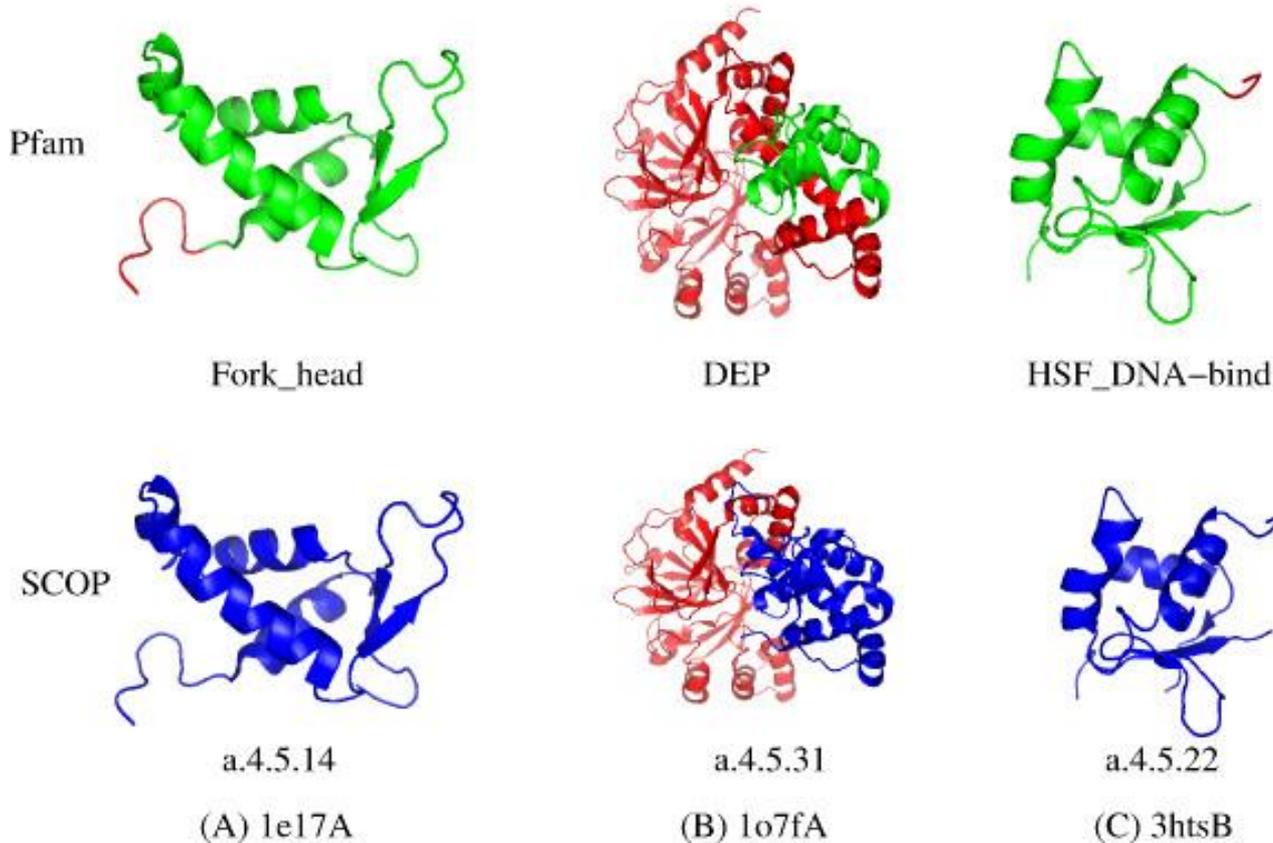
Sixteen distinct architectural variants containing the CheW-like domain were identified



Che белки участвуют в хемотаксисе бактерий

Доменные архитектуры белков, содержащих домен CheW-like

Домены Pfam часто, но не всегда соответствуют структурным доменам SCOP



Такое соответствие наблюдается не для всех доменов Pfam

Examples of one-to-one exact mapping between Pfam families and SCOP domain families. The domains are graphed onto the PDB structures of their corresponding member proteins using Pymol. The first row shows Pfam domains and the second row shows their corresponding SCOP domains. The structure regions of Pfam domains are marked in green and those of SCOP domains are marked in blue. Red regions lie outside the SCOP or Pfam domains.

Эволюционные домены

- Имеют определенную функцию (не всегда известна)

DUF – Domain of Unknown Function

- Часто совпадают со структурными доменами (но не всегда)

Гомеодомен – ДНК связывающий домен

Homeodomain proteins regulate gene expression and cell differentiation during early embryonic development, thus mutations in homeobox genes can cause developmental disorders.^[1]



Наблюдаемый результат крупных перестроек генома

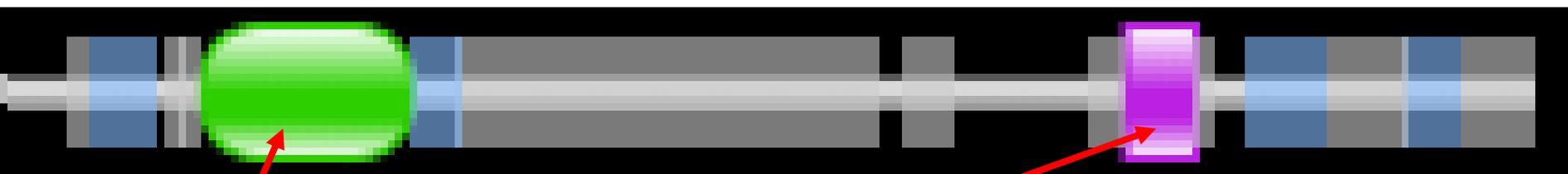
Часто белки с одинаковой доменной архитектурой имеют сходство последовательностей в границах доменов, а вне них сходство не детектируется или его НЕТ. Участки белков не гомологичны, а что происходило в эволюции – неизвестно.

ДВА ДОМЕНА гомеобелков: гомеодомен и OAR домен

SW: PMX1_CHICK/1	-----MASSYAHAMERQALLPARLDGPACLDNLQAKNFSVSHLLDLEEAG-DMVAAQDCEGGGPRGRSLLLESP-GLTSGSDTPQQD	: 80
SW: PMX2_HUMAN/1	-----MDSAAAAFALDKPALGCPGPPPPALGCPGDCQAQRNFSVSHLLDLEEVAAAGRLAARPGARABAREGAAREPSCGSSGSEAAAPQD	: 86
SW: PMX1_HUMAN/1	-----MTSSYGHVLEQPALQGRDLSPGMLDTLQAKNFSVSHLLDLEEAG-DMVAAQADENVGACRSLLLESP-GLTSGSDTPQQD	: 80
SW: ARX_BRARE/1	ISQAPQVVISRSKSYREN-APFSQS---D-EGQSP--EHMAQLVELST-----LKFEEDEVVKEEACQDN-----S-----LSPKDEESLH-NDGVDKCDSDSVCLS	: 84
SW: ARX_MOUSE/1	ISQAPQVVISRSKSYRENGAPVFPVPPALD-ELSGPCGVHPEERLSAASGPGSAPAAGCGTCAEDDEEELLEDEEEDEREELLEDDDEELLEDDARALLKEPERRCVATTCTVAAAAAATAVATEGGELSPKRELLLHPEDARCKDGEDSVCLS	: 157
SW: AL_DROME/1-1	-----MGISEEIKLEELPQAKLHAHPDAVVLVDRAPGSSAASAGAAALTVSMVSYSGAPSCASGASGCTNSPVSVDGNS	: 72
SW: ALX4_MOUSE/1	-TFLSAGAKQCPCDAKSRARYGACQDLaAPLESSSGARGSPNKFQPPQPTQP-----PPAPPAPPAHLYLQRCACKTPDQCSLKLQEGSSGCHNAALQVPCYAKRESNLCEPELPPDSFVPCVMDNSYLSVKRTGARCPQDASARIPSP	: 145
SW: ALX4_HUMAN/1	-TFLSAAAQAQPCDAKSRARYGACQDLaTPLESAGARGSPNKFQPPQPTQPQPPQPPQPPQPPQPPQPPHLYLQRCACKTPDQCSLKLQEGSSGCHNAALQVPCYAKRESNLCEPELPPDSFVPCVMDNSYLSVKREAGVRCQDRASSDLPSP	: 157
SW: RX2_CHICK/1	-----NPSRLHSIEAILGFTKDDGLLGFQFP-----DGGAGSAAEAADKRGPRHCLPKGPAEPPPAEHQGRFQEPYPCGASAPF-----LPAGCGDG	: 83
SW: RX2_BRARE/1	-----GISCRVHSIDVILGFSKDDPPLLEPSGR-----HKVDLEDQLEEQEKQVADPYSHLQIPDQIQQQQSVYH---DTGLFSTDKCADLGDPRSINVEDSRS	: 92
SW: RX1_XENLA/1	-----NPSRLHSIEAILGFKEDS-VLGSFQSEIISPRNAKEVDKRSRHLHMTREIHPQEHLEDG-QADCYG--DPYSGRTSSECLP-PCGST--SNSDN	: 91
SW: RX_HUMAN/1-1	-----STSRLHSIEAILGFTKDDG-ILGTFPAERGARGAKEBRRLGARCPACPKPAEREGSEPSPPAPAPAPYEAPRPPYCPKPEWRAPSPGLPVGPAATGEA	: 97
SW: PIX2_BRARE/1	-----MTSHKPLSLDHHHHHHVTCGSHAPLSMASSLQLPQRSVDSKHLRDLVHTVSDTSSPEVKEKRCQ--	: 66
SW: PIX2_HUMAN/1	-----METNCRKLVSAVGLQVPAAEVCLFSKDSSEIKKVFETDPSRKRKAASAKFPFPHQPCANEKRSKQ--	: 68
SW: PIX1_HUMAN/1	-----MDAFKGMSELRLEPGFPPPPPHDMGPAFHLLRAPDPRPLEN-SASESSDTELEPEKRGCEP	: 64
SW: OTP_MOUSE/1	-----MLSHADLLDARLCHKDAEALLGHREAVKRLGVGCSDPGCHPCDLAPNSDPVREGATLLPREDITTVGSTPASLAVSAKDPDKQPGPQCGP	: 90
SW: PMX1_CHICK/1	NDQLNSEE-----KKKRRRRRRTFTFNSSQLQALERWERETHYDPAFVRBDLARRVNLTEARVQVVFQMRRAKFRNNEAMLSKMASLLKSYSGDVTAVEQIIVPRPAPRPTDYLWGTASPYSAMATYSTTCTMAS-----	: 213
SW: PMX2_HUMAN/1	GCPCSPGRCG-----AAKRRRRRRTFTFNSSQLQALERWERETHYDPAFVREELARRVNLTEARVQVVFQMRRAKFRNNEAMLSRSASLLKSYSG-ATEQVPAPRPTALSPTDYLWGTASSPYSTVPPYPCSSGCP-----	: 221
SW: PMX1_HUMAN/1	NDQLNSEE-----KKKRRRRRRTFTFNSSQLQALERWERETHYDPAFVRBDLARRVNLTEARVQVVFQMRRAKFRNNEAMLANKNASLLKSYSGDVTAVEQIIVPRPAPRPTDYLWGTASPYSAMATYSATCANNS-----	: 213
SW: ARX_BRARE/1	AGSDSEEG-----HLKRRRRRRTFTFTSYQLEELERAFQTHYDPAFVREELARRLDLTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAPPSPPLPPPG-SASLPSCAPLG	: 230
SW: ARX_MOUSE/1	AGSDSEEG-----LLKRRRRRRTFTFTSYQLEELERAFQTHYDPAFVREELARRLDLTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAPPSPPLPPPG-SASLPSCAPLG	: 303
SW: AL_DROME/1-1	EKADSEY-----PKRRRRRRTFTFTSYQLEELERAFQTHYDPAFVREELARRLDLTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAPPSPPLPPPG-SASLPSCAPLG	: 212
SW: ALX4_MOUSE/1	EKTDSESN-----KCKRRRRRRTFTFTSYQLEELERAFQTHYDPAFVREELARRLDLTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAPPSPPLPPPG-SASLPSCAPLG	: 290
SW: ALX4_HUMAN/1	EKADSESN-----KCKRRRRRRTFTFTSYQLEELERAFQTHYDPAFVREELARRLDLTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAPPSPPLPPPG-SASLPSCAPLG	: 302
SW: RX2_CHICK/1	KPSDEEQ-----PKRRRRRRTFTFTSYQLEELERAFQTHYDPAFVREELARRVNLTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAPPSPPLPPPG-SASLPSCAPLG	: 215
SW: RX2_BRARE/1	PDIPDEDQ-----PKRRRRRRTFTFTSYQLEELERAFQTHYDPAFVREELARRVNLTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAPPSPPLPPPG-SASLPSCAPLG	: 225
SW: RX1_XENLA/1	KLSDEEQ-----PKRRRRRRTFTFTSYQLEELERAFQTHYDPAFVREELARRVNLTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAPPSPPLPPPG-SASLPSCAPLG	: 224
SW: RX_HUMAN/1-1	KLSEEQ-----PKRRRRRRTFTFTSYQLEELERAFQTHYDPAFVREELARRVNLTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAPPSPPLPPPG-SASLPSCAPLG	: 242
SW: PIX2_BRARE/1	SKNEDSW-----DDPSKRRRRRRTFTFTSYQLEELERAFQTHYDPAFVREELARRVNLTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAPPSPPLPPPG-SASLPSCAPLG	: 212
SW: PIX2_HUMAN/1	GRNEDVGA-----EDPSKRRRRRRTFTFTSYQLEELERAFQTHYDPAFVREELARRVNLTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAPPSPPLPPPG-SASLPSCAPLG	: 215
SW: PIX1_HUMAN/1	KCPEDSCAGCTGCCGADDPAKRRRRRRTFTFTSYQLEELERAFQTHYDPAFVREELARRVNLTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAPPSPPLPPPG-SASLPSCAPLG	: 218
SW: OTP_MOUSE/1	NPSQACQQ-----CQQQKRRRRRRTFTFTSYQLEELERAFQTHYDPAFVREELARRVNLTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAPPSPPLPPPG-SASLPSCAPLG	: 236
	k 4 4R RT Ft QL EL E R F 4 HYPD RE 6A L E R6qVWFQNRRAK54 e4	
SW: PMX1_CHICK/1	-----PAQGMNMANSLALPLAKRQYSLQRNQVPTVN-----	: 245
SW: PMX2_HUMAN/1	-----ATPCVMNMANSLALPLAKRQYSLQRNQVPTVN-----	: 253
SW: PMX1_HUMAN/1	-----PAQGMNMANSLALPLAKRQYSLQRNQVPTVN-----	: 245
SW: ARX_BRARE/1	LGTFLGTAAMFRHFAFIPTFCRLFFSSMCLPTSASTAAALLRQTAPPVSPVQSAALPEPPSSSSSTAADRRASSIAALPLAKRQYSLQRNQVPTVN-----	: 336
SW: ARX_MOUSE/1	LSTFLGAAVFRHFAFISPAFCRLFFSTMAPLTSASTAAALLRQTAPPVGAASGALADP-----ATAAADRRASSIAALPLAKRQYSLQRNQVPTVN-----	: 404
SW: AL_DROME/1-1	PPTSPASGAXPQVLQVIGIALTQQAASSLPT---QTSFVALTSHSPQRQLPSPHQAPPPPPRAATPPEDRRTSSIAALPLAKRQYSLQRNQVPTVN-----	: 313
SW: ALX4_MOUSE/1	DFL-----SVSGACSHVQTHMCSLFGAACISPLNGCYELNCEPDRKTSIAALPLAKRQYSLQRNQVPTVN-----	: 354
SW: ALX4_HUMAN/1	DFL-----SVSGACSHVQTHMCSLFGAACISPLNGCYELNCEPDRKTSIAALPLAKRQYSLQRNQVPTVN-----	: 366
SW: RX2_CHICK/1	LPASYTPPPFL-----NSPVTGHALQPLGAMGPPPPYQCGAFAVDFKPLDEGDPNRTSSIAALPLAKRQYSLQRNQVPTVN-----	: 290
SW: RX2_BRARE/1	LQPTYTAHPCFL-----NTSPGMHNTQKPL---PPPPYQVFPVNDKYLEEDV---RSSIAALPLAKRQYSLQRNQVPTVN-----	: 297
SW: RX1_XENLA/1	LPASYTPPPFI-----NPVSVGHALQPLGAMGPPPPYQCGAFAVDFKYLEEDV---RMSIAALPLAKRQYSLQRNQVPTVN-----	: 296
SW: RX_HUMAN/1-1	LPASYTPPPPPFL-----NSPPLGCPQLPL---APPPSYPCGCFGDKRPLDEADPNSSIAALPLAKRQYSLQRNQVPTVN-----	: 319
SW: PIX2_BRARE/1	SISMSMSMSMVPASVTCVPGSSL-----NSLNNLNLNSPNSLNSVPTPACPYAPPTPPY-VYRDTCNSSLASLPLAKRQYSLQRNQVPTVN-----	: 314
SW: PIX2_HUMAN/1	SISMSMSMSMVPASVTCVPGSSL-----NSLNNLNLNSPNSLNSVPTPACPYAPPTPPY-VYRDTCNSSLASLPLAKRQYSLQRNQVPTVN-----	: 317
SW: PIX1_HUMAN/1	SISMTMPSSMCPGAVPGMNSCL-----MNIN---MLTGSSLNSAMSGPCPYCTPASPYVYRDTCNSSLASLPLAKRQYSLQRNQVPTVN-----	: 314
SW: OTP_MOUSE/1	SQCSLAAGPPPNMCLNSLNSLNSGACGLQ---SHLYQAPFGMVPASLPGSPNSVSGSLQCLSSPDSVDRGTSIAALPLAKRQYSLQRNQVPTVN-----	: 325

Домены принято изображать так

[X1WJ92 ACYPI](#) [Acyrtosiphon pisum (Pea aphid)]
Uncharacterized protein (408 residues)



There are 1836 sequences with the following architecture:
Homeodomain, OAR

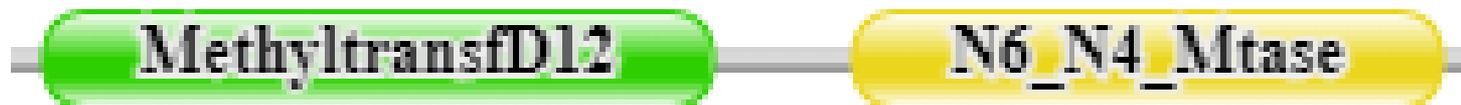
Как выровнять эти две последовательности?

25

There are 9 sequences with the following architecture:

MethyltransfD12, N6_N4_Mtase

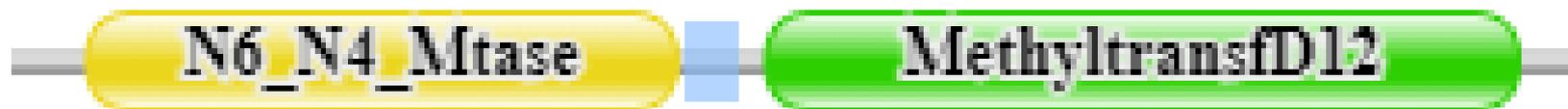
[A0A2Z5QVW5](#) [9MICC](#) [**D12-N6_N4**]



There are 5 sequences with the following architecture:

N6_N4_Mtase, MethyltransfD12

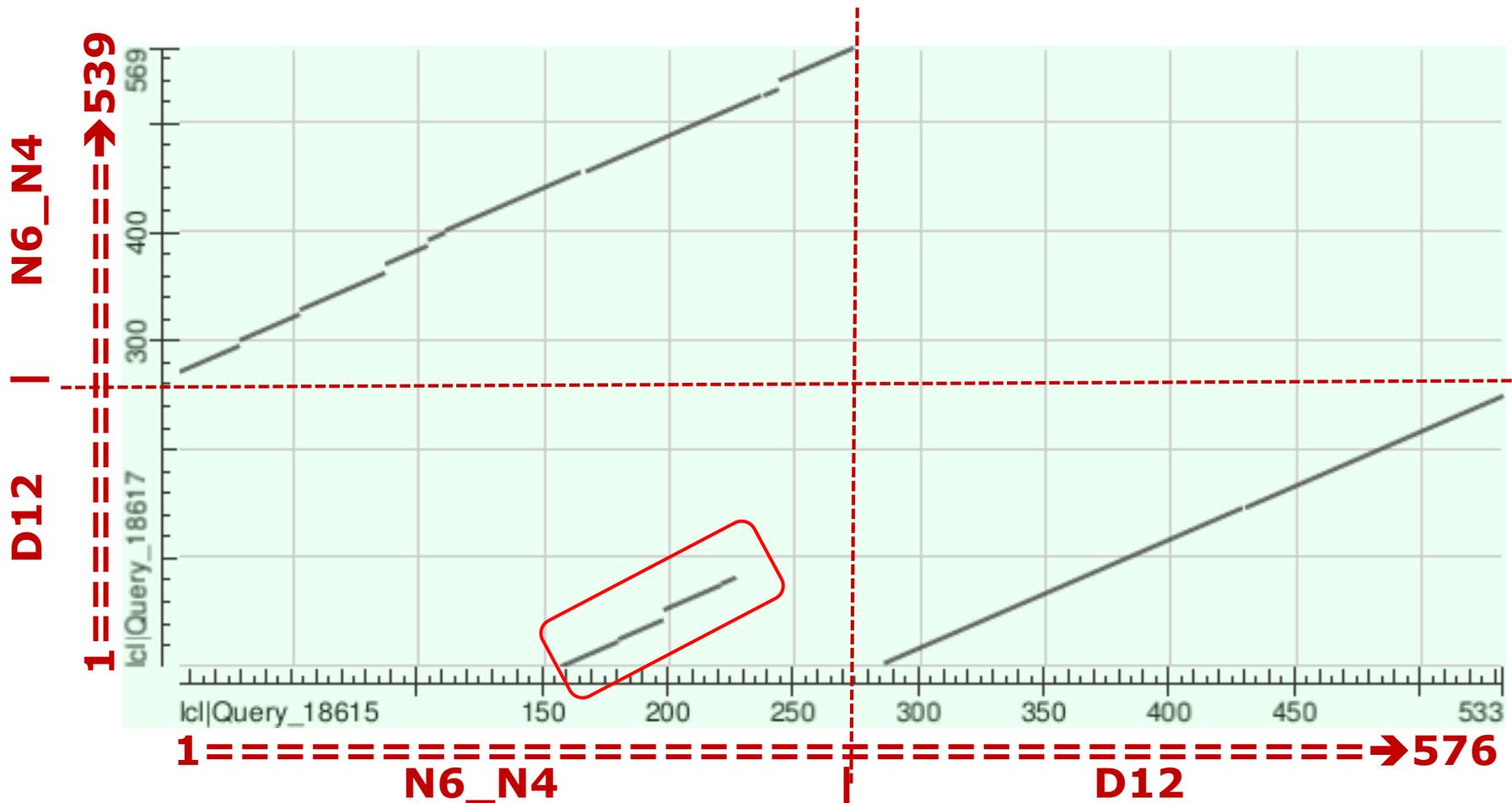
[A0A1I7GYG0](#) [9CLOT](#) [**N6_N4-D12**]



Как такое может возникнуть?

Лучшее парное выравнивание:
(выдача BLAST – набор парных выравниваний.

40



Программа BLASTp. Визуализация Dot Plot

Вопрос. Что значит лишняя диагональ ?

41

JaView – редактор выравниваний

Демонстрация параллельно выполнению упражнений

Список умений в Jalview

42

Действие	меню	подменю	варианты	комментарии
Импорт последовательностей	File	Add sequences	<ul style="list-style-type: none"> Из файла From textbox url 	
Команды относятся к выделенному	Ctrl-A			
Выравнивание	Web services	Alignment	Выбираете программу	Mafft быстро работает Результат в новом окне
Раскраска	Color	Clustal и By conservation	Conservation threshold	Меняя порог можно увидеть консервативные участки
Создание групп по выделенным колонкам	Select <hr/> Calculate	Make groups for selection <hr/> sort	by groups	Удобно для поиска гомологичных последовательностей Меньше колонок – больше группы
поиск	select	find	Пишете посл. Или паттерн	Пример [FY].[GA].{1,2}[GA] “.” – любая буква {от,до} раз [FY] – и F, и Y годятся

Действие	меню	подменю	варианты	комментарии
Выделение прямоугольного блока в отдельное окно	Мышкой выделяете прямоугольный блок	Правой кнопкой selection	Output to text block	Сразу new windows Текстовое окно можно закрыть
	Edit	Remove empty columns		После создание окна из выделенного блока
	Edit	Remove all gaps		Для перевыравнивания
Сократить выравнивание за счёт удаления почти идентичных склейки	Edit	Remove redundancy	Поставить порог сходства	Из высокосходных последовательностей оставить одну
Сохранение выравнивания				
Сохранение проекта со всеми окнами				

КОНЕЦ ПРЕЗЕНТАЦИИ

Но не занятия!

Упражнения

Выполняются в классе

Можно просить подсказать

- a. Заполните таблицу с информацией о домене:
 - a. Сколько всего белков с доменом (подсказка k = тысяча)
 - b. Сколько из них из SwissProt (reviewed)
 - c. Сколько из бактерий
 - d. Для скольких определена пространственная структура
 - e. Сколько в выравнивании seed, которое использовалось для поиска всех доменов в белках

- b. Сохраните выравнивание последовательностей домена C-5 cytosine-specific DNA methylase (PF00145) из SwissProt (revised)
 - a. Pfam => proteins => Generate Fasta => download (имя файла начните с вашей фамилии)
 - b. Откройте в Jalview
 - c. Выровняйте последовательности. Ужаснитесь!!!
 - d. Раскрасьте Clustal и by conservation
 - e. Найдите консервативные участки понижая порог conservation пока их не увидите.
 - f. Сделайте группы по консервативным позициям 2го участка
 - g. Сортировка по группам
 - h. Сколько последовательностей не имеют канонических консервативных а.к...
 - i. Сохраните – пока не знаю что и как.

1. Задание по теме "гомология и выравнивание"

Результат:

- **Тривиальная часть** - описание одного семейства по информации из Pfam
- **Нетривиальная** (самому или самой надо думать и принимать решения) - одно выравнивание двух подгрупп белков семейства с обоснованием их различий

Методы:

- Сервисы базы данных Pfam
- Редактор выравниваний Jalview
- Blast выравнивание 2х последовательностей, в формате Dot Plot
- Uniprot поиск и скачивание результата в табличном формате

1. Выберите семейство доменов из Pfam для анализа

От выбора зависит всё дальнейшее

Ограничения, направлены на то, чтобы обезопасить вас от больших технических трудностей, они не являются абсолютными

2. Опишите семейство доменов

Укажите число доменных архитектур с этим доменом

Выберите две достаточно представленные доменные архитектуры и укажите какие именно выбрали, их названия и число белков с каждой из них

Укажите число разных белков с доменом семейства, для которых известна 3D структура. *Разные структуры одного и того же белка (по Uniprot ID) считать за одну.*

Укажите число белков с доменом по таксонам самого высокого ранга. Типично - по суперцарствам(они же домены жизни) - бактерии, археи, эукариоты.

3. Постройте карту локального сходства (Dot Plot) двух белков из семейства, но с разной доменной архитектурой

Придумайте эволюционный сценарий наблюдаемого

4. В выравнивании семейства выделите на основании сходства две подгруппы доменов Pfam

В ответе - выравнивание, содержащее обе подгруппы и обоснование различий подгрупп

5. Сохраните таблицу со всеми белками из Uniprot семейства Pfam