

Гомология

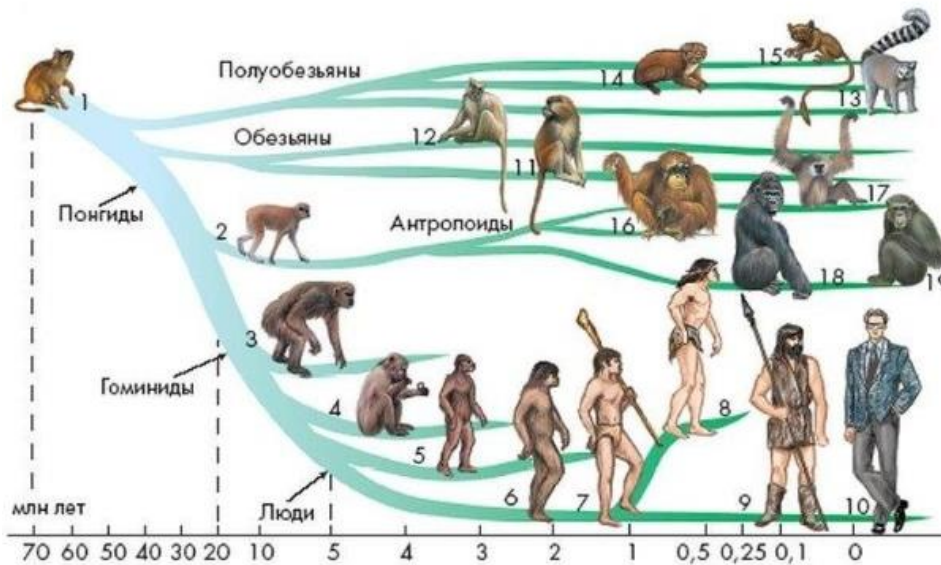
и

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

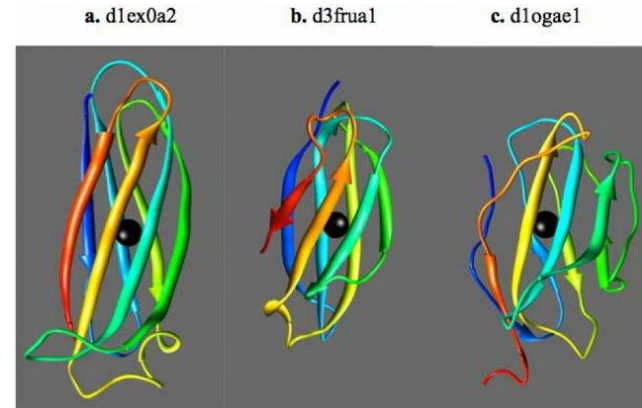
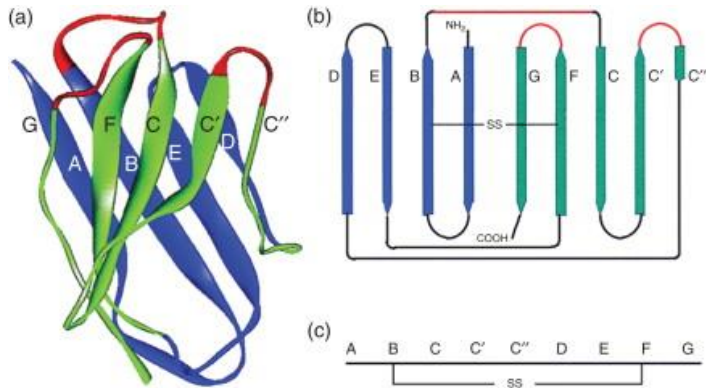
Гомология в молекулярной биологии

- 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

- 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

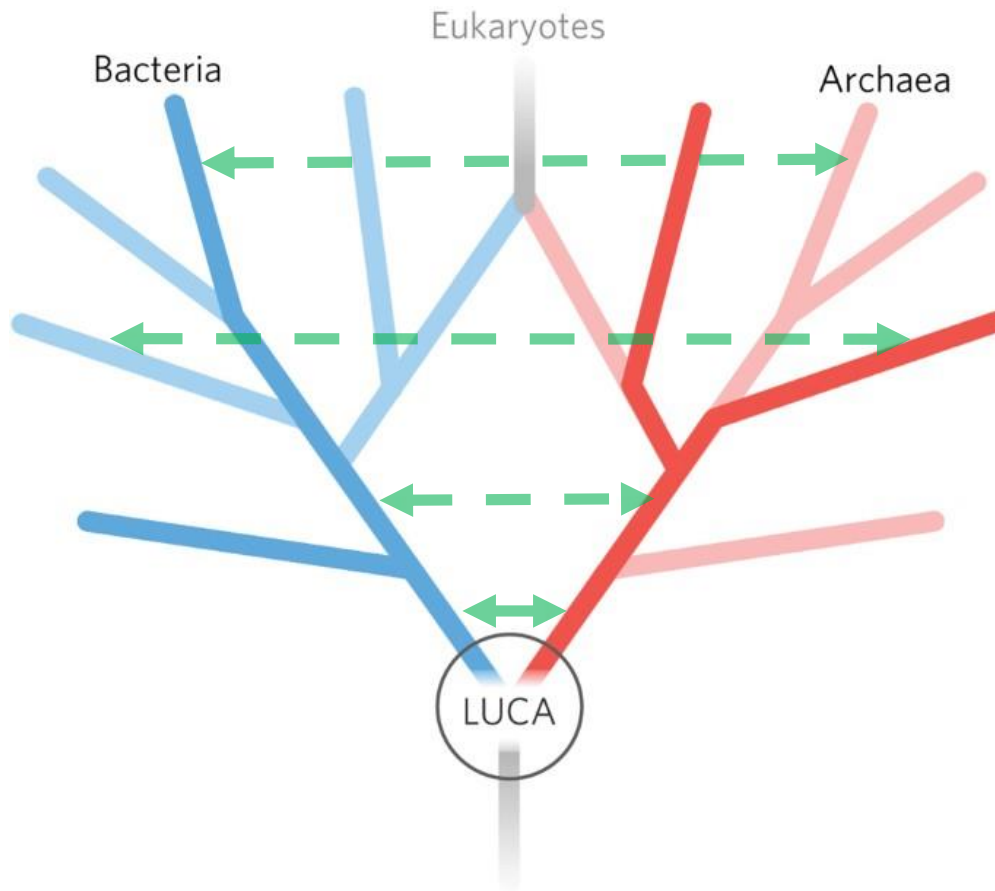
ImFold встречается во многих белках одного организма, связанных с иммунитетом и не только, и у всех животных

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.



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

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	

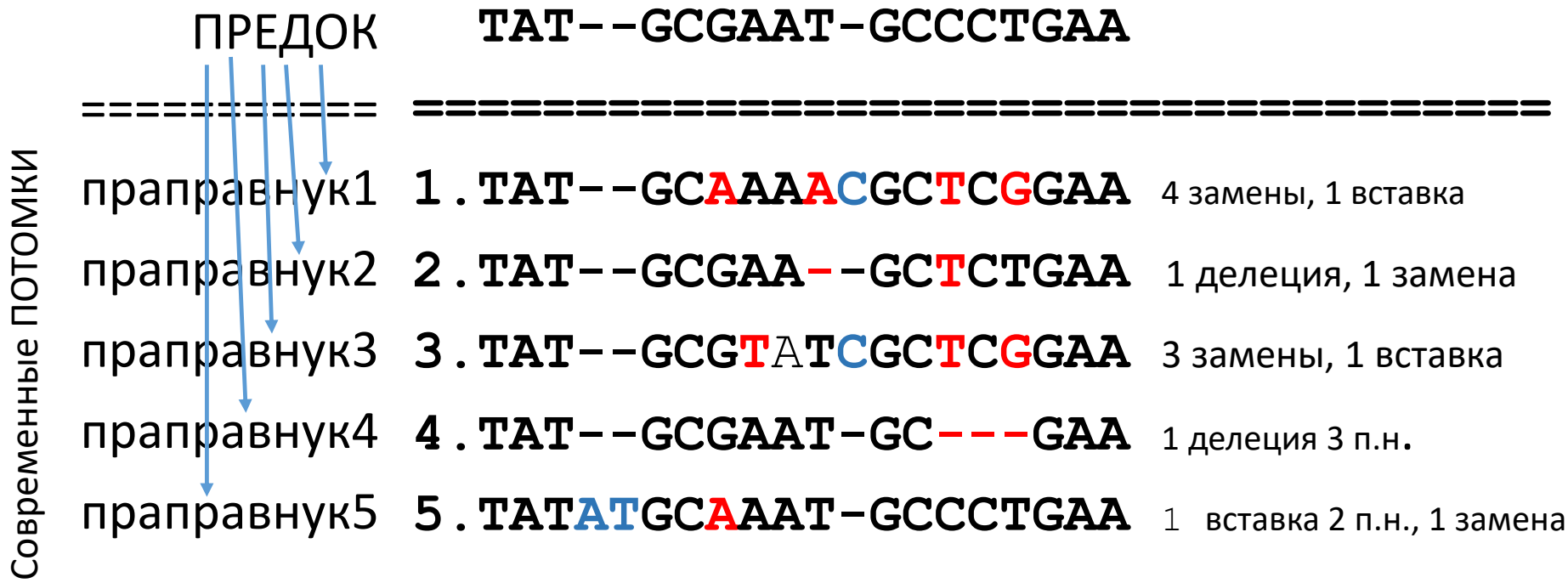
Нукл-ы потомка с номерами как у предка являются гомологами нукл-в предка

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.

Единоновременные 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)

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-LH DPR-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 GGPPCPDFSVGGKNKGHLGDN-GKLSASYIELICQNLPDFFLFENVKGLWRTTKHR
EDN77159.1 SLIGFI GGPPCPDFS IAGKNKGKDGDN-GKLSLSYTNLI IEMKPDFFLFENVKGLWRTARHR
SOD91684.1 EVSLVV GGAPCQPF SNIGK KLGKNDERNGDLFLEFVRMVKGIQPEAFIFENVVGI TQNKHSD
QCS48280.1 NVVGF I GGPPCPDFS IGGKNRGRQGDH-GKLSSESYIDLIIQHQPDFFI FENVKGLYRTKKHR
SMB95934.1 GLFGII GGPPCPDFSVGGKNRGGENGEQ-GRLSKVFDKIDLQPVFFLYENVPGLIRTAKHR
RUO38876.1 SPVGF I GGPPCPDFSVGGKNRGHEGEN-GR LTRTYVDGIIKYAPDFFI FENVKGLWRTKRHR
OIP70538.1 TIDLIC GGPPCQGFSTIGTND-KKDHR-NFLFFFLRMVETFKPNFII LENV TGLLAKKNES
AFY60915.1 NLVGFV GGPPCPDFS IGGKNKGQYGDN-GKLTKVYVDII IENQPDFFVFENVKGLWRTRSRHR
CUR30340.1 DLIGFI IAGPPCPDFSVGGKNRGKNGDQ-GKLTACYVELICQQRPDFFVFENVKGLWSTKKHR
TAK03971.1 QAALVV GGAPCQPF SNL GSKRGTADSR-GTLFQDFIRIVKGV RPKGFI FENVEGLTQDKHKG
AEE51071.1 KVALVV GGAPCQPF SNIGKKEGENDA KNGDLFLEFVRMVKGIQPEAFIFENVAGIIQSKHSG
RTR31666.1 RLVGFV GGPPCPDFSVGGKNKGSEGEN-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

```
200          210          220          230          240          250          260          270          280
YG-VPQDRKRVFIVGYREDLNLK-----FEFPKPLNKKVTLRD-----AIGDLPE-F
YG-VAQDRERVFYVGFVKDLNIN-----FE-FYPYISEKERKYLKD-----SIWDLKDNA
YG-VAQERKRVFYIGFRKDLEIKF-----SFPKGSTVEDKDKITLKD-----VIWDLQDTA
YG-VAQDRKRVFYIGFRKELNIN-----YLPPIPHLIKPTFKD-----VIWDLKDNF
YG-IPQQRDRLLVFAAKQG-----VIKIIPPTHTPENYR-----TVRDVIGSLATNY
YG-VPQSRQRVFFIIGLKSDRPLNQQ-----ILTP-----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-SSQARRRVFMI STLNEF-----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 FYLVAFLNQN-----IHFEFPPK-----PMISKDIGEVLES
FG-LPQRRERIVIVGFHPDLG-----INDFSFPKGN-----PDNKVPINAIL E
YG-IPQKRERIYMICFRNDLN-----IQNFQFPKP-----FELNTFVKDLLLP
YG-NAQRRRRVFI FGYKQDLNYSKAME-----ESPLDKI IYHNGLFAEAFP I EDYANKNR-----VNRTHITHDIVDISDN
YG-TPQRRKRAI IRLNKKGT IWN-----LPLKQNI VSVEQ-----AIGNLPSIESG
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						
<i>EJL77459.1</i>	GVDLVF	GGPPCQGF	SQIGMRR	-LDDER-	NELYQQYTR	IVAKLKP	RVFLMENV	PNLAL	LMNKG	GH		
<i>RXK67093.1</i>	DLDVVF	GGPPCQGY	SQIGTRR	-LDDER-	NELYLQYAR	IVEKQRP	RMFLMENV	PNMVLL	LNKGH			
<i>OJY44288.1</i>	NVDLVF	GGPPCQGY	SQIGTRD	-LH DPR-	NRLFEEFAR	VVATLKP	KLFLMENV	PNLLLL	LNKGH			
<i>TRU90449.1</i>	NPEMIV	GSPPCQDF	SSAGKR	NEGLGR	-ANLTLT	FAEIVTR	VSPQWF	VMENV	D--RI	EKSK		
<i>OXI46696.1</i>	GTDLVF	GGPPCQGF	SQIGMRR	-LDDER-	NELYKQYTR	VVSTLR	PRVFLMENV	PNLAL	LMNKG	GH		
<i>AVZ30243.1</i>	EIDVVF	GGPPCQGF	SLIGKRS	-FEDPR-	NSLVFHYI	RLVLELS	PKFFVI	ENVKGM	TAGNHQA			
<i>AFZ12381.1</i>	DIIGFI	GGAPCPDF	SVGGKNR	GSEGDK	-GKLSASY	IELICQQ	KPDFFL	FENVKGL	YKTKK	HR		
<i>HCQ21462.1</i>	HIIGFI	GGPPCPDF	SVGGKNG	HLDGN	-GKLS	SAYIELI	CQNLPDF	FLFENV	KGLWRT	TKHR		
<i>EDN77159.1</i>	SLIGFI	GGPPCPDF	SVAGKNG	KDGDN	-GKLS	S	YTNLI	IEMKPDF	FLFENV	KGLWRT	ARHR	
<i>SOD91684.1</i>	EVSLVV	GGAPCQPF	SNIGKKL	GKNDER	NGDLF	LEFVR	MVKG	IQPEAF	IFENV	VGITQ	NKHSD	
<i>QCS48280.1</i>	NVVGFI	GGPPCPDF	SVGGKNR	GROGDH	-GKLS	ESYIDL	I IQHQ	PDFFL	FENVKGL	YRTK	KHR	
<i>SMB95934.1</i>	GLFGII	GGPPCPDF	SVGGKNR	GENG EQ	-GRLS	KVFVDK	LDLQP	VFFLY	ENV	PGLIRT	AKHR	
<i>RUO38876.1</i>	SPVGFI	GGPPCPDF	SVGGKNR	GHEGEN	-GRLT	RTYVDG	I I KYA	PDFFL	FENVKGL	WRTKR	HR	
<i>OIP70538.1</i>	TIDLIC	GGPPCQGF	STIGTND	-KKDHR-	NLFF	FEFLRM	VETFK	PNFII	LENT	GLLAK	KNES	
<i>AFY60915.1</i>	NLVGFI	GGPPCPDF	SVGGKNG	QYGDN	-GKLT	KVYVDI	I I ENQ	PDFFL	VFENV	KGLWRT	RSRHR	
<i>CUR30340.1</i>	DLIGFI	AGPPCPDF	SVGGKNR	GKNGDQ	-GKLT	ACYVEL	ICQQR	PDFFL	VFENV	KGLWS	TKKHR	
<i>TAK03971.1</i>	QAALVV	GGAPCQPF	SNLGS	KRGTADSR	-GTLF	QDFIR	IVKGV	RPKGFI	FENVE	GLTQD	KHKG	
<i>AEE51071.1</i>	KVALVV	GGAPCQPF	SNIGKKE	GENDA	KNGDLF	LEFVR	MVKG	IQPEAF	IFENV	VAGI	IQSKH	SK
<i>RTR31666.1</i>	RLVGFI	GGPPCPDF	SVGGKNG	GSEGEN	-GKLT	RTYIDL	IVKDN	PDYFI	FENVKGL	WRTTR	HR	
<i>PTU64472.1</i>	NIDLVF	GGPPCQGF	SQIGTRR	-LDDER-	NELYKQYTR	IVKTLKP	RVFLMENV	PNLAM	MNKG	GH		

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

```
200      210      220      230      240      250      260      270
YVG-VPQDRKRVFIVGYREDLNLK-----FEFPPKPLNKKVTLRD-----AIGDL
YVG-VAQDRERVFYVGFGRKDLNISN-----FE-FYPYISEKERKYLKD-----SIWDL
YVG-VAQERKRVFYIGFRKDLEIKF-----SFPKGSTVEDKDKITLKD-----VIWDL
YVG-VAQDRKRVFYIGFRKELNIN-----YLPPIPHLIKPTFKD-----VIWDL
YVG-IPQQRDRLVLF AAKQG-----VIKIIPPTHTPENYR-----TVRDVIGSL
YVG-VPQSRQRVFFIGLKS DRPLNQQ-----ILTP-----PSKVI ESEYTSLEEAISDLPVIE-----AGEGGEVQDY
CG-VPQLRKRTFVIGHRHGSIAD-----LANVLQQR LAKQSL-----TVRDY
CG-VPQSRTRFSLIGKLNSEHNF-----LIPTLSRKLSDKPM-----TVRDY
YVG-VPQRRHRIIVGIRK DQD-----VAFRVPEPTHKEKYR-----TASEALADI
IG-AHHQRHRWFCLAIRKDY EPEE-----IIVSVNATKFDWENNEPPCQVDNK-----SYENSTLVRL
YFGNIPQNRERIYIVGFRN-----IEHYKNFNFPMPQP-----LTLTIKDM
YFN-VPQNRERLYIIGIREDLIKNEE-----WSLDFKRKDI LQKGKQRLVELDIKSFNFRWTAQ-----SAATKRLKDL
YFG-IPQNRERVF C I SILN-----PNEDFTFPQKQ-----NLTL SMNDL
YFG-SSQARRRVFMISTLNEF-----VELPKGDKKPKS-----IKKVLNKI
YFG-IPQNRERIYLVGF-----LNHDVDFRFPQP-----IGQATAVGD I
YFG-LPQNRERIYIVGFDRKS-----ISNYSDFQMPTP-----LQEKTRVGN I
YFG-VPQNRERIYIVGFNKEK-----VRNHEHFTFPTP-----LKT KTRVGD I
YFG-VPQNRERIYIVGFHKS-----TGVNSFSY PEP-----LDKIVTFADI
YFQ-VPQNRRLRVYIVGLDQSQPELT-----ITSHIGATDSHKFKQLSNQASLFD-----TNKIMLVRDI
YFG-VPQNRVRIYILGILGSKPKLT-----LTSNVGAADSHKYK--NEQISLFD-----ES-YATVKDI
YFG-IPQKRKR FYLVAF LNQN-----IHFEFPPK-----PMISKDIGEV
YFG-LPQRRERIVIVGFHPDLG-----INDFSFPKGN-----PDNKVPINA I
YVG-IPQKRERIYMICFRNDLN-----IQNFQFPKP-----FELNTFVKDL
YVG-NAQRRRRVFI FGYKQDLNYSKAME-----ESPLDKI IYHNGLFAEAFP IEDYANKNR-----VNRTHITHDIVDI
YVG-TPQRRKRAIIRLNKKGTIWN-----LPLKQNI VSVEQ-----AIGNLPSI
YGG-TPQVRERVFITATLVPERMRDERIPRTETGEIDAEAIGPKPVATMNDRFP I KGGTEL FHPGDRKSGWNLLTSGI I REG
YVG-VAQNRDRVFIIGIQQKLGVPD-----FSFPEYSESEQRLYDILDNLQTPSII-----PESLPIQRNL
YFG-VAQNRDRVFI VGIQQKLDLNG-----FSFPEYAFSDQRIYHILDNI FAPETK-----LESIP IQRNL
YID-VAQKRERIV IIGIREDLVK-----EQKYPFRFPLAQ-----VYKPVLKDV
YVG-VSQLRPRVLFVALKNEYTN-----FFKWPEPNSEQPK-----TVGELLFDL
```

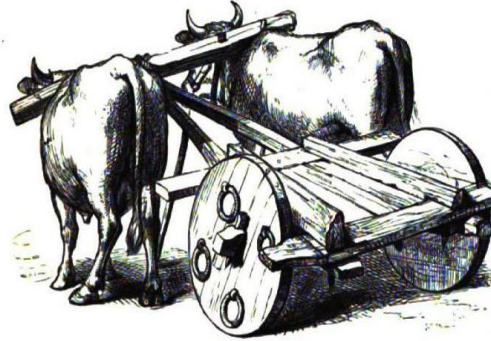
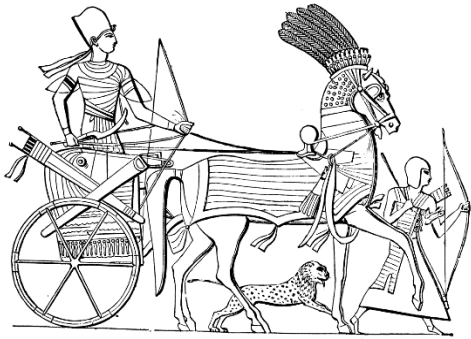


Выравнивание другой программой и раскраска по а.к.о.

	10	20	30	40	50	60	70	80
-GVPQDRKR VFI	-VG Y	-RE DL	-NL -KFEFPKP	-LN -	-KKVTLRDA	I	G	D
-GVAQDRER VFY	-VG F	-RK DL	-NISNFEPYP	-ISEK -	-ERKYLKDS	I	W	D
-GVAQDRKR VFY	-IG F	-RK DL	-EI -KFSFPKGSTVEDK	-DKITLKDVI	W	D	L	
-GVAQDRKR VFY	-IG F	-RK EL	-NI -NYLPP IPHLI	-KPTFKDVI	W	D	L	
-GIPQQRDR LVL	-FA A	-KQGV I	-KI I PPTH	-TPENYRTVRDV	I	G	S	
-GVPQSRQR VFF	-IG L	-KSD R	-PLNQQIL T PPSK	-VIES -	-EYTSLEEAI	S	D	
-GVPQLRKRT FV	-IG H	-RHGS I	-ADLANVLQ	-QRLAKQSLT	V	R	D	
-GVPQSRTR FSL	-IG K	-LNSEH	-NFL IPTLS	-RKLSDKP	M	T	V	
-GVPQRRHR I I I	-VG I	-RK DQ	-DV -AFRVPEP	-THKE -	-KYRTASEAL	A	D	
-GAHHQRHR WFC	-LA I	-RK DYEPEE I I	-VSVNATKFDWENNE	-PPCQVDNKS	Y	E	N	
GNIPQNRER I Y I	-VG F	-RNIE	-HYKNFNFPMP	-QPLTLTI	K	D	M	
-NVPQNRER L Y I	-IG I	-RE DL I KNE	-EWSLDFKRKDI LQKGKQRLVELDIKSFNFRWT	-AQSAATKRL	K	D	L	
-GIPQNRER VFC	-IS I	-LNP NE	-DFTFPQK	-QNLTL	S	M	N	
-GSSQARRR VFM	-IS T	-LNEFVELPKGD	-KKPKS	I	K	K	V	
-GIPQNRER I Y L	-VG F	-LNHDV	-DFRFPQP	-IGQATAV	G	D	I	
-GLPQNRER I Y I	-VG F	-DRKS I S	-NYSDFQMPTP	-LQEKTRV	G	N	I	
-GVPQNRER I Y I	-VG F	-NKEKVR	-NHEHFTFPTP	-LKT	K	T	R	
-GVPQNRER I Y I	-VG F	-HKST	-GVNSFSYPE	-PLDKIVT	F	A	D	
-QVPQNR LRVY I	-VGLD	-QSQPELTITSHI	-GATDS	-HKFKQ	-LSNQASL	-FD	-TNKIMLVRDI	
-GVPQNRVRI Y I	-LGI L	-GSKPKLTLTSNV	-GAADS	-HKYK	-NEQISL	-FD	-ESYATVKDI	
-GIPQKRKR FYL	-VA F	-LNQNI	-HFEFPKP	-PMISKDI	G	E	V	
-GLPQRRER I VI	-VG F	-HPDL	-GINDFSFPK	-GNPDNKVP	I	N	A	
-GIPQKRER I YM	-IC F	-RNDL	-NIQNFQFPK	-FELNTFVKDL				
-GNAQRRR R VFI	-FG Y	-KQDLNYSKAMEESPLD	-KI I YHN	-GLFAEAFPIEDYANKNRVNRTHITHDIVDI				
-GTPQRRKRAI IRLNKKGT	-IWNL	-PLKQNI	-SVEQA	I	G	N	L	
-GTPQVRER VFI	-TATLVPERMRDERI	PR TET	-GEIDA	-EAIGPK	-PVATMNDRFP	I	K	
-GVAQNRDR VFI	-IG I	-QKQL	-GVPDFSFP	EYSESEQRLYDILDNLQTPSII				
-GVAQNRDR VFI	-VG I	-QKQL	-DLNGFSFP	EYAESEQRLYHILDNLEAPETK				
-GVAQNRDR VFI	-VG I	-QKQL	-DLNGFSFP	EYTESEQRLYHILDNLEVPETK				
-DVAQKRER I VI	-IG I	-RE DLVKE	-QKYPFRFPLA	-QVYKPV	L	K	D	
-GVSQRLRPR VLF	-VA L	-KNEY	-TNFFKWPEP	-NSEQPKTVGELLFDL				

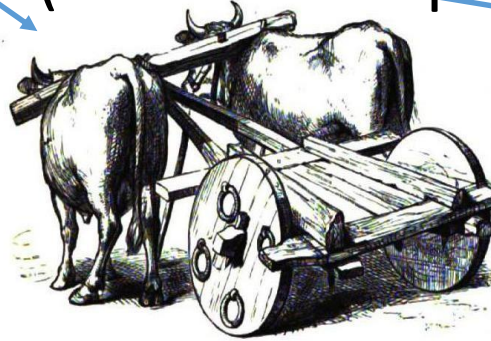
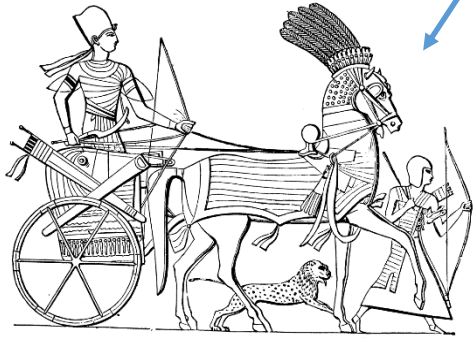
Консервативное значит важное

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



LUCA

(колесного транспорта)

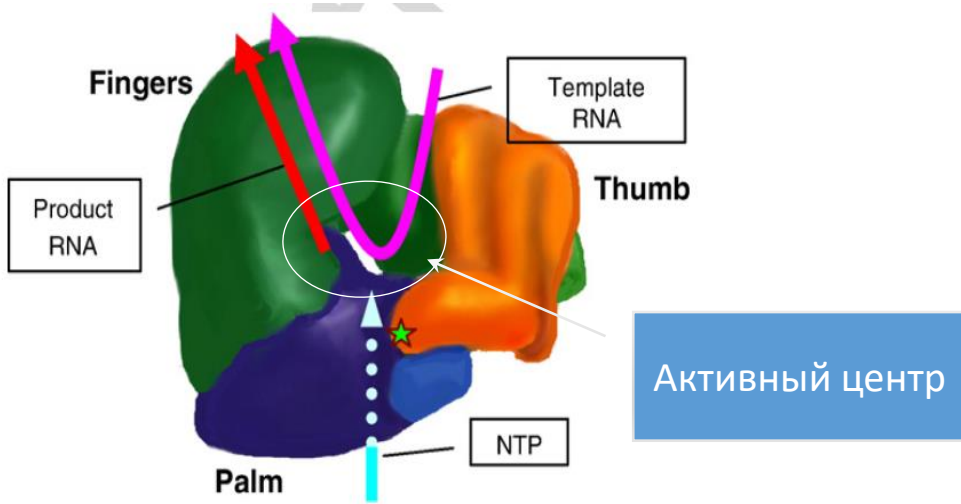


В выравниваниях белков – то же самое:
Сохраняющееся в эволюции (консервативное) – важно

РНК зависимая РНК полимераза (RdRP), консервативные участки

```

*      320      *      340      *      360      *      380      *      400      *      420      *      440      *      460
FKTMRIRFGDVGDLDDFFSADASLSPFMIREA..GRIMSELS...GTPSHFGTALINTIIYSKHLIYNCCY...HVCGSMPSGSPCTALNSTINNVLIIYVFSKIFGKSPVFF.....CQALKILC.YGDDVLIIVFSRDV
EVAMQG.FERVYDVIDYSNEDSTHVSAMFRLL..A...EEFF.TPENGFDPLTREYLESLAISTHAFEEKRF.....LTGGLPSGCAATSMNTIMNIIIRAGLYLTYKNFEFDD.....VKVLS.YGDDLIVATNYQL
ETHFAQ.YKNVWDVLYSADANHCSDAMNMFEEVFERTEFG.....FHPNAEWILKTLVNTTEHAYENKRI.....VVEGCMPSGCSATSIINTILNIIYVLYALRRHYEGVELDT.....YTMIS.YGDDIVVASYDYL
.....WSLCVATIDVSDHDTFWPGWLRDLICDELINMGYA.PWVVKLFETSLKLPVYVGAFAPEQGHTLLGDPSNPDLVGLSSGQGATDLMGTLIMSTIYLVMLQDHTAPHLNSRIKDMPSACRFLDSYWQGHEETROIS.KSDDAILGWTKGR
LRLRLE.NWVYCDADGSOEDSSLTPYLINAV..LTLRSTYMEDWDVGLQMLRNLYTEIVYTPISTPDGTIV.....KKFRGNNSGQPSIVDNLSLMVVIAMHYALIKECFEVEEID.....STCVFFV.NGDDLIAVNPEK
HDKLNRPGLWLGSGDGRDSSIDPFFFDVV..KTKRKHFL..PSEHHRaidLIYDEILNTTICLANGMVI.....KKNVGTQR.QPSTVDNTLVMITAFLYAYIHKTGDRELAL.....LNERFIFVC.NGDDNKFAISPQF
AISLASFSYPYGFNGDFANEDGMFHPSSFSMV..SELANIFY...GNFLSTERDNLTRMLTNRFSIMKGAIL.....RVPGGSPSGFFMTVFNSEINLFYLQSAWIMLARFNGRQDISH.....PCNFPKYVRACV.YGDDNIIVAIKMEV
AARMKEKGNVDVLCODYSSEFDGLLSKQVMDVI..ASVINELC.GGEDQLKNARRNLMACCSRIAICKNTVW.....RVECGIPSGFFMTVFNSEINLFYLRHYHKIMREQQAPELMV.....QSFDKLIGLVT.YGDDNLSVNAVV
YAEHAK.YKNHFDADYIANDSTQNRQIMTES..FSIMSRILT...ASPELAEVVAQDLLAPSEMDVGDYVI.....RVKEGLPSGFFPCTSQVNSINHWITLICALSEATGLSPDVV.....QMSYFYSFYGDDIVSTDIDF
NNLTSKASDFLCLDYSKFDSTMSPCVVRIA..IDLADCC...EQTELTKSVVLTILKSHFMTILAMIV.....QTKRGLPSGMPFTSVINSHCHWLLWSAAVYKSCAEIGLHCS.....NLYEDAPFYT.YGDDGVYAMTEMM
IQRIKS.AAKVYAVDYSKWDSTQSPRVSAAAS..IDLRYFS...DRSPIVDSAANTLKSPPIAIFNGVAV.....KVSSGLPSGMPFTSVINSHCHWLLWSAAVYKSCAEIGLHCS.....NLFSTFLMMT.YGDDGVYMFPMFM
TKRLERPKHDRYCVLYSKWDSTQPPKVTSSQS..IDILRHFT...DKSPIVDSACATLKSNPIGIFNGVAF.....KVAGGLPSGMPFTSVINSHCHWLLWSAAVYKSCAEIGLHCS.....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 : FQVFFIVNDDVELALSIQADGHIHVGQKDTAVETILRNTRNKPIIIGLSINTLAQALANKDRQDIDYFVGVPPIFPTNSKADH : 15
THEIE_STRA3 : YQVFFIIDDIDLVELIDADGHIHGQNDLPVDEARRRLPDKI-IGLSVSTMAEYQKSQ-LSVVDYIIGIGPFNPQSKADA : 14:
THEIE_LISIN : YQVFFIINDDDVALALEIGADGHIHVGQNDDEEIRQVIASCAGKMKIIGLSVHSVSEAEBAERLGSVDYIIGVGPPIFPTISKADA : 14:
THEIE_ANOFW : YNIFFFIVNDDVDLALALQADGVHVGQDEVEAERVDRDRIGDKY-LGVSVHNLNEVKKAL-AACADYVGLGPIFPPTVSKEDA : 14:
THEIE_GEOTN : YGVFFIVNDDVELAIAIDADGVHVGQDDEADARRVREKIGDKI-LGVSAHNVVEEARAAI-EAGADYIIGVGPPIYPTRSKDDA : 14
THEIE_BACSU : AGVFFIVNDDVELALNLKADGHIHGQEDANAERVRAAIGDMI-LGVSAMTSEVVKQAE-EDGADYVGLGPIYPTETTKDT : 14
THEIE_BACA2 : AGIFFIINDDDVELALRLEADGVHIGQDDADAETRAAIGDMI-LGVSAMTSEVVKRAE-AAGADYVGMGPVYPTETTKDA : 14
THEIE_OCEIH : FQIFFIINDDDVDLAKQLDADGHIHGQDDQPVVVRKQFENKI-IGLSISTNNELNQSP-LDLVDYIIGVGPPIFDTNTKEDA : 14
THEIE_STAAB : YNVFFIVNDDVSLAKEINADGHIHVGQDDAKVKEIAQYFTDKI-IGLSISDLGEYAKSD-LTHVDYIIGVGPPIYPTPSKHDA : 14
THEIE_STACT : YNVFFIVNDDVALAEEIDADGHIHVGQDDEAVDDFNRRFEGKI-IGLSIGNLEELNASD-LTYVDYIIGVGPPIFATPSKDDA : 14:
      6pFI61DD6 La 6 ADG6H6GQ D 6G6S 2 DY G6GP pT 3K Da
```

```
      *           180           *           200           *           220           *           240
THEIE_LACLS : AKETGKIDLR-LMLLENQLPIVGIIGGITQDSLTELSAIGLDGLAVISLLTEAENPKKVAQMIRQKITKNG~~~~~ : 21
THEIE_MANSM : SPIVGMNFIRQIRQLGIDKPCVAIGGITKEESAAILRRLGADGVAVISAIHSHSVNIANTVKTLAQK~~~~~ : 22
THEIE_STRA3 : KPAVGNRTTKAVREINQDIPVVAIGGITSDVFVDIIESEGADGLAVISAIISKANHIVDATRQLRYEVEKALVNRQKRSDVI : 22:
THEIE_LISIN : EPVSGTAILEEIRRAGIKLPIVGIIGGITNETNSAEVLTAGADGVSVISAITRSEDCQSVIKQLKNPGSPS~~~~~ : 21
THEIE_ANOFW : KQACGLTMEHIRAEKRVPLVAIGGITETQAKQVIEAGADGLAVISAIKRAEHIYEQTKRLYEMVMRAKQKQKQDR~~~~~ : 21
THEIE_GEOTN : NEAQQPGILRHLRREQGITIPIVVAIGGITADNTRAVIEAGADGVSVISAIASAPEPKAAAAALATAVREANL---R~~~~~ : 22
THEIE_BACSU : RAVQGVSLIEAVRRQGISIPIVGIIGGITIDNAAPVIEAGADGVSMISAISSAEDPESAARKFREEIQTYKTG--R~~~~~ : 22:
THEIE_BACA2 : EAVQGVTLIEEVRQGITIPIVGIIGGITADNAAPVIEAGADGVSMISAISSAEDPKAAARKFSEEIRRSKAGLSR~~~~~ : 22
THEIE_OCEIH : KTAVGLEWISLKKQHPSLPIVVAIGGITNTNAQEIIEAGADGVSVISAITETDHIHQAVQRL~~~~~ : 20
THEIE_STAAB : HTPVGPMEIATFKEMNPQLPIVVAIGGITSNVAPIVEAGANGISVISAIKXSENIKTVNRFKDFFN~~~~~ : 21:
THEIE_STACT : SEPVGPKMIETLRKEVGDLEIPIVVAIGGISLDNVQEVAKTSADGVSVISAIARSPHVTETVHKFLQYFK~~~~~ : 21:
```

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

```
      160           *           180           *           200           *           220
THEIE_LACLS : AAKPTG---TKDLRLMLLENQLPIVGIIGGITQDSLTELSAIGLDGLAVISLLTEAENPKKVAQMIRQKITKNG : 218
THEIE_MANSM : HSPVGMNFIRQIRQLGIDK--PCVAIGGITKEESAAILRRLGADGVAVISAIHSHSVNIANTVKTLAQK----- : 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

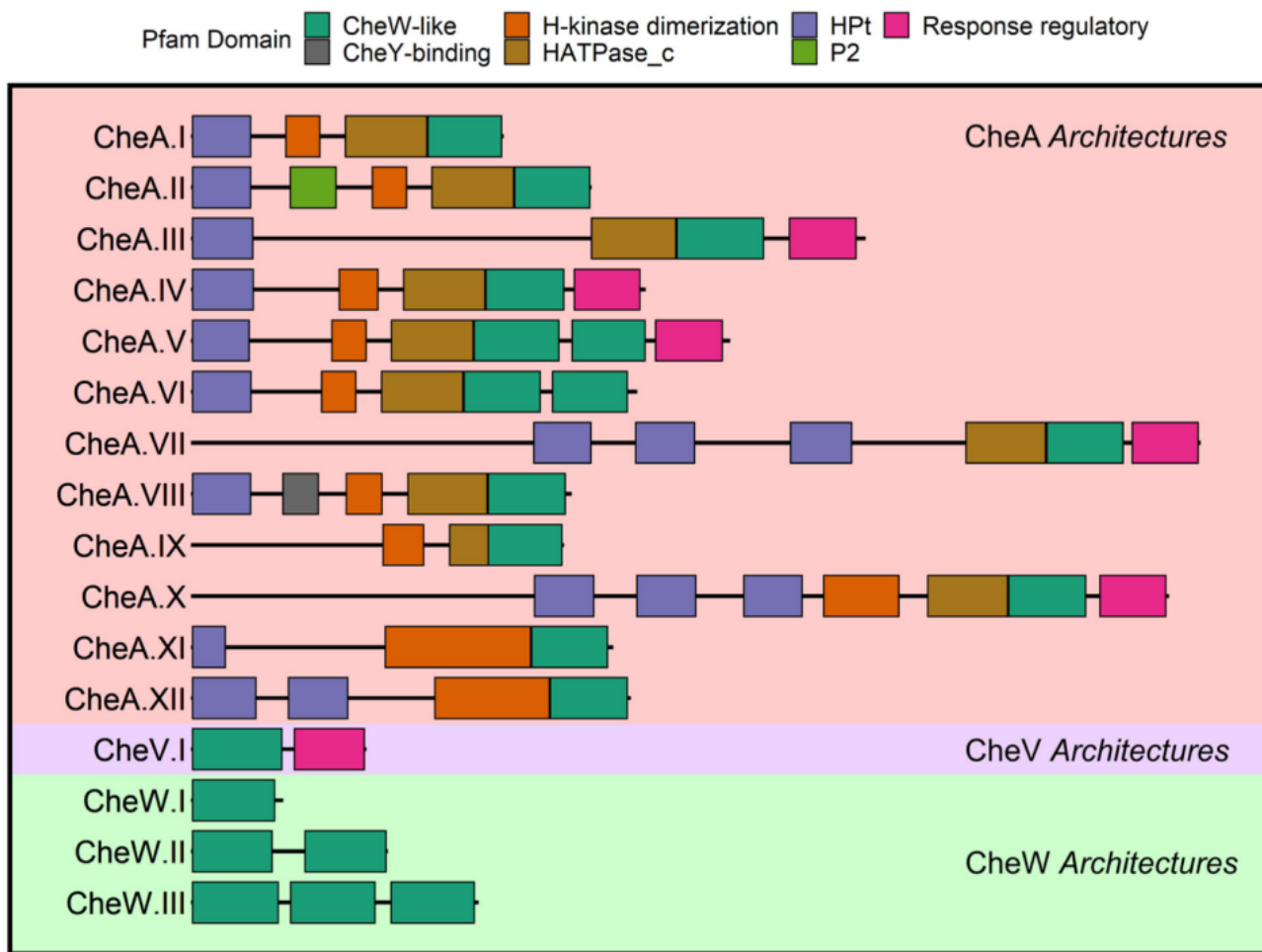
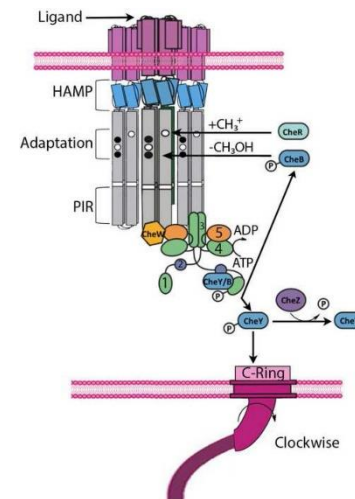


FIGURE 2. Schematic of the most abundant CheW-containing domain Architectures seen in nature.

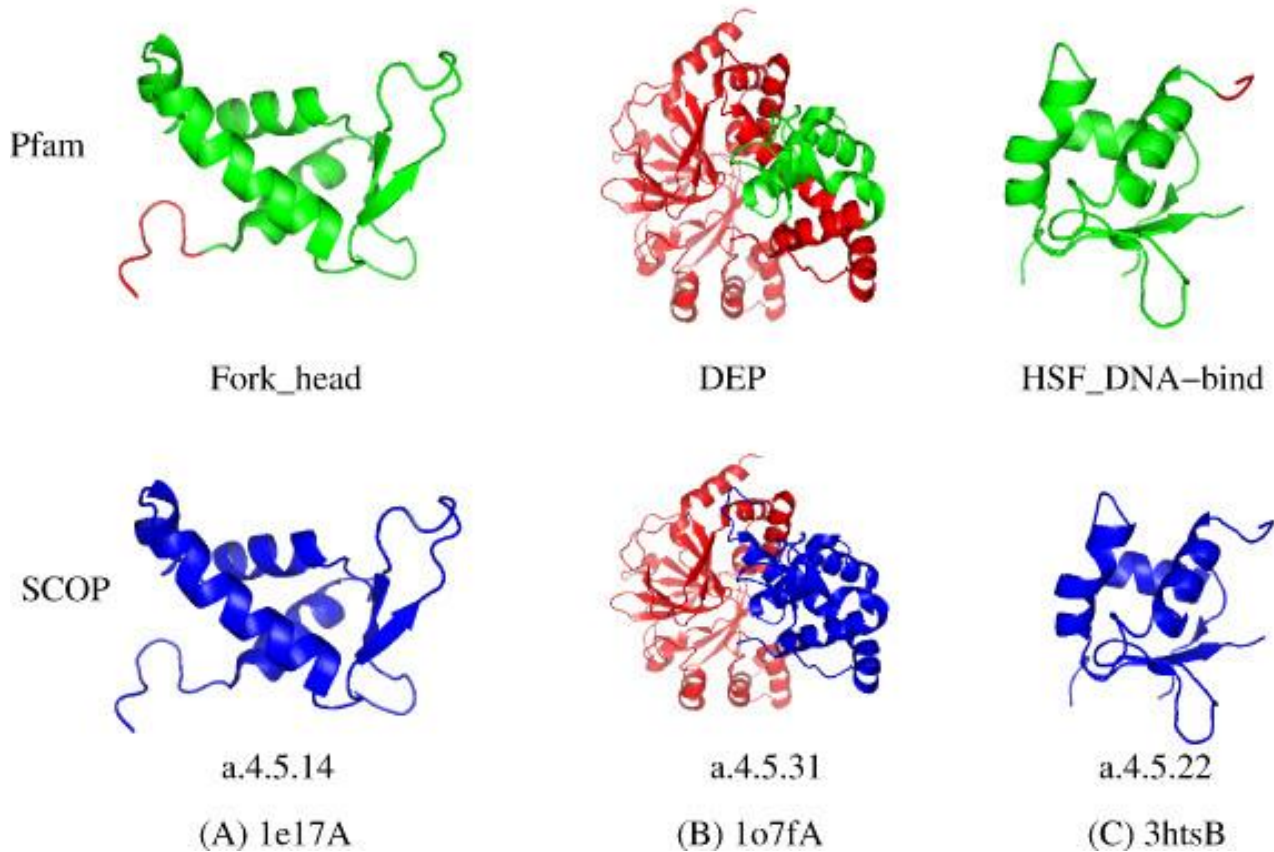
Sixteen distinct architectural variants containing the CheW-like domain were identified



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

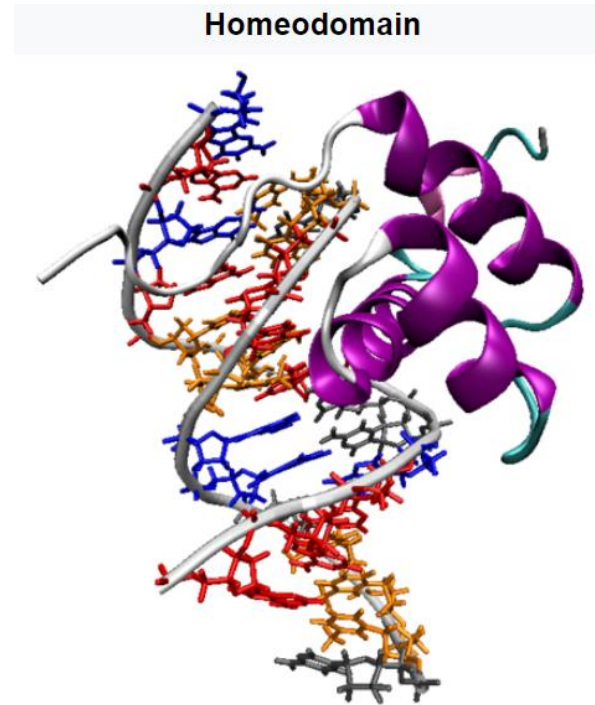
- Имеют определенную функцию (не всегда известна)

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-DMVAAQDEEGCGPGRSLLLESP-GLTSGSDTPQQD	: 80
SW: PMX2_HUMAN/1	-----MDSAAAFAALDKPALGCPGPPPPALGCPGDCQAQRNFSVSHLLDLEEVAAAGRLAARPGARABAREGAAAREPSCGSSGSEAAAPQD	: 86
SW: PMX1_HUMAN/1	-----MTSSYGHVLEQPALGCRRLDSPGLMDTLQAKNFSVSHLLDLEEAG-DMVAAQADENVGACRSLLLESP-GLTSGSDTPQQD	: 80
SW: ARX_BRARE/1	ISQAPQVVISRSKSYREN-APFSQS---D-EQSP--EHAQELVELST-----LKFEEDEVVKEEACQDN-----S-----LSPKDEESLH-NDGVDKCDSDSVCLS	: 84
SW: ARX_MOUSE/1	ISQAPQVVISRSKSYRENGCAPVFPVPPALD-ELSGPCGVVAHPEERLSAASGPGSAPAAGCGTCAEDDEEELLEDEEEDEREELLEDDDEELLEDDARALLKEPERRCVATTCTVAIAAAAAAAAAAVATEGGELESPKRELLHHPEDAREKDCDSDSVCLS	: 157
SW: AL_DROME/1-1	-----MGISEEIKLEELPQAKLAHPDAVVLVDRAPGSSAASAGAAALTVSMVSYSGAPSCASGASGCTNSPVSDGNS	: 72
SW: ALX4_MOUSE/1	-TFLSAGAKQCPCDAKSRARYGACQDLAAPLESSSGARGSPNKFQPPQPTQP-----PPAPPAPPAHLYLQRCACKTPDDCSLKLQEGSSGCHNAALQVPCYAKRESNLCEPELPPDSFVPCVMDNSYLSVKRETGARCPDRASAEIPL	: 145
SW: ALX4_HUMAN/1	-TFLSAAAQAQPCDAKSRARYGACQDLATPLESSGARGSPNKFQPPQPTQPQPPQPPQPPQPPQPPQPPHLYLQRCACKTPDDCSLKLQEGSSGCHNAALQVPCYAKRESNLCEPELPPDSFVPCVMDNSYLSVKREAGVRCQDRASSDLPSP	: 157
SW: RX2_CHICK/1	-----NPSRLHSIEAILGFTKDDGLLGFQFP-----DGGAGSAAKAAADKRGPRHCLPKGPAEPPPAEHQGRFQEPYPCGASAPF-----LPAGCGDC	: 83
SW: RX2_BRARE/1	-----GISCRVHSIDVILGFSKDDPFLLEPSGR-----HKVDLEDQLEEQEKQVADPYSHLQIPDQIQQQQSVYH---DTGLFSTDKCADLGDPRSINVEDSRS	: 92
SW: RX1_XENLA/1	-----NPSRLHSIEAILGFKEDS-VLGSFQSEIISPRNAKEVDKRSRHLCHMTREIHPQEHLEDG-QADCYG--DPYSGRTSSECLS-PGLST--SNSDN	: 91
SW: RX_HUMAN/1-1	-----STSRLHSIEAILGFTKDDG-ILGTFPAERGARGAKEDRRLGARCPACPKPAEREGSEPSPPAPAPAPYEAPRPPYCPKPEWEARPSPLPVGPATGEA	: 97
SW: PIX2_BRARE/1	-----MTHSKDPLSLDHHHHHHVTCGKHAFLSMASLLQPLQRSDVSKHRLDVHTVSDTSSPESVKEKRCQ--	: 66
SW: PIX2_HUMAN/1	-----MTNCRKLVSAVGLQVPAAEVCLFSDKSEIKKVFETDPSRKRKAASAKFPFPHQPCANEKRSQQ-	: 68
SW: PIX1_HUMAN/1	-----MDAFKCGMSLERLPEGFPPPPPHDMGPAFHLLARPADPREPLEN-SASESSDTLPEKEKRGCEP	: 64
SW: OTP_MOUSE/1	-----MLSHADLLDARLCHKDAEALLGHREAVKRLGVGCSDPGCHPCDLAPNSDPVEGATLLPREDITTVGSTPASLAVSAKDPKQPGPQCGP	: 90
SW: PMX1_CHICK/1	NDQLNSEE-----KPKRQRNRRTFTFNSSQLQALERWERETHYDPAFVRDLARRVNLTEARVQVVFQNRRAKFRNNEAMLSKMASLLKSYSGDVTAVEQPIVPRPAPRPTDYLWGTASPYSAMATYSTTCTMAS-----	: 213
SW: PMX2_HUMAN/1	GCPCSPGRCG-----AAKRRKQRNRRTFTFNSSQLQALERWERETHYDPAFVREELARRVNLTEARVQVVFQNRRAKFRNNEAMLSRSASLLKSYSGE-AAEQVPAPRPTALSPTDYLWGTASSPYSTVPFPYPCSSGP-----	: 221
SW: PMX1_HUMAN/1	NDQLNSEE-----KPKRQRNRRTFTFNSSQLQALERWERETHYDPAFVRDLARRVNLTEARVQVVFQNRRAKFRNNEAMLANKNASLLKSYSGDVTAVEQPIVPRPAPRPTDYLWGTASPYSAMATYSATCANNS-----	: 213
SW: ARX_BRARE/1	AGSDSEEG-----MLKRRQRNRRTFTFTSYQLEELERAFQETHYDPAFVREELARRLDLTEARVQVVFQNRRAKWRREERCAQTHPPGLPFPFPCPLSATHPSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPSCAPLG	: 230
SW: ARX_MOUSE/1	AGSDSEEG-----LLKRRQRNRRTFTFTSYQLEELERAFQETHYDPAFVREELARRLDLTEARVQVVFQNRRAKWRREERCAQTHPPGLPFPFPCPLSATHPSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPSCAPLG	: 303
SW: AL_DROME/1-1	EKADSEY-----PKRQRNRRTFTFTSYQLEELERAFQETHYDPAFVREELARRVNLTEARVQVVFQNRRAKWRREERCAQTHPPGLPFPFPCPLSATHPSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPSCAPLG	: 212
SW: ALX4_MOUSE/1	EKTDSESN-----KCKRQRNRRTFTFTSYQLEELERAFQETHYDPAFVREELARRLDLTEARVQVVFQNRRAKWRREERCAQTHPPGLPFPFPCPLSATHPSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPSCAPLG	: 290
SW: ALX4_HUMAN/1	EKADSESN-----KCKRQRNRRTFTFTSYQLEELERAFQETHYDPAFVREELARRLDLTEARVQVVFQNRRAKWRREERCAQTHPPGLPFPFPCPLSATHPSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPSCAPLG	: 302
SW: RX2_CHICK/1	KPSDEEQ-----PKRQRNRRTFTFTSYQLEELERAFQETHYDPAFVREELARRVNLTEARVQVVFQNRRAKWRREERCAQTHPPGLPFPFPCPLSATHPSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPSCAPLG	: 215
SW: RX2_BRARE/1	PDIPDEDQ-----PKRQRNRRTFTFTSYQLEELERAFQETHYDPAFVREELARRVNLTEARVQVVFQNRRAKWRREERCAQTHPPGLPFPFPCPLSATHPSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPSCAPLG	: 225
SW: RX1_XENLA/1	KLSDDDEQ-----PKRQRNRRTFTFTSYQLEELERAFQETHYDPAFVREELARRVNLTEARVQVVFQNRRAKWRREERCAQTHPPGLPFPFPCPLSATHPSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPSCAPLG	: 224
SW: RX_HUMAN/1-1	KLSEEEQ-----PKRQRNRRTFTFTSYQLEELERAFQETHYDPAFVREELARRVNLTEARVQVVFQNRRAKWRREERCAQTHPPGLPFPFPCPLSATHPSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPSCAPLG	: 242
SW: PIX2_BRARE/1	SKNEDSW-----DDPSKRRQRNRRTFTFTSYQLEELERAFQETHYDPAFVREELARRVNLTEARVQVVFQNRRAKWRREERCAQTHPPGLPFPFPCPLSATHPSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPSCAPLG	: 212
SW: PIX2_HUMAN/1	GRNEDVGA-----EDPSKRRQRNRRTFTFTSYQLEELERAFQETHYDPAFVREELARRVNLTEARVQVVFQNRRAKWRREERCAQTHPPGLPFPFPCPLSATHPSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPSCAPLG	: 215
SW: PIX1_HUMAN/1	KCPEDSCAGCTGCCGADDPAKRRQRNRRTFTFTSYQLEELERAFQETHYDPAFVREELARRVNLTEARVQVVFQNRRAKWRREERCAQTHPPGLPFPFPCPLSATHPSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPSCAPLG	: 218
SW: OTP_MOUSE/1	NPSQACQQ-----CQQCQRQRNRRTFTFTSYQLEELERAFQETHYDPAFVREELARRVNLTEARVQVVFQNRRAKWRREERCAQTHPPGLPFPFPCPLSATHPSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPSCAPLG	: 236
k 4 4R RT Ft QL EL E R F 4 HYPD RE 6A L E R6qVWFQNRRAK54 e4		
SW: PMX1_CHICK/1	-----PAQCMMNMANSLALPLAKRHSYSLQRNQVPTVN-----	: 245
SW: PMX2_HUMAN/1	-----ATPCVNMNMANSLALPLAKRHSYSLQRNQVPTVN-----	: 253
SW: PMX1_HUMAN/1	-----PAQCINMANSLALPLAKRHSYSLQRNQVPTVN-----	: 245
SW: ARX_BRARE/1	LGTFLGTAAMFRHFAFIPTFCRLFSSMCLPTSASTAAALLRQTAPPVSPVQSAALPEPPSSSSSTAADRASSIAALPLAKRHSYA-QLTQLNLIPSGTACKKEVC-----	: 336
SW: ARX_MOUSE/1	LSTFLGAAVFRHFAFISPAFCRLFSTMAPLTSASTAAALLRQTAPPVGAVASGALADP-----ATAAADRASSIAALPLAKRHSYAQLTQLNLIPGCTCKKEVC-----	: 404
SW: AL_DROME/1-1	PPTSPASGHAXPQVLQVIGIALTQQASSLSPT---QTSFVALTSHSPQRQLPPSHQAPPVPPPPAAATPPEDRRTSSIAALPLAKRHSYSLQRNQVPTVN-----VS	: 313
SW: ALX4_MOUSE/1	DFL-----SVSGACSHVQTHMCSLFCFAAGISPLNGCYELNCEPDRKTSIAALPLAKRHSYAASISWAT-----	: 354
SW: ALX4_HUMAN/1	DFL-----SVSGACSHVQTHMCSLFCFAAGISPLNGCYELNCEPDRKTSIAALPLAKRHSYAASISWAT-----	: 366
SW: RX2_CHICK/1	LPASYTTPPFL-----NSPAMTHALQPLGAMGPPPPYQCGAFAVDKFLDEGDPNRTSSIAALPLAKRHSYSLQRNQVPTVN-----	: 290
SW: RX2_BRARE/1	LQPTYTAHPCFL-----NTSPGMHNTQIPM---PPPPYQCPVFNDRKFLDEVD---RSSIAALPLAKRHSYSLQRNQVPTVN-----	: 297
SW: RX1_XENLA/1	LPASYTTPPFI-----NPVSVGHALQPLGAMGPPPPYQCGAFAVDKFLDETDPRMSSIAALPLAKRHSYSLQRNQVPTVN-----	: 296
SW: RX_HUMAN/1-1	LPASYTTPPFPFL-----NSPPLGCPQLPL---APPYPTFCGPFCDKFLDEADPNSSIAALPLAKRHSYSLQRNQVPTVN-----	: 319
SW: PIX2_BRARE/1	SISMSMSMSMVPASVTCVPGSSSL-----NSLNNLNLNLSNPSLNSAVTTPACPYAPPTPPY-VYRDTCNSSLASLPLAKRHSYSLQRNQVPTVN-----	: 314
SW: PIX2_HUMAN/1	SISMSMSMSMVPASVTCVPGSSSL-----NSLNNLNLNLSNPSLNSAVTTPACPYAPPTPPY-VYRDTCNSSLASLPLAKRHSYSLQRNQVPTVN-----	: 317
SW: PIX1_HUMAN/1	SISMTMPSMCPGAVPGMNSCL-----MNIN---MLTGSSLNSAMSFGACPYCTPASPYVYRDTCNSSLASLPLAKRHSYSLQRNQVPTVN-----	: 314
SW: OTP_MOUSE/1	SQCSLAAGPPPNMCLNSLNSAGNCAQLQ---SHLYQAPFGMVPASLPGSPMNSGSPQLCCSSPDSVDVRCSTASLPLAKRHSYSLQRNQVPTVN-----	: 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

Гомеодомен является ДОМЕНОМ

Доказательство - выравнивание

There are 25976 sequences with the following architecture [X2JL88 DROME](#) [Drosophila melanogaster (Fruit fly)] Uncharacterized

Show all sequences with this architecture.

There are 2311 sequences with the following architecture [X2JDY7 DROME](#) [Drosophila melanogaster (Fruit fly)] POU domain pr

Show all sequences with this architecture.

There are 2108 sequences with the following architecture [W6NCH4 HAECO](#) [Haemonchus contortus (Barber pole worm)] Zinc f

Show all sequences with this architecture.

There are 1903 sequences with the following architecture [MOU1E3 MUSAM](#) [Musa acuminata subsp. malaccensis (Wild banana)]

Show all sequences with this architecture.

There are 1836 sequences with the following architecture [X1WJ92 ACYPI](#) [Acyrtosiphon pisum (Pea aphid)] Uncharacterized p

G4VRE6_SCHEMA/203-259
HME2B_DANRE/173-229
HM1N_BOMMO/373-429
T1EHE5_HELRO/41-97
HM05_CAEE/36-92
G3IBX4_CRIGR/25-81
F1QFR3_DANRE/136-192
B8A5N9_DANRE/135-191
Q91967_CHICK/77-133
DLX3B_DANRE/126-182
HM43_CAEE/103-159
HM23_CAEE/212-268
MSX3_MOUSE/88-144
HM30_CAEE/96-152
BARH2_RAT/230-286
BARH1_DROME/300-356
BARX2_MOUSE/138-194
BSH_DROME/275-331
H2XXU6_CIOIN/470-524
HM19_CAEE/95-151
SLOU_DROME/546-602
F6VUQ6_XENTR/112-168
TIN_DROME/302-358
NKX25_RAT/138-194
H0XR12_OTOGA/100-156
HM09_CAEE/71-127
H2VEX2_TAKRU/13-69
U3K517_FICAL/59-115
TLX3_CHICK/173-229
U3JZQ6_FICAL/136-188
LBX1_MOUSE/126-182
G4VGG4_SCHEMA/38-94
BCD_DROME/98-153
BCD_DROME/98-153 (SS)
VENTX_HUMAN/92-148
VENT1_XENTR/128-184
Q804C9_XENTR/190-246
K4B8Z1_SOLLC/24-79
PHO2_YEAST/78-134
WOX9_ATH/52-113
WOX9_ORYSJ/11-72
WOX2_ORYSJ/24-85
WOX4_ATH/87-148
WOX1_ATH/73-134
WOX2_ATH/11-72
WOX5_ORYSJ/41-102
MUS_SOLLC/25-85
WOX6_ATH/58-119
YHP1_YEAST/174-230
YOX1_YEAST/177-233
WARA_DICDI/163-219
PHX1_SCHPO/169-223
CUT_DROME/1746-1802
CUX2_MOUSE/1114-1170
CUX1_MOUSE/1240-1296
Q22810_CAEE/212-268
HBX2_DICDI/486-542
BRX1_HUMAN/234-293

```

KRPRTSFTVPOLKRLSSEFE...K..NRYLDELRRKKLATE...DLRESQVKIWFONKRAKTKK
KRPRTAFTAELQRLKNEFQ...N..NRYLTERQRRQALAE...GLNESQIKIWFONKRAKIKK
KRPRTAFSGPOLARLKHEFA...E..NRYLTERRRQSLAAEL...GLAEAQIKIWFONKRAKIKK
KRPRTAFTGDOLARLKREFN...E..NKYLTERRRTCLAKEL...SLNESQIKIWFONKRAKMKK
KRPRTVFTDEQLKLEESFN...T..SEVLSGSTRAKLAESL...GLSDNQKVIWFONRRTKQKK
KRVRTVFTAEQLYRLLEMEFQ...R..CQYVVGERTERLARQL...NLSEITQVKVIWFONRRTKQKK
KRIRTAFSPSQLLRLERAFE...K..NHVYVGAERKQLANG...CLTETQVKVIWFONRRTKHKK
KRMRTSFTNDQLSRLKEFE...R..CQYVVGSERFLASAL...QLTEAQKVIWFONRRTKKWR
KRVRTVFKPEQLERLEQEF...K..CQYVVGERTERLARQL...GLSPQVRIWFONRRSKHRR
RKPRTIYSSYQALQRRFQ...K..AQYLALPERAEALAE...GLTQVQKVIWFONRRSKFKK
RKPRTIYSSQQLMLQKKFQ...K..TOYALPDRAALAE...GLSQTQKVIWFONRRSKQKK
RKARTIYGTQTQQLQEDMF...K..QYVYVGAERENLAQRL...GLSPQVRIWFONRRSKHRR
RKPRTPTTAQQLLALERKFFH...Q..KQYLSIAERAEFSSSL...SLTETQVKIWFONRRKAKKR
RKARTIFTDKLQLEENTFE...K..QKYLSDQDRMDLAHRM...GLTDTQVKTTHYONRRTKKWR
RKARTAFSDHQLNQLERSFE...R..QKYLSDQDRMDLAAL...NLTDQVKTTHYONRRTKKWR
RKARTAFDHLQTLLEKSF...R..QKYLSDQERQELAHKL...DLSDCQKTHYONRRTKKWR
RRSRTIFTELQMLGELKFFQ...K..QKYLSTPDRDLAQSL...GLTQLQVKTTHYONRRMKWKK
RKARTVFSDDQLSGLKRF...K..QRYLSTPERVELATAL...GLSEITQVKIWFONRRMKHKK
..SRAVFSLMQRRGLEKSFQ...V..QKYVAKPERRLAEL...GLTDAQVKIWFONRRMKWRQ
RKPQAYSARQLDRLETFEQ...T..DKYLSVNRKRIQLSQT...NLTEITQKTHYONRRTKKWK
RRARTAFTYEQLVSLNFK...T..TRYLSVGERENLALS...SLTETQVKIWFONRRTKKWK
KR3RAAFSHAQVYELERRF...L..QRYLSGERADLAAL...KLTETQVKIWFONRRYKTKR
RKPRLVFSQAQVLELECFR...L..KKYLTGAERELIAQKL...NLTSATQVKIWFONRRYKTKR
RKPRLVFSQAQVYELERRF...Q..QRYLSPAERDQLASVL...KLTSTQVKIWFONRRYKTKR
RKRRLVFSQAQVYELERRF...Q..QKYLSPAERHLASMI...HLTPTQVKIWFONRRYKTKR
KKARTTFSGQVLELEKQFE...A..KVTLSSSDRSELAKRL...DVTETQVKIWFONRRTKKWK
KHTRPTFSGQVLELEKQFE...Q..TKYLAGPERARLAYS...GLTESQVKIWFONRRTKKWK
KKTRTFSRSQVLELEKQFE...V..KRYLSSEERGLAASL...HLTETQVKIWFONRRMKHKK
KKPRTSFVSFRVQICELEKRF...R..QRYLSASAERAAALAKL...KMTDAQVKIWFONRRKWR
...VRFNSDQITIELEKFF...T..QKYLSPPERKRLAKML...QLSERQVKIWFONRRKWR
RKSRTAFTNHQIYELERFL...Y..QKYLSPADDDQIAQL...GLTNAQVITWFONRRKLR
RKRTRTFSNQCINLEENFN...R..QRYLTPDORDRIAKHL...GLTNTQVITWFONRRKLR
RRTRTFTSSQIAELQHF...Q..GRYLTAPRLADLSAKL...ALGTAQVKIWFONRRRKH.
S--S---HHHHHHHHHHH...T...T...S--S---HHHHHHHHHHH...TS-HHHHHHHHHHHHHH.
PRVRTAFTMEQVRLTEGVFQ...H..HOYLSPLERKRLAREM...QLSEVQIKIWFONRRMKHKK
RRLRTAFTPQQITRLEQAFN...K..QRYLSASERKKLATS...QLSEIQVKIWFONRRMKLKR
RRLRTAFTSDQISTLEKTFQ...K..HRYLSASERKKLAAK...QLSEVQIKIWFONRRMKYKR
.PKRONKTPFQLETLERVYA...M..ETYPSEATRAELSEKL...GLTRDQLQKWFONRRKLDKN
RKPRTRAKGEALDVLKRF...I..NPTPSLVERKKISDLI...GMPKENVRIWFONRRKLRK
PKPRNPKPEQITRLEIAFN...S..GMVNPREEIRRIAQLEQYGVGDANVFYWFONRKSRSKH
KCBRWNPTEAQVKVLELFR...A..GLRTPSTEQIQIRISTHLSAFKVKESKNVYWFONHKARERH
STTRWCPTPEQLMMLLEMYR...G..GLRTPNAAIQQITIAHLSTYGRIEGKNVYWFONHKARDRQ
GGTRWNPTEQIGILEMLYK...G..GMRTPNAAIQQIETLQGLKYGKIEGKNVYWFONHKARERQ
VSSRWNPTEQDQIRVLEELYR...Q..GTRTPSADHIQQITIAQLRRYGKIEGKNVYWFONHKARERQ
SSSRWNPTEQDQITLLENLYK...E..GIRTPSADQIQQITIRLRYGKIEGKNVYWFONHKARERQ
ANARWPTKEQITAVLELYR...Q..GLRTPTEAQIQQITIRLREHGKIEGKNVYWFONHKARERQ
..SRWPTSDQITRILKDYLS...NNGVRSPTAEQIQIRISAKLRQYKIEGKNVYWFONHKARERQ
ATLWNPTEQITLLELYR...S..GTRTPTEQIQQITIAKLRKYGKIEGKNVYWFONHKARERL
RRKRRRTSSYELGILQAFD...E..CPTPNKAKRIELSEQ...NMSEKSVQIWFONKROAAK
RKRKRRTSSQELSLQAEF...K..CPAPSKEKRIELAES...HMTKAVQIWFONKROAVKR
KKKRKRTSPDQLKLEKIFM...A..HQPMLNLRSQLAVEL...HMTARSVQIWFONRRAKARN
KKQR...LTADQLAYLREFS...K..DTNPPPAIREKIGREL...NTPERSVQIWFONRRAKSKL
KKQRVLFSEEQKEALRLAFA...L..DPYPIVGTIEFLANEL...GLATRTITVFNHNRMLKQ
KRPVVLAPAEKEALRAYQ...L..EPYPSQOITIELLSFQL...NLKTNITVFNHNYRSRMR
KRPVVLAPAEKEALRAYQ...Q..KPYPSPTKIEELATQL...NLKSTVFNHNYRSRMR
KKTSPTEHETAVMMALE...I..NKSPIHNEVQKLAVQL...NLGYRSVAFVFNHNRKAREK
KRRTRLKKEQADIKTFD...N..DDYPTKDDKETLANRL...GMSYCAVTIWFONKROEKKR
RKPRTAFTMEQVRLTEGVFQ...H..HOYLSPLERKRLAREM...QLSEVQIKIWFONRRMKHKK

```

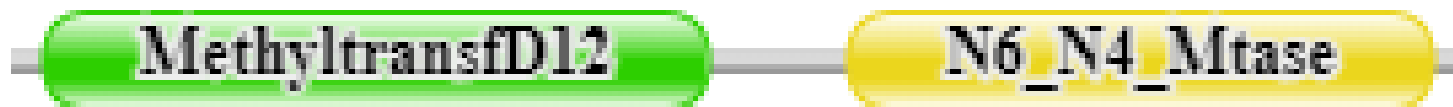
Гомеодомен (зелёный) представлен и в негомологичных белках, как следует из разнообразия доменных архитектур с гомеодоменом

Как выровнять эти две
последовательности?

There are 9 sequences with the following architecture:

MethyltransfD12, N6_N4_Mtase

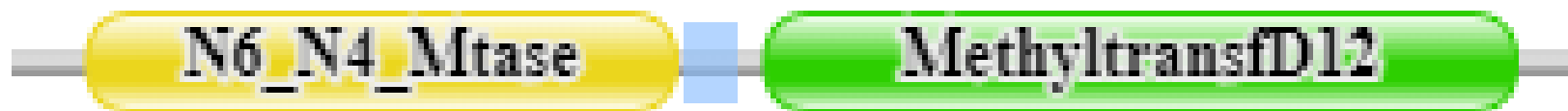
[A0A2Z5QVW5](#) [9MICC](#) [**D12-N6_N4**]



There are 5 sequences with the following architecture:

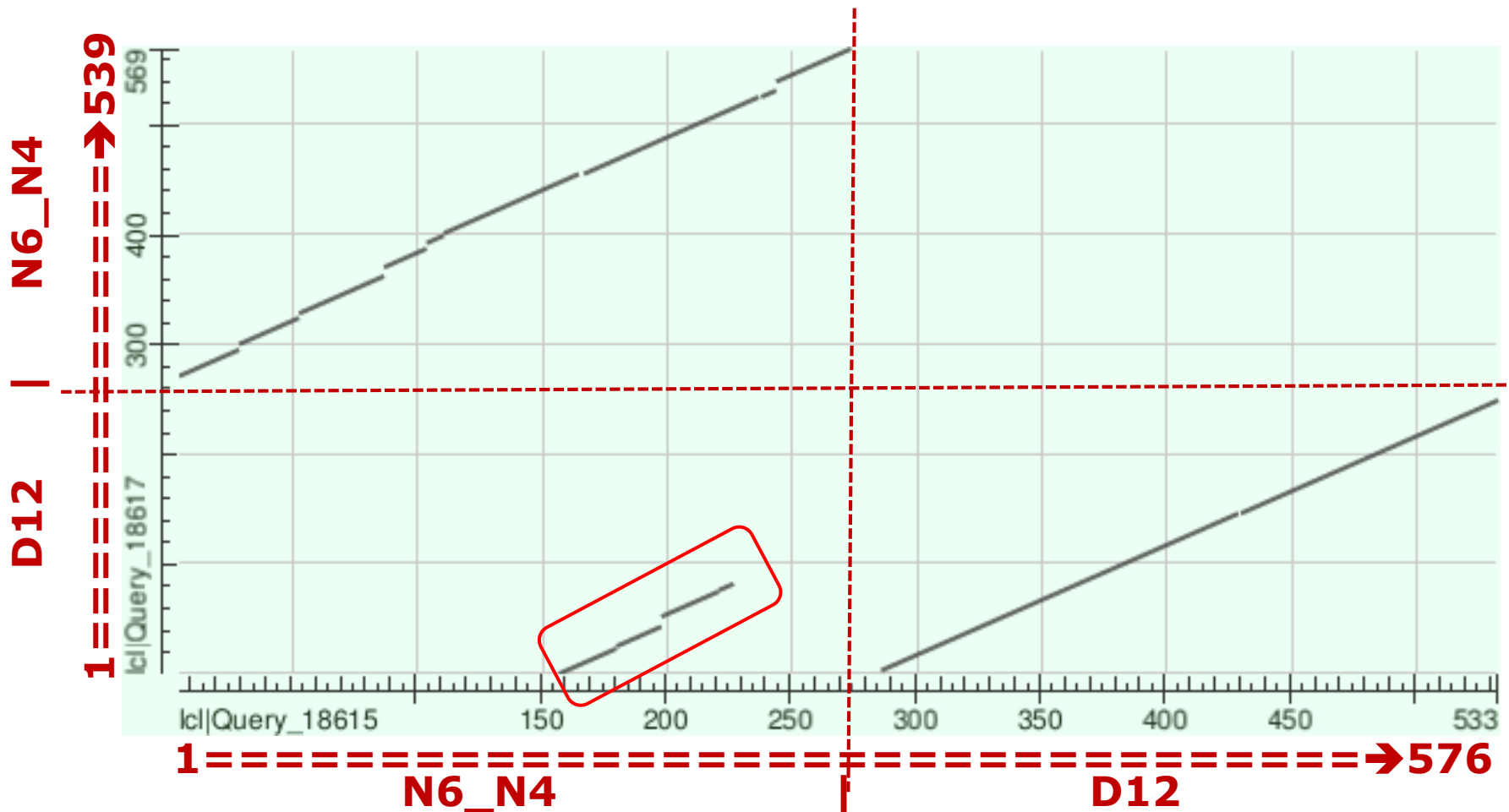
N6_N4_Mtase, MethyltransfD12

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



Как такое может возникнуть?

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



Программа BLASTp. Визуализация Dot Plot

Вопрос. Что значит лишняя диагональ ?

JaView – редактор выравниваний

Демонстрация параллельно выполнению упражнений

Список умений в Jalview

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