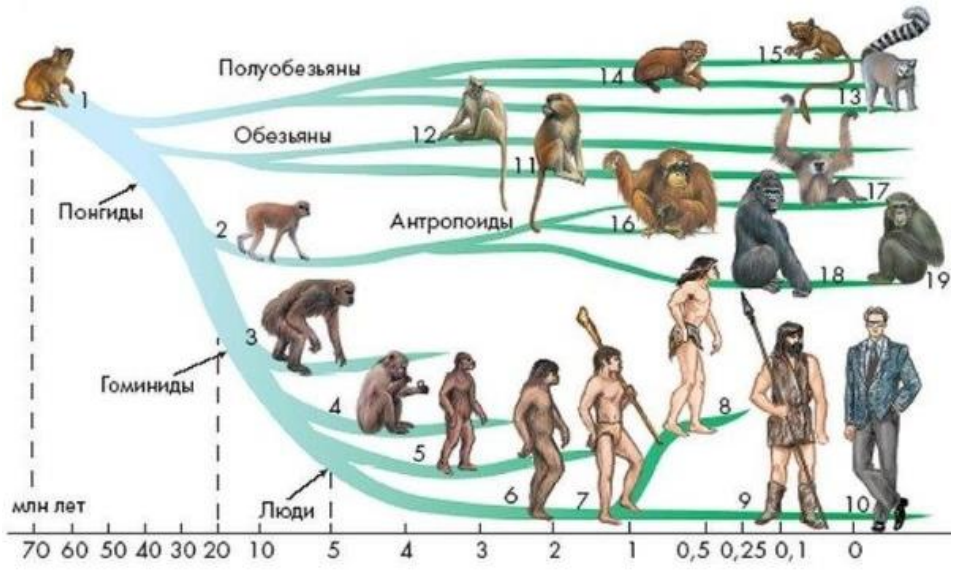
Гомология и выравнивание

Множественное выравнивание последовательностей гомологичных белков

I. Гомология и сходство

1. Общность происхождения. Эволюция



В словарик:

*** 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

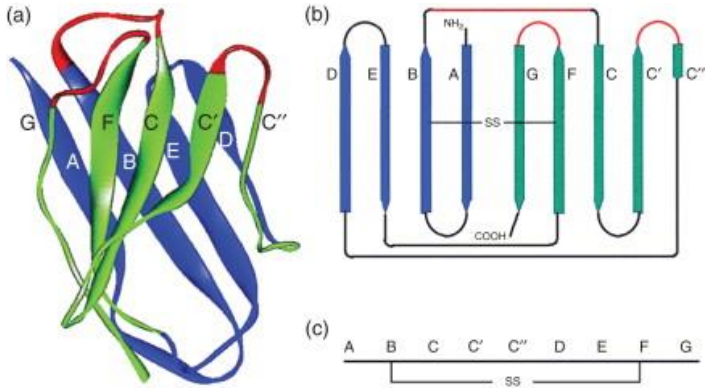
- Гомология того, что закодировано в геномах – генов **белков** и РНК, и просто участков ДНК
- **Белки гомологичны**, если их гены произошли из гена их общего предка.
- При каждом делении клетки ген дочерней клетки – копия гена материнской клетки.
- Ошибки – мутации – в ДНК случаются. Самая распространённая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

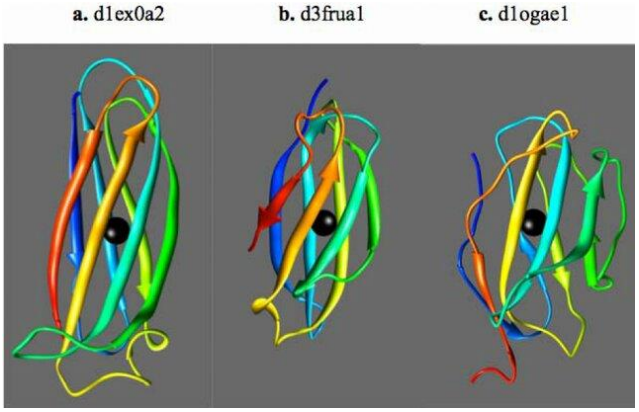
Гомология в молекулярной биологии

- О гомологии белков судят по сходству их последовательностей
- ГОМОЛОГИЧНЫЕ белки, давно разошедшиеся от общего предка, МОГУТ иметь отличающиеся функции
- ГОМОЛОГИЯ ≠ 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 ≠ ГОМОЛОГИЯ Бывает случайное совпадение (аналогия). Чем длиннее фрагмент послдовательности, тем менее вероятно случайное совпадение

Иммуноглобулиновая укладка (ImFold)





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

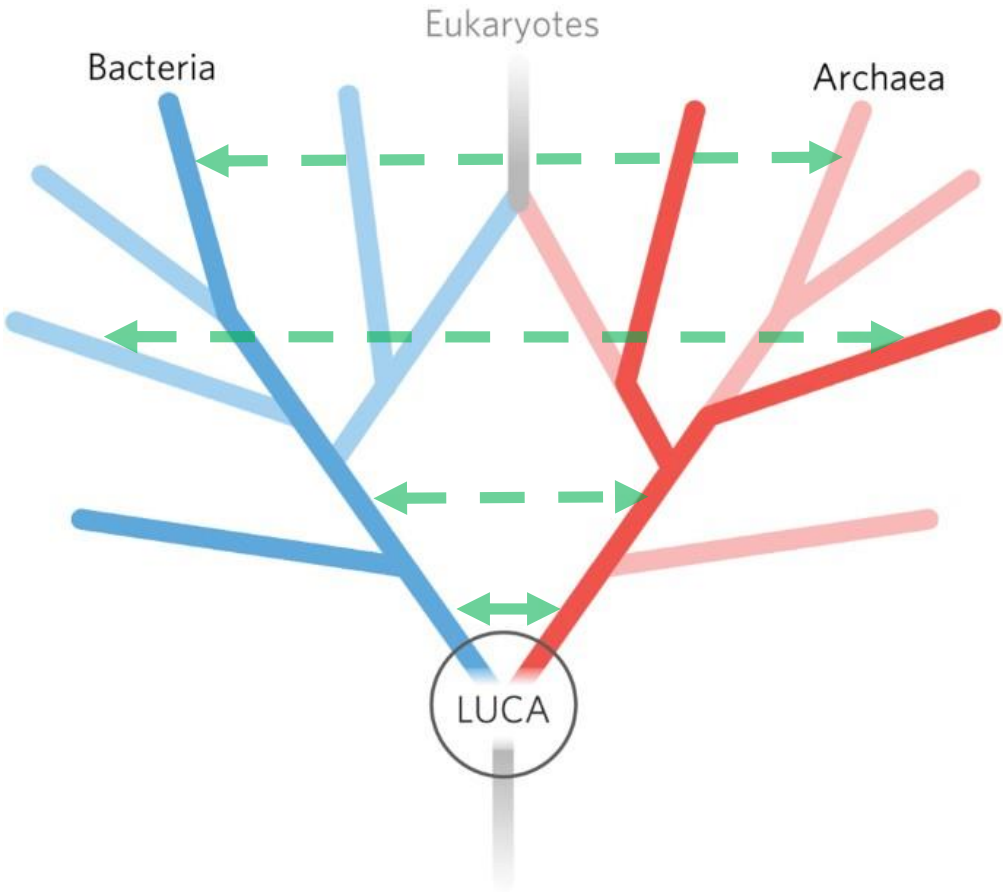
ImFold встречается во многих белках одного организма, связанных с иммунитетом и не только, и у всех животных

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

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.

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)))

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**GAATGC**T**CTGAA	замены
правнук	TATG**C**GAAT**C**GC**T**CTGAA	вставка 1 п.н.
праправнуку	TATG**A**GAA**A****C**GC**T**CTGAA	замены
прапраправнук	TG**A**GAA**AC**GC**T**CTGAA	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-GC**T**CTGAA
правнук	TATG**C**GAAT**C**GC**T**CTGAA
праправнуку	TATG**A**GAA**A****C**GC**T**CTGAA
прапраправнук	T--G**A**GAA**AC**GC**T**CTGAA

Выравнивание: гомологичные нуклеотиды - друг под другом

Идеальное выравнивание потомков относительно общего предка

Синий – вставка
Красный – замена
- Делеция
относительно ПРЕДКА

Современные ПОТОМКИ

```
ПРЕДОК          TAT--GCGAAT-GCCCTGAA
===========     =================================
праправнук1  1.TAT--GCAAAACGCTCGGAA   4 замены, 1 вставка
праправнук2  2.TAT--GCGAA--GCTCTGAA   1 делеция, 1 замена
праправнук3  3.TAT--GCGTATCGCTCGGAA   3 замены, 1 вставка
праправнук4  4.TAT--GCGAAT-GC---GAA   1 делеция 3 п.н.
праправнук5  5.TATATGCAAAT-GCCCTGAA   1 вставка 2 п.н., 1 замена
```

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)

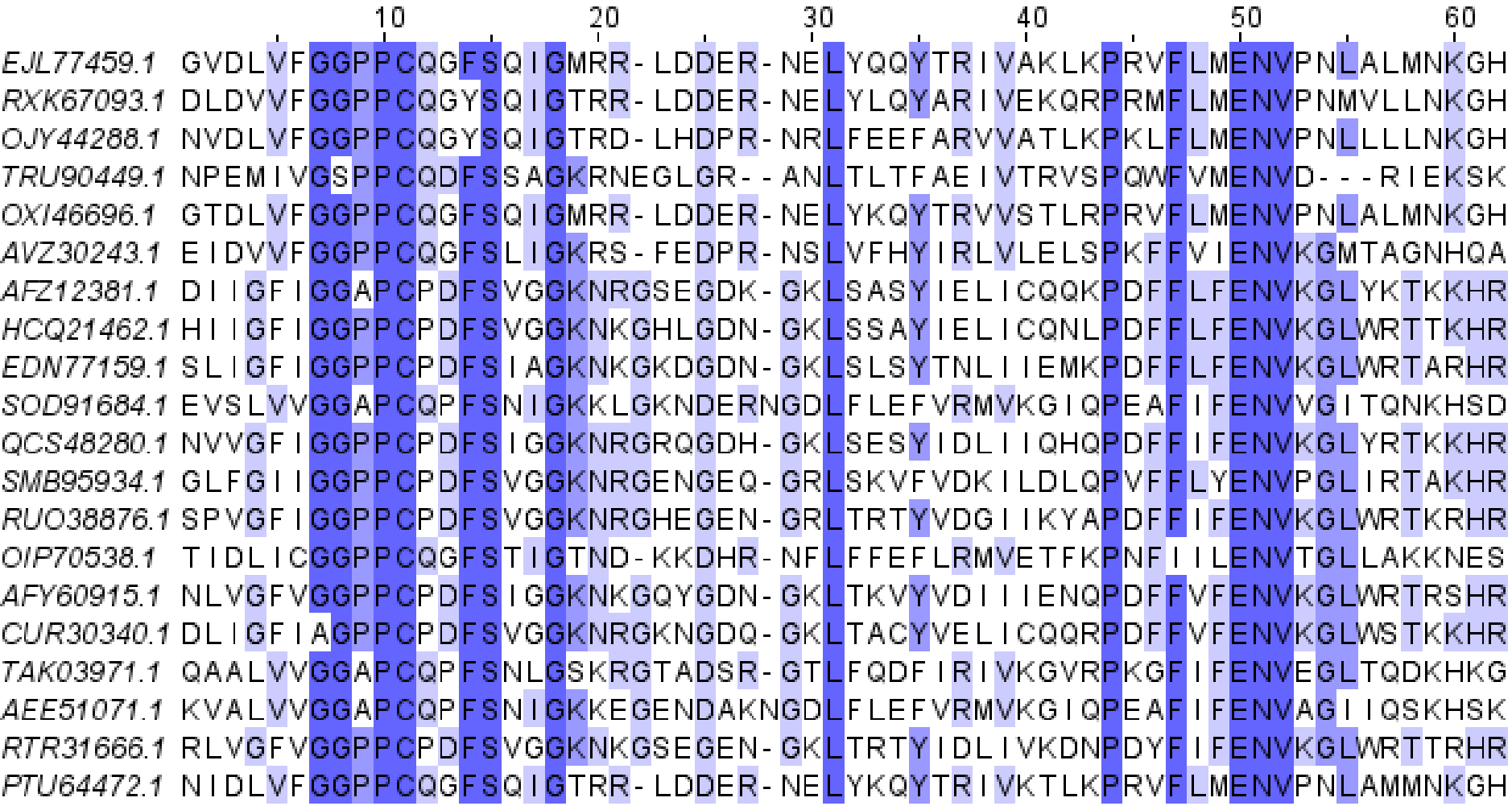
Смысл выравнивния последовательностей гомологичных белков

- 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!

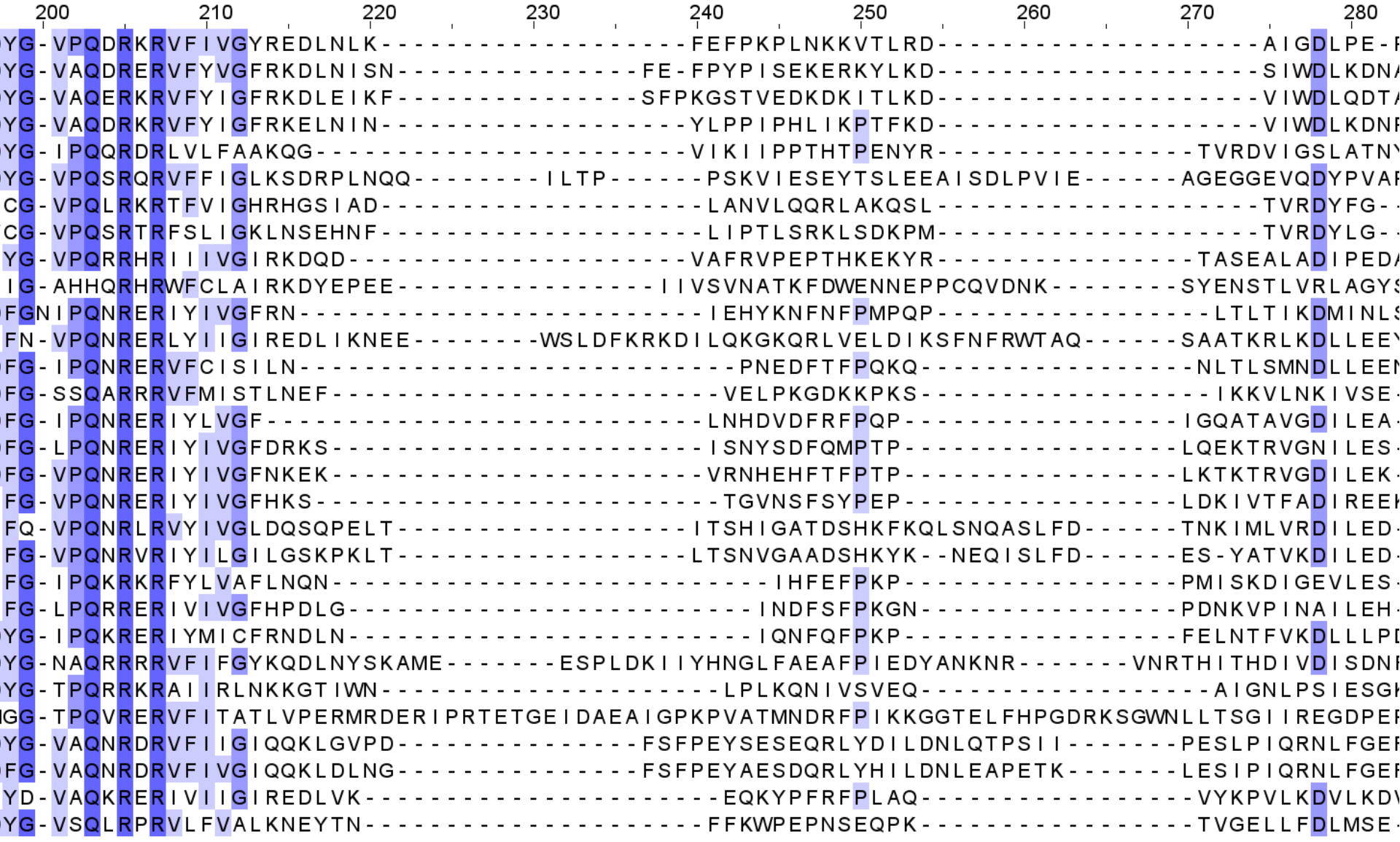
Для этого нужно набираться опыта на примерах.

Биологический смысл выравнивания



DNA (cytosine-5-)-methyltransferases

Продолжение того же выравнивания
Не всегда так хорошо как на предыдущем слайде

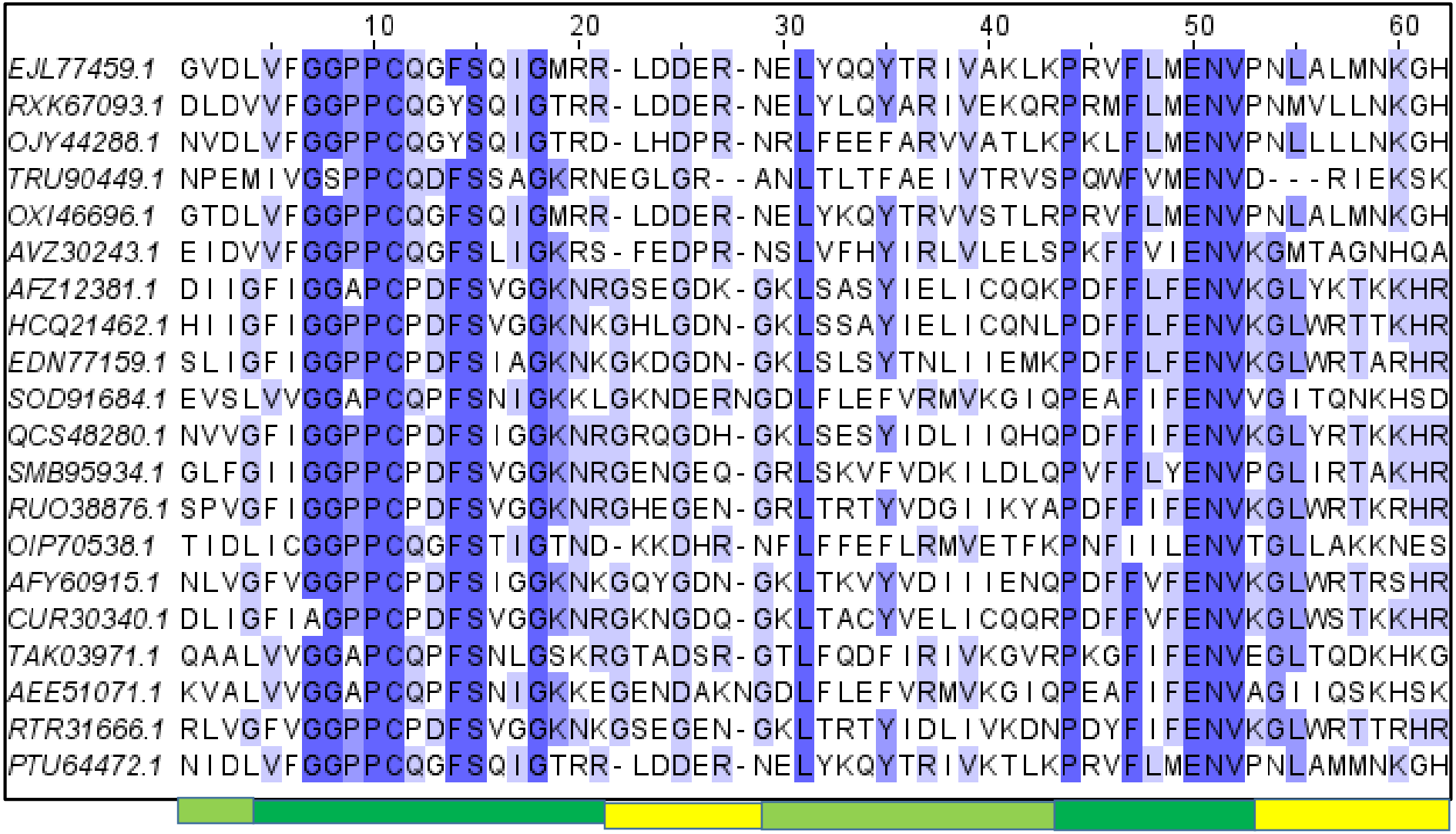


Почему такая неоднородность качества колонок в выравнивании?

- 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



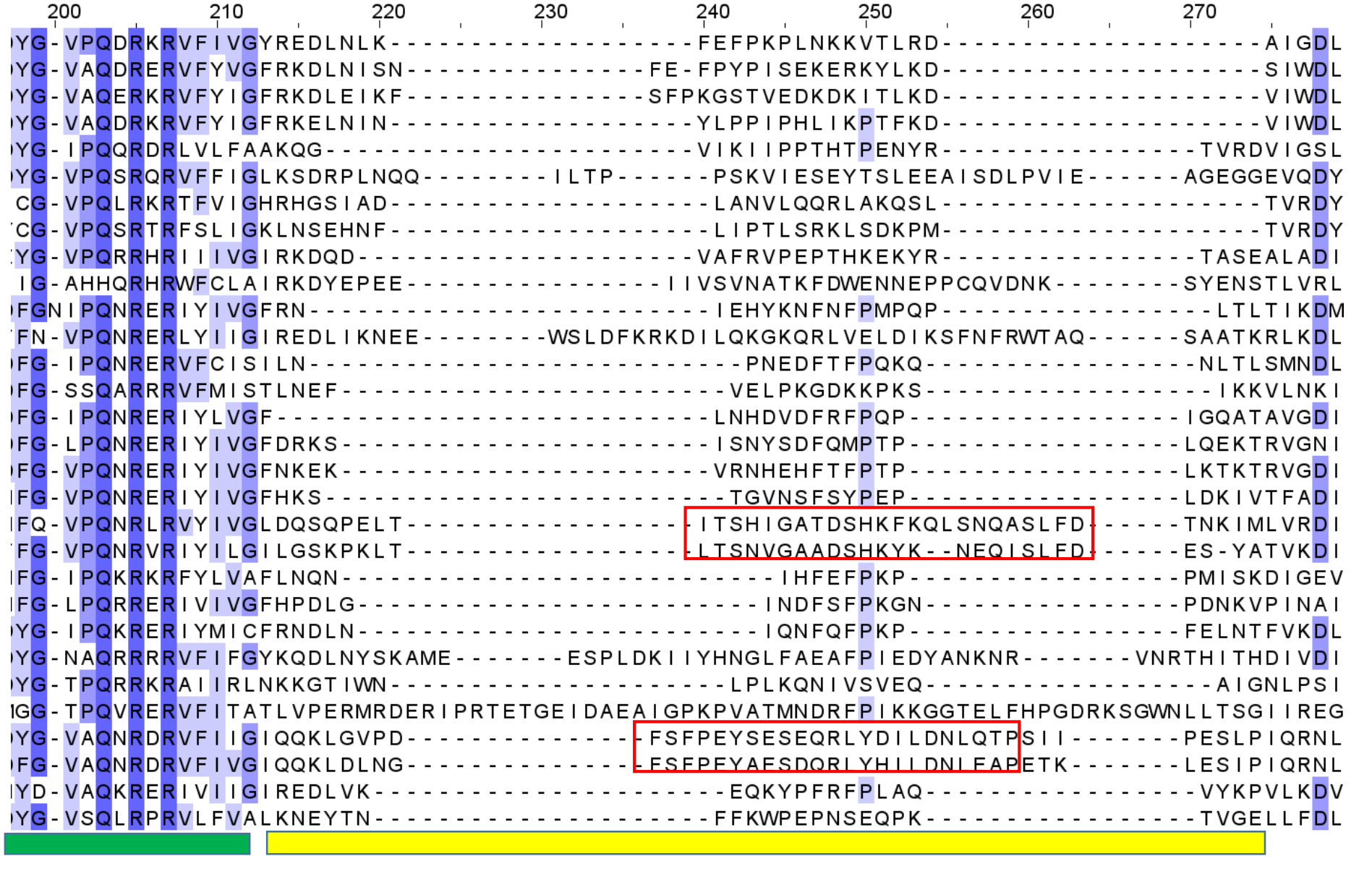
Учитывать * Есть конс. позиции

* Нет гэпов

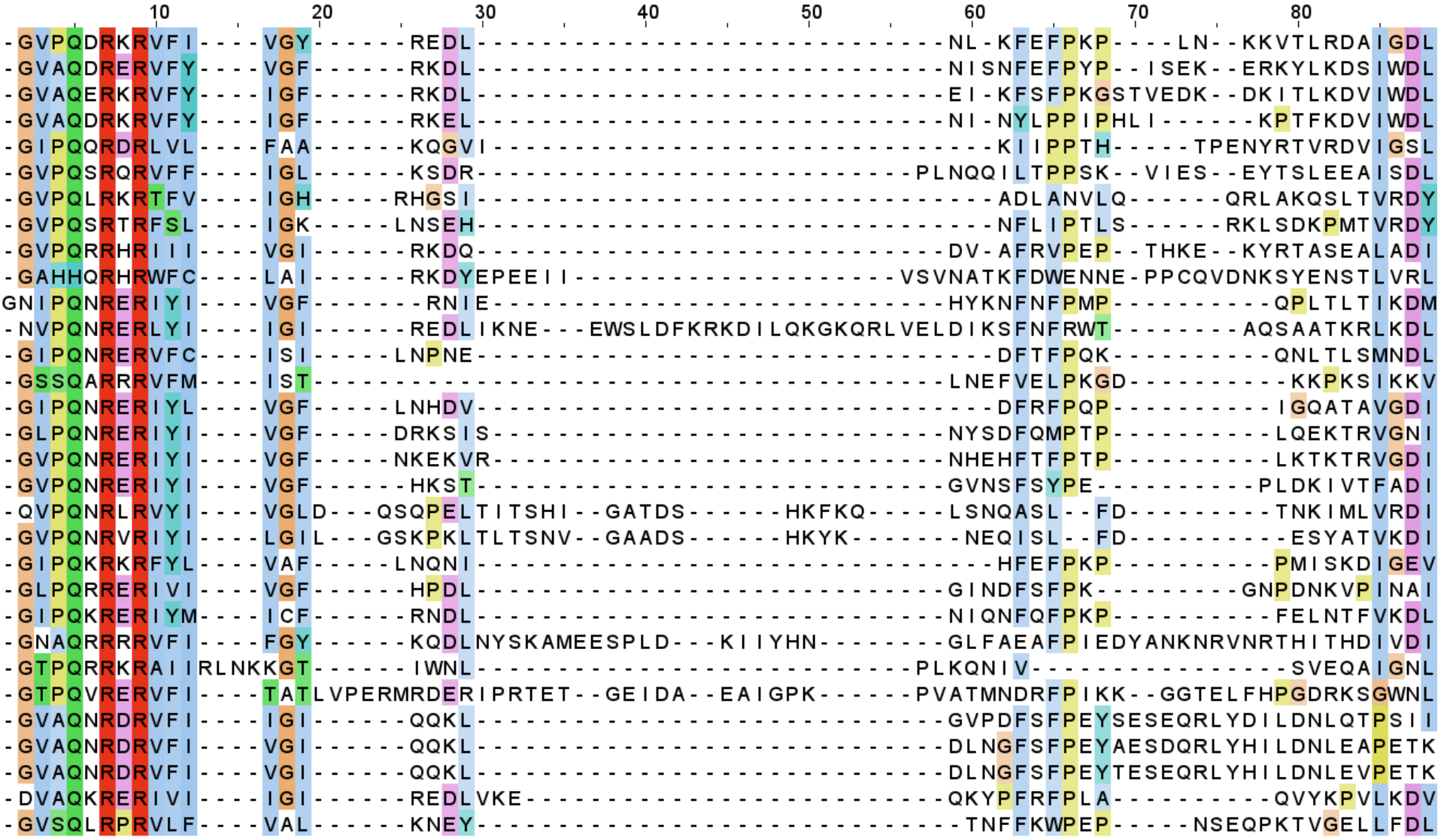
Зелёный – ОК.

жёлтый – есть выравнивания отдельных последовательностей (блоки);

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

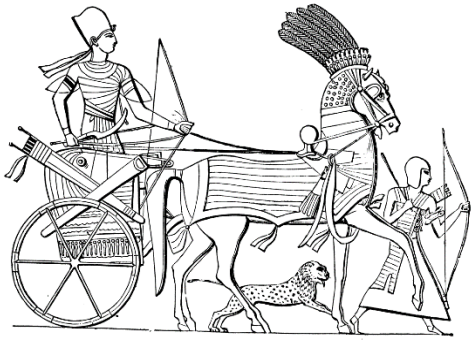


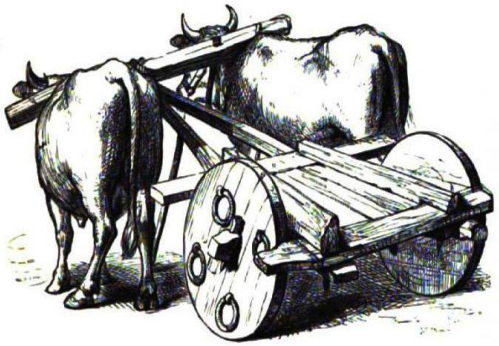
Выравнивание другой программой и раскраска по а.к.о.



Консервативное значит важное

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

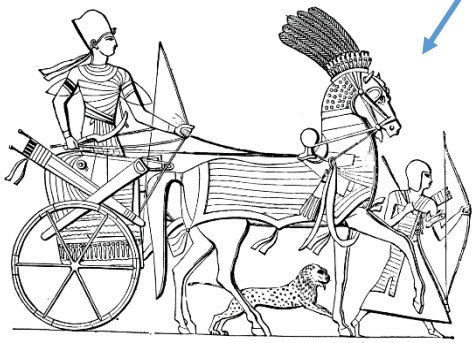


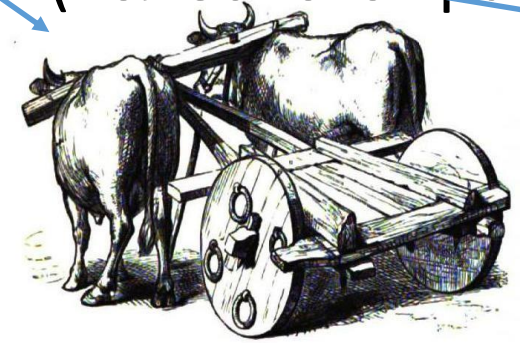



iStock
EDITORIAL USE ONLY

LUCA

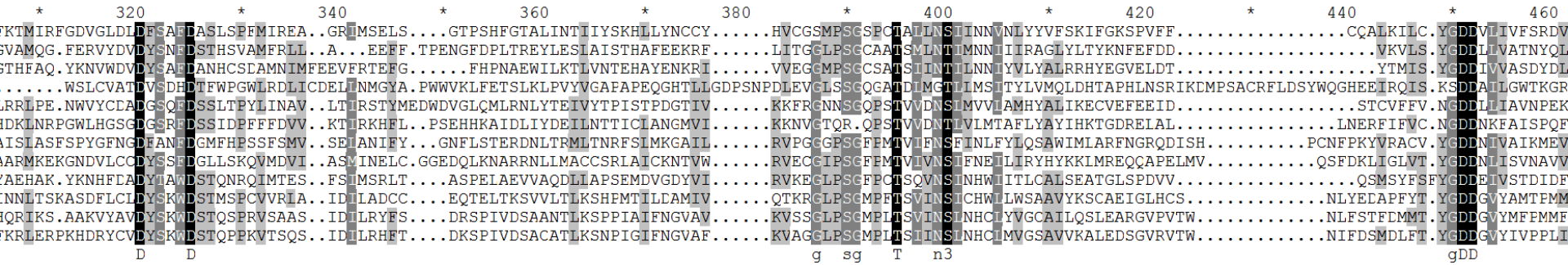
(колесного транспорта)

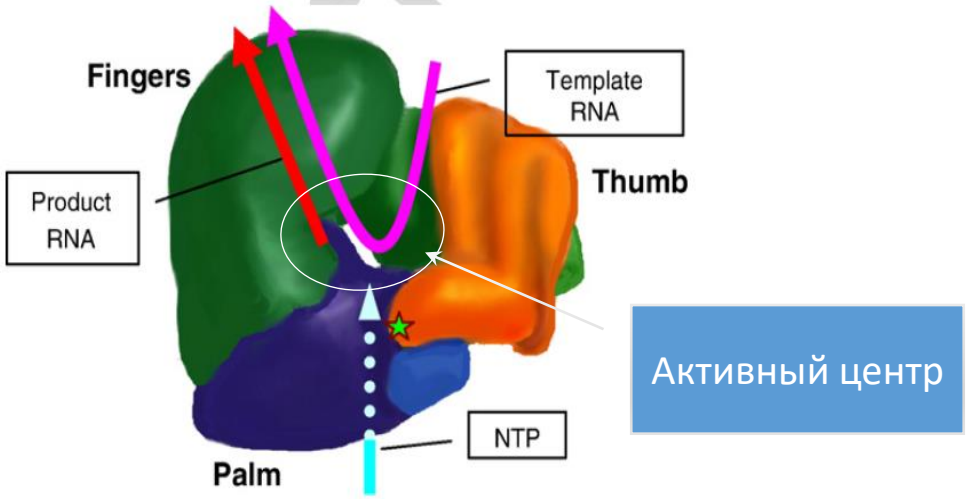




В выраниваниях белков –то же самое:
Сохраняющееся в эволюции (консервативное) – важно

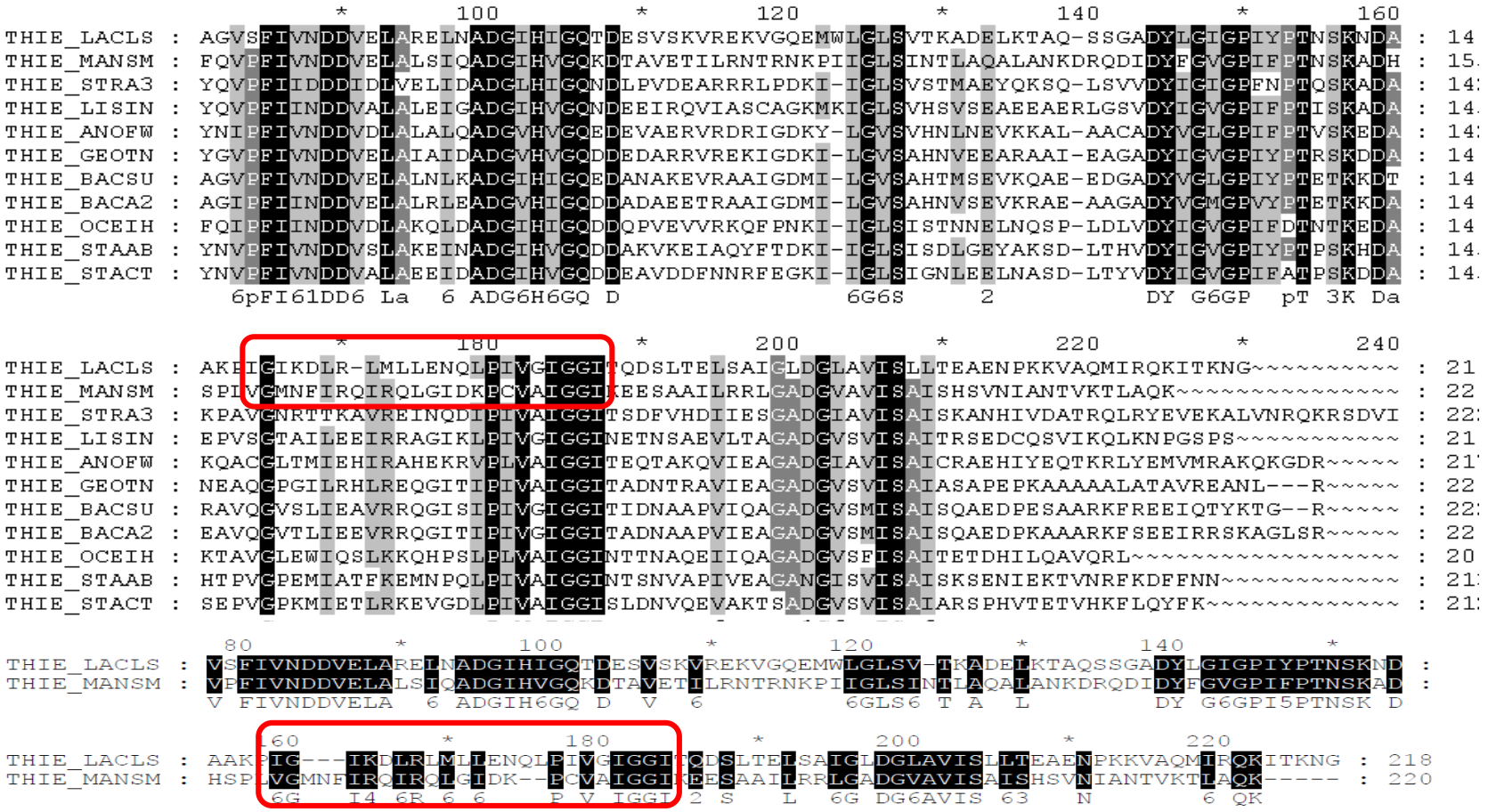
РНК зависимая РНК полимераза (RdRP), консервативные участки





На каких участках выравнивание правильное –
совпадает с эволюционным?

Множественное даёт аргументы, опровергающие оптимальное парное выравнивание. Пример.



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

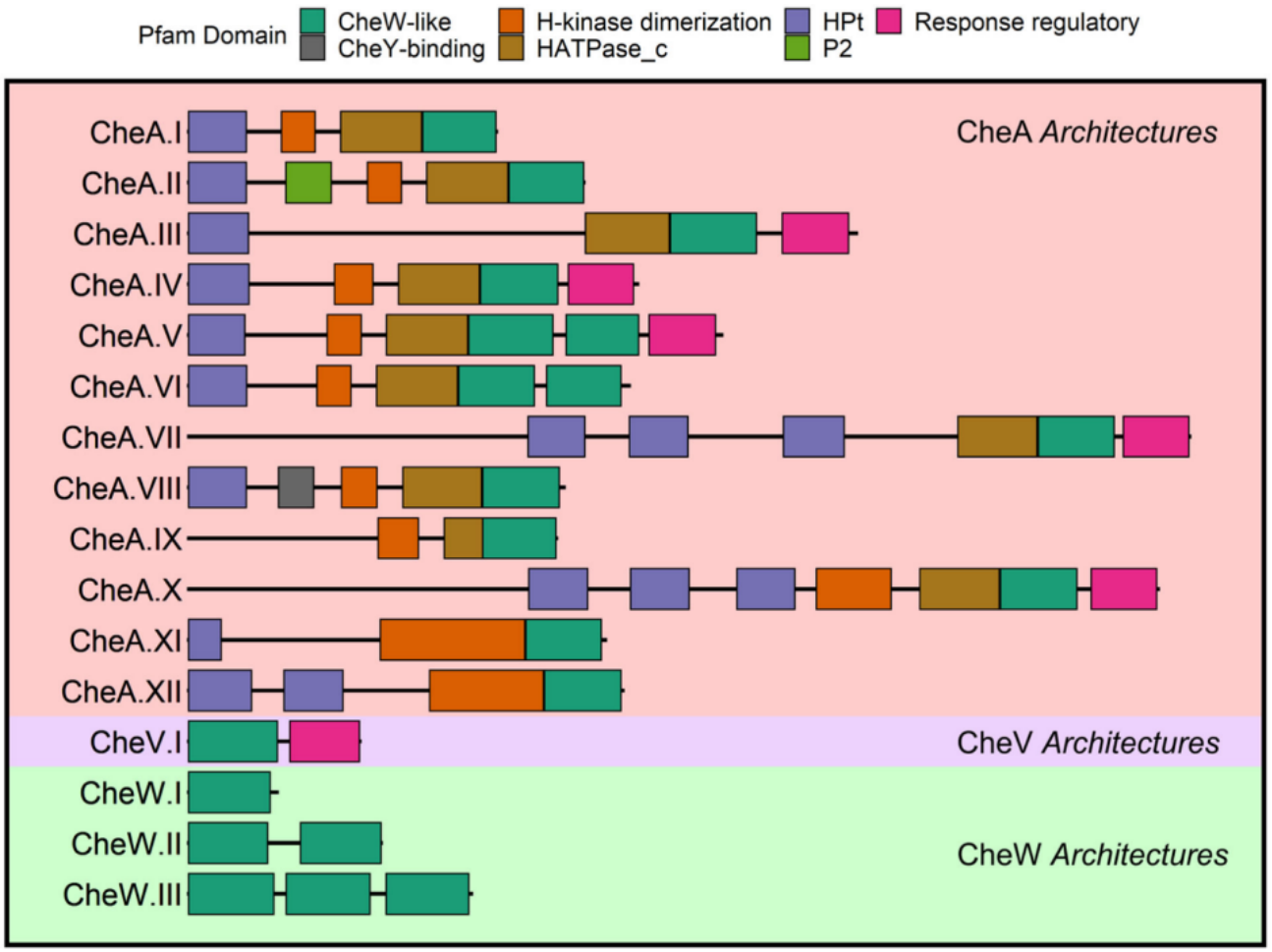
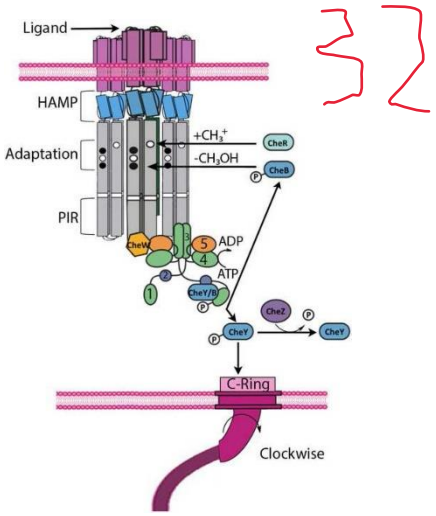


FIGURE 2. Schematic of the most abundant CheW-containing domain *Architectures* seen in nature.
Sixteen distinct architectural variants containing the CheW-like domain were identified

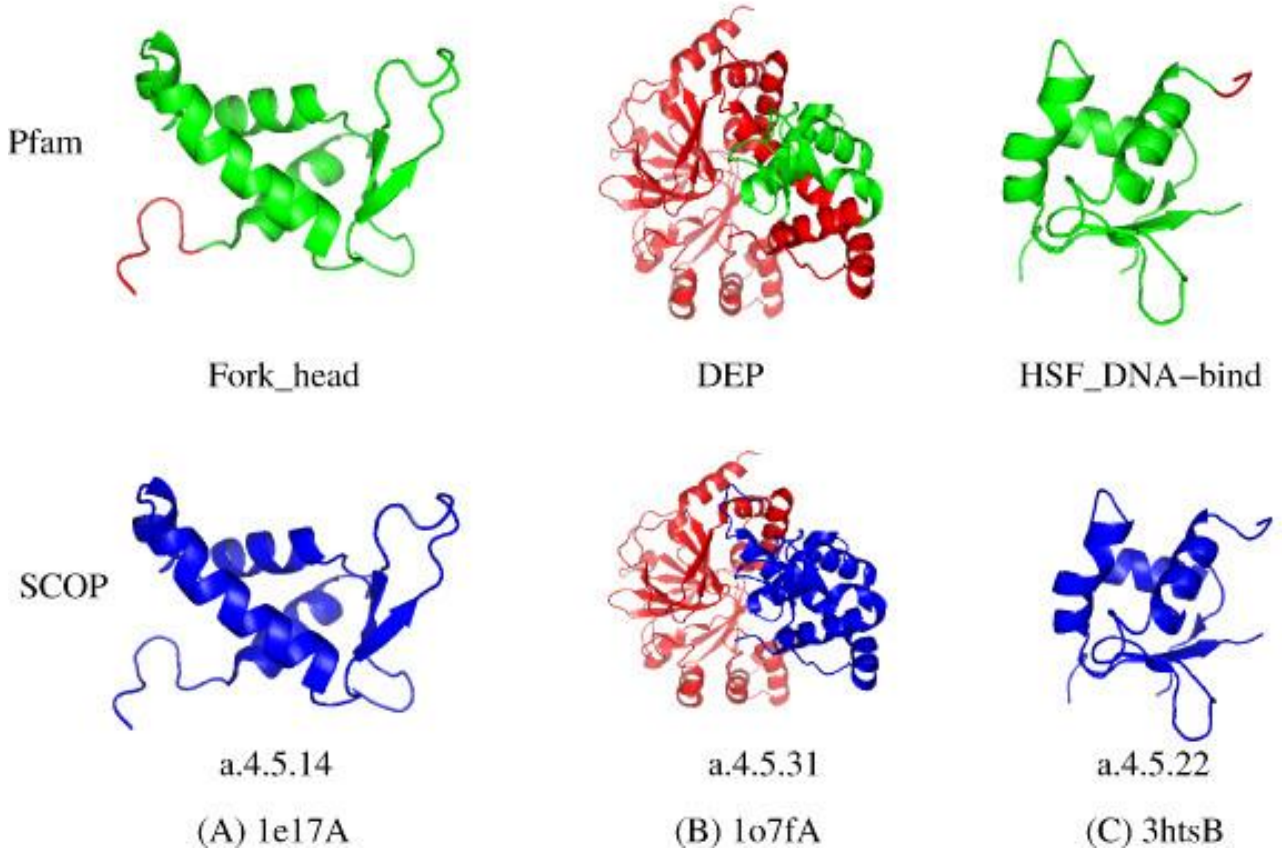


Che белки участвуют в хемотаксисе бактерий

Доменные архитектуры белков, содержащих домен CheW-like

Vass LR, Branscum KM, Bourret RB, Foster CA. Generalizable strategy to analyze domains in the context of parent protein architecture: A CheW case study. Proteins. 2022

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.

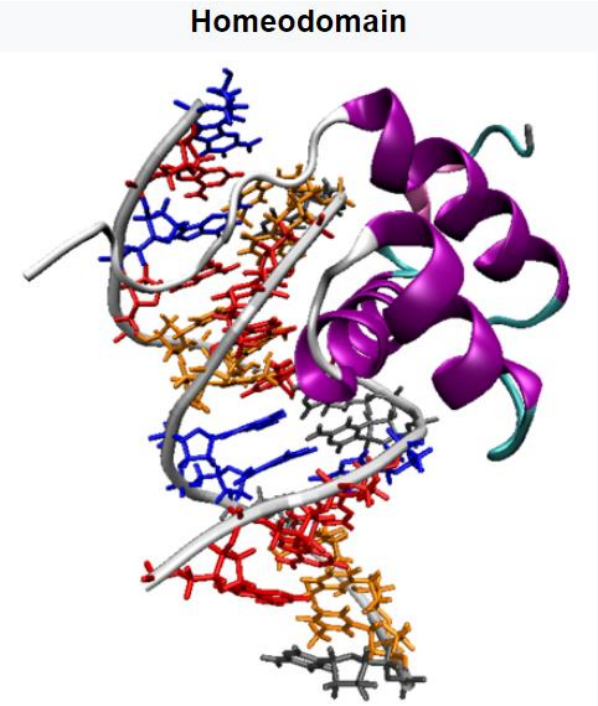
Zhang Y, Chandonia JM, Ding C, Holbrook SR. Comparative mapping of sequence-based and structure-based protein domains. BMC Bioinformatics. 2005

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

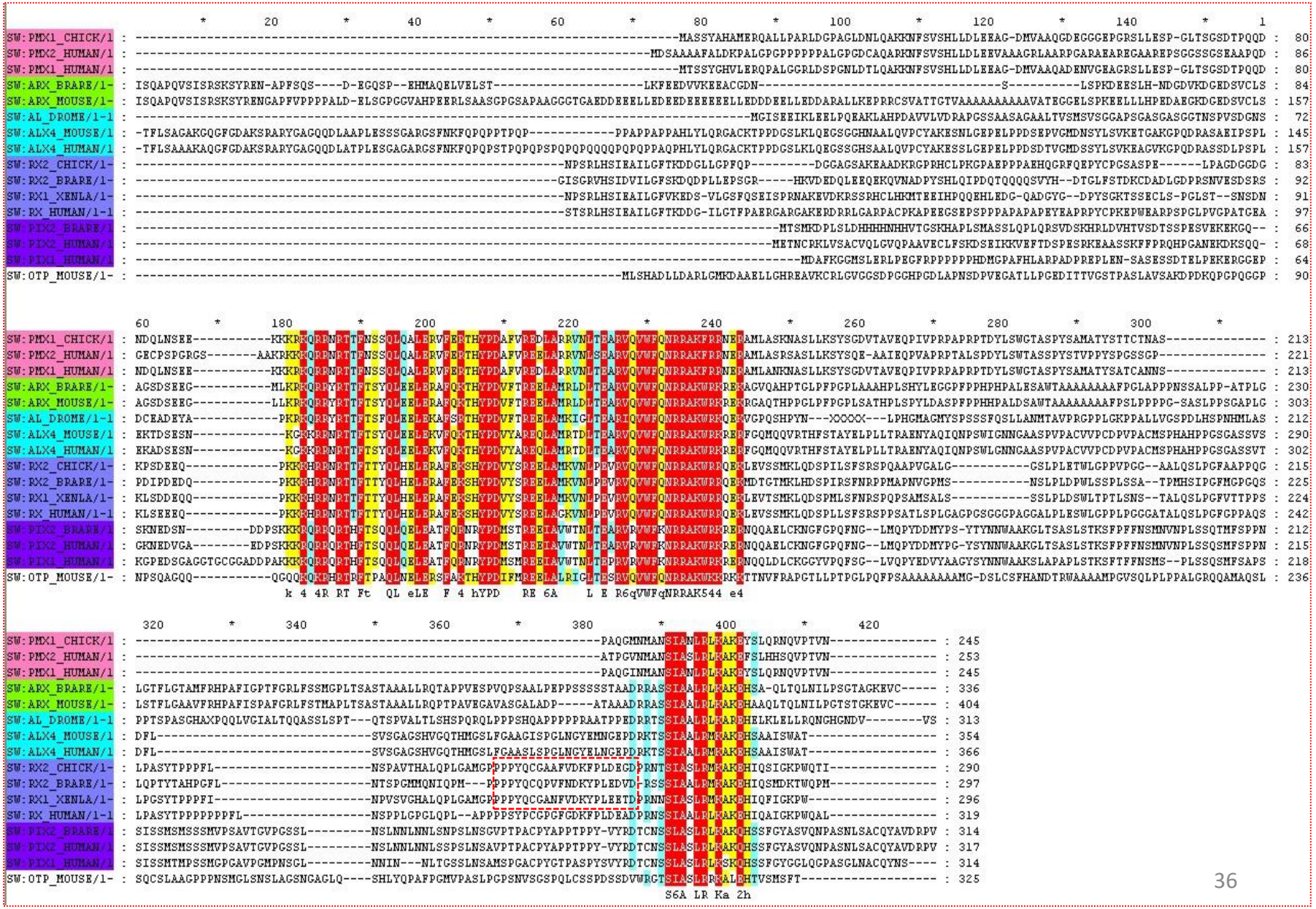
Homeodomain proteins regulate gene expression and cell differentiation during early embryonic development, thus mutations in homeobox genes can cause developmental disorders.[1]



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

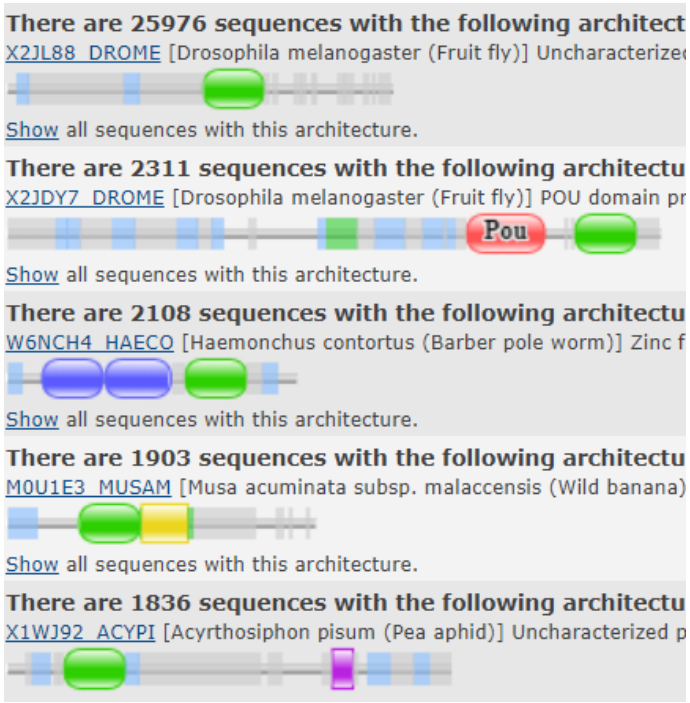


Домены принято изображать так

X1WJ92_ACYPI [Acyrthosiphon pisum (Pea aphid)]
Uncharacterized protein (408 residues)

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

Гомеодомен является ДОМЕНОМ

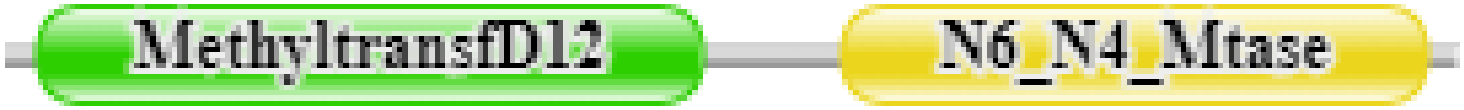


Гомеодомен (зелёный) представлен и в негомологичных белках, как следует из разнообразия доменных архитектур с гомеодоменом

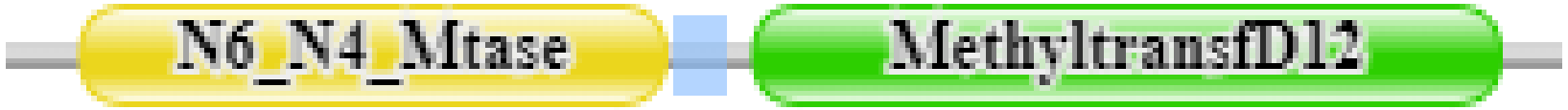
Доказательство - выравнивание

Как выровнять эти две последовательности?

There are 9 sequences with the following architecture:
MethyltransfD12, N6_N4_Mtase
A0A2Z5QVW5_9MICC [**D12-N6_N4**]

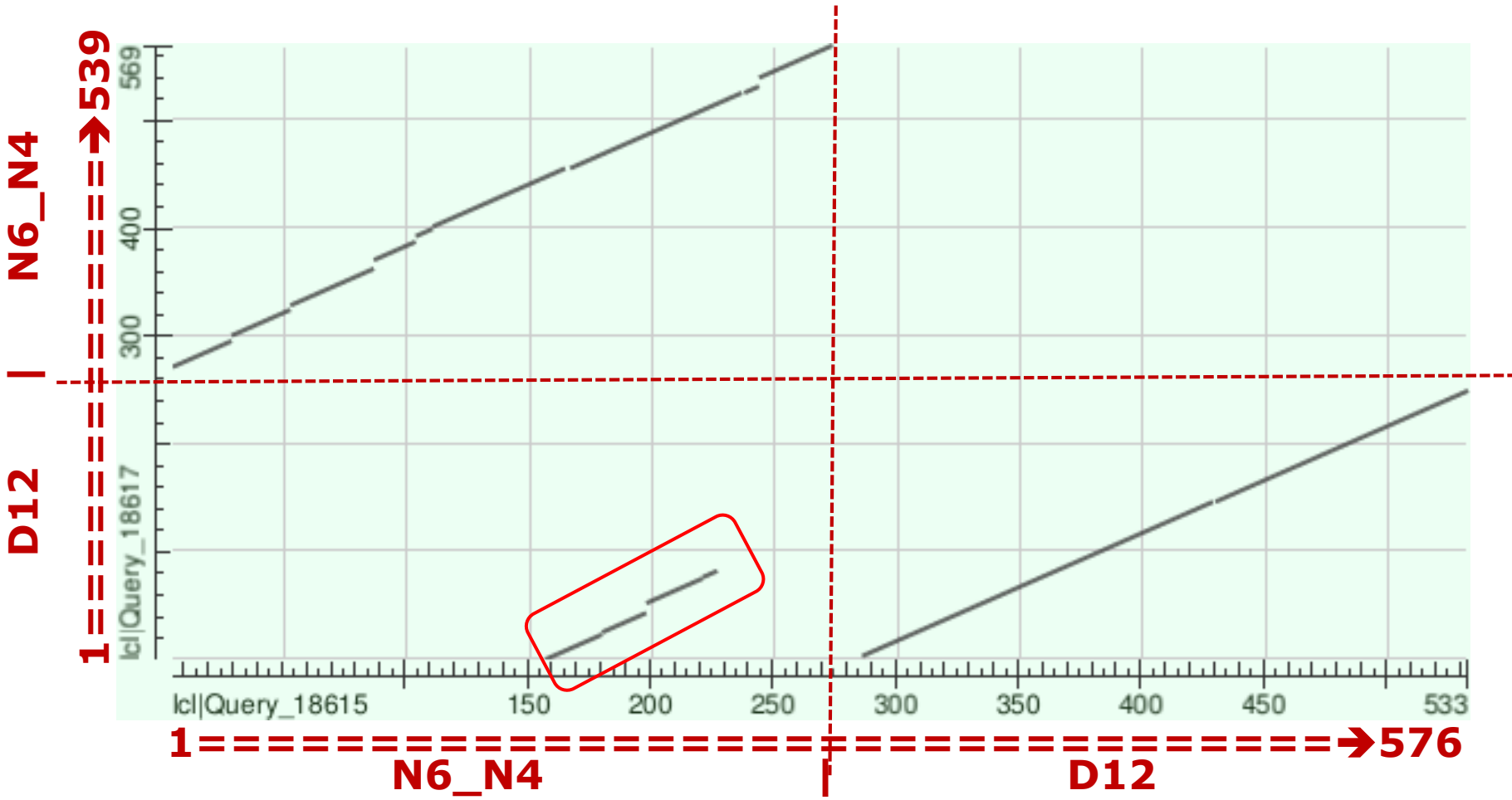


There are 5 sequences with the following architecture:
N6_N4_Mtase, MethyltransfD12
A0A1I7GYG0_9CLOT [**N6_N4-D12**]



Как такое может возникнуть?

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



Программа BLASTp. Визуализация Dot Plot

Вопрос. Что значит лишняя диагональ ?

JalView – редактор выравниваний

Демонстрация параллельно выполнению упражнений

Список умений в Jalview

Действие	меню	подменю	варианты	комментарии
Импорт последовательностей	File	Add sequences	• 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 _____ Calculate	Make groups for selection _____ 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 окноможно закрыть
	Edit	Remove empty columns		После создание окна из выделенного блока
	Edit	Remove all gaps		Для перевыравнивания
Сократить выравнивание за счёт удаления почти идентичныхскл ейки	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 использовалаось 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