Гомология и выравнивание

Множественное выравнивание последовательностей гомологичных белков

План

1. Задание
2. Гомология (3 слайда)
   a. Термин
   b. Эволюция геномов бактерий
   c. Гомологию белков и фрагментов геномов определяют по сходству
3. Эволюционное выравнивание (2)
   1. Отражение эволюции с помощью выравнивания
   2. Смысл выравнивния последовательностей гомологичных белков
4. Множественное выравнивание (10 + | Jalview demo)
   1. Анализ выравнивания
   2. Консервативное значит важное
   3. Множественное выравнивание против парного
   4. Алгоритмы.
5. Эволюционные домены
   1. Определение
   2. Примеры из Pfam
   3. Как происходят изменения доменной архитектуры
6. Крупные перестройки в геномах
   1. Следствия для белков. Карты локального сходства белков
   2. Эволюционные домены
   3. БД Pfam
7. Jalview

   a. Эволюция белков
   b. Эволюционные события
   c. Эволюционные домены

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

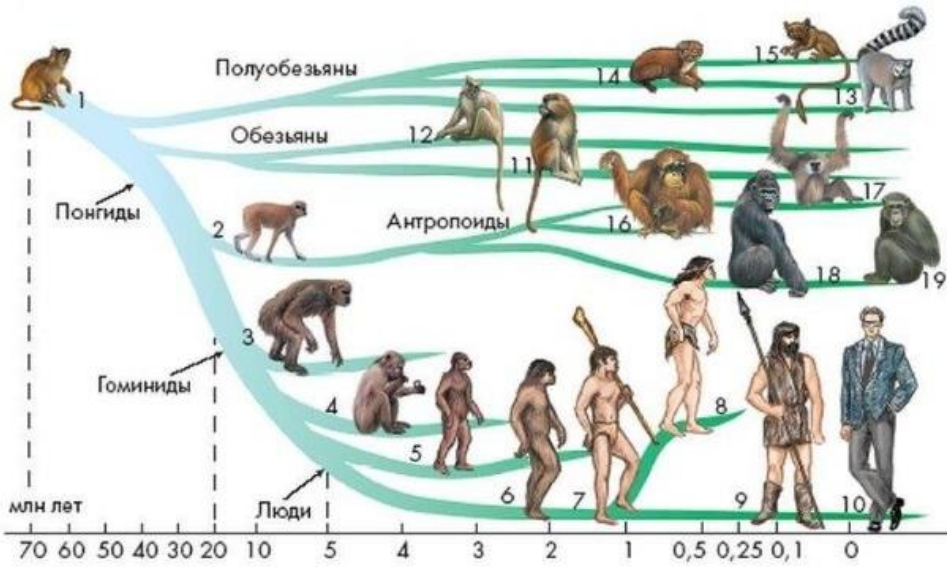
Придумайте эволюционный сценарий наблюдаемого

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

2. Гомология – наличие общего предка



В словарик:

*** Последний общий предок (LCA)**

*** Гомология**

Последний общий предок ныне живущих обезьян.

Гомоло́гия (др.-греч. ὅμοιος «подобный, похожий» + λογος «слово, закон») в биологии — сопоставимость частей сравниваемых биологических объектов, обусловленная общностью происхождения

wiki

Для целых организмов термин «гомология» не употребляют

Эволюция геномов бактерий

Половое размножение отсутствует

Небольшие, локальные изменения от поколения к поколению.
Замены нуклеотидов, короткие делеции и вставки
Изменения локальные, но их может быть много в эволюции

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

Крупные единовременные изменения генома
тем более – под отбором! Сохраняются только удачные.
Но из-за огромного количества организмов, мы их видим

*) У *Deinococcus radiodurans* частота повыше. Почему?

Гомологию последовательностей нуклеотидов и белков выводят из сходства последовательностей

Геномы и белки, как молекулы, определяются (почти *) однозначно) своей последовательностью

Поэтому их гомология определяется похожестью последовательностей

Как говорить можно, и как нельзя

- Высокая ~~ГОМОЛОГИЯ~~ последовательностей

+ Высокое СХОДСТВО последовательностей

*) почему почти?

3. 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	T	A	T	G	C	G	A	A	T	G	C	C	C	T	G	A	A	
сын	T	A	T	G	**A**	G	A	A	T	G	C	C	C	T	G	A	A	замена
внук	T	A	T	G	**C**	G	A	A	T	G	C	**T**	C	T	G	A	A	замена
правнук	T	A	T	G	**C**	G	A	A	T	**C**	G	C	**T**	C	T	G	A	A	вставка 1 п.н.
прапраправнуку	T	A	T	G	**A**	G	A	A	**A**	**C**	G	C	**T**	C	T	G	A	A	замена
прапраправнук	T	G	**A**	G	A	A	**A**	**C**	G	C	**T**	C	T	G	A	A		делеция 2 п.н.
потомок	1		4	5	6	7	8		9	9a	10	11	12	13	14	15	16	17

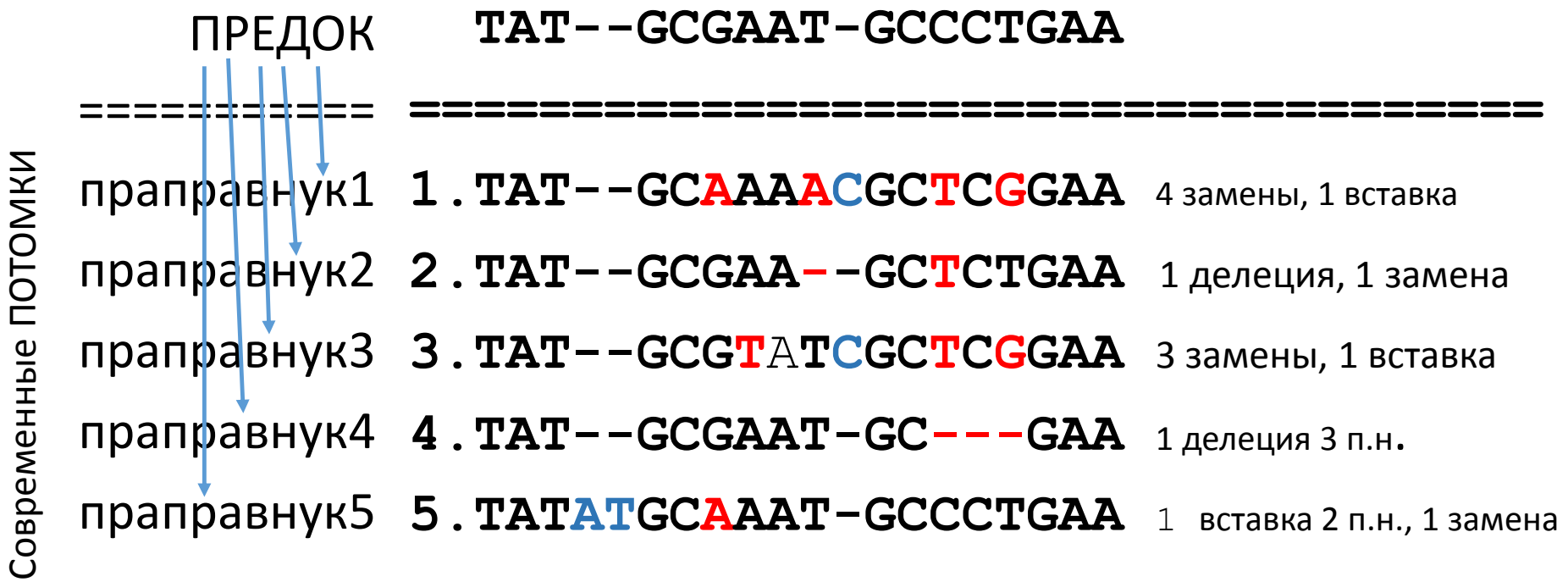
Нукл-ы потомка с номерами как у предка являются гомологами нукл-в предка

предок	T	A	T	G	C	G	A	A	T	-	G	C	C	C	T	G	A	A
сын	T	A	T	G	**A**	G	A	A	T	-	G	C	C	C	T	G	A	A
внук	T	A	T	G	**C**	G	A	A	T	-	G	C	**T**	C	T	G	A	A
правнук	T	A	T	G	**C**	G	A	A	T	**C**	G	C	**T**	C	T	G	A	A
прапраправнуку	T	A	T	G	**A**	G	A	A	**A**	**C**	G	C	**T**	C	T	G	A	A
прапраправнук	T	-	-	G	**A**	G	A	A	**A**	**C**	G	C	**T**	C	T	G	A	A

Выравнивание: гомологичные нуклеотиды - друг под другом

Идеальное выравнивание потомков относительно общего предка

Синий – вставка
Красный – замена
- Делеция
относительно ПРЕДКА



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

Сокращения
* а.к.о. – аминокислотный остаток
* aa – amino acid residue

4. Множественное выравнивание последовательностей гомологичных белков (анализ результата выравнивания)

Эволюционное выравнивание (редко достижимый идеал) : в каждой колонке стоят гомологичные аминокислотные остатки (или “-” – символы гэпа)

Смысл выравнивния последовательностей гомологичных белков

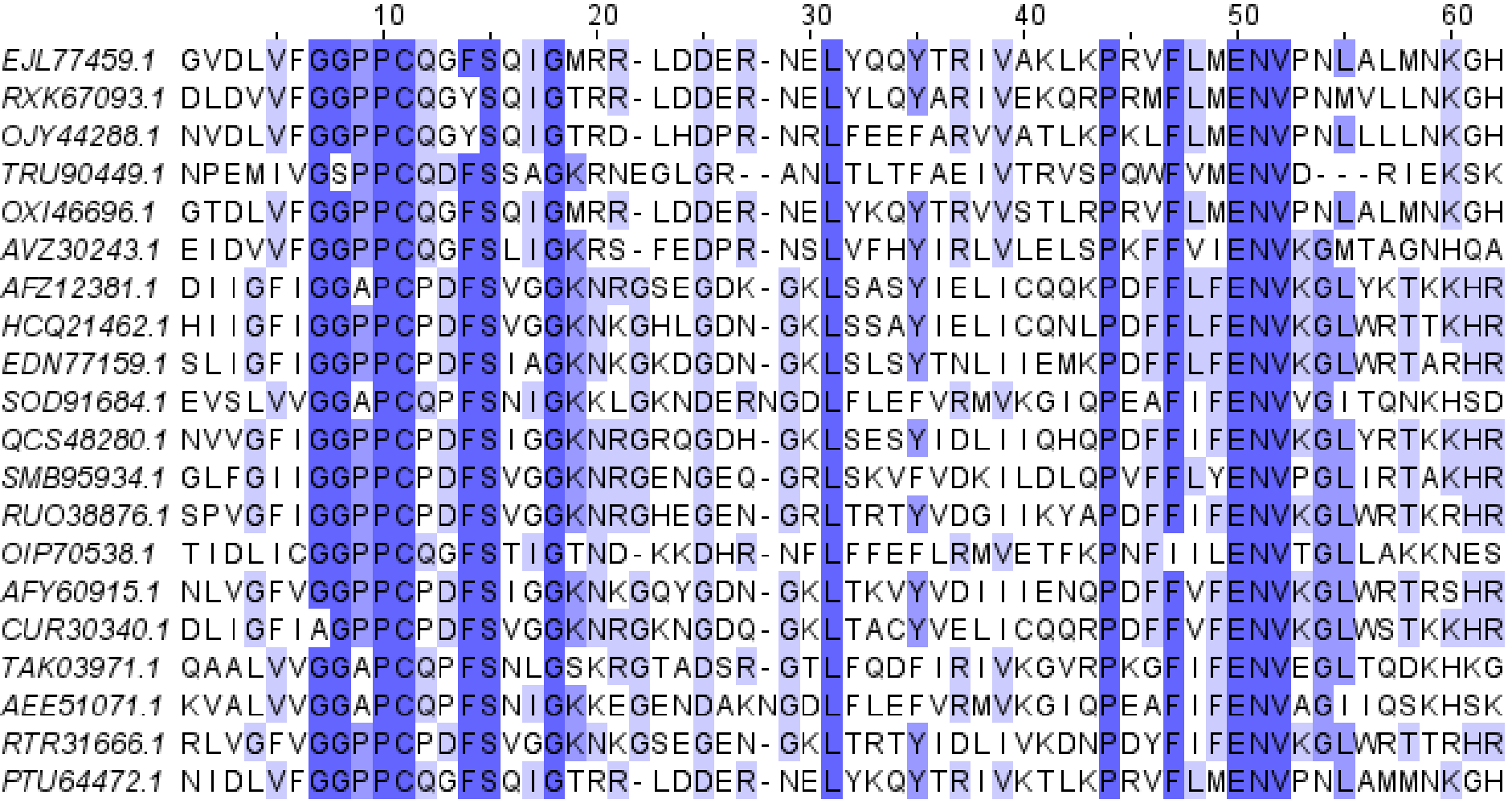
- Некоторые кодоны а.к.о. гомологичных белков потомков *произошли из одного кодона последнего общего предка* этих белков, или были делятированы в эволюции, или появились в результате вставки новых кодонов
- Цель программ множественного выравнивания *последовательностей гомологичных белков* воспроизвести 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

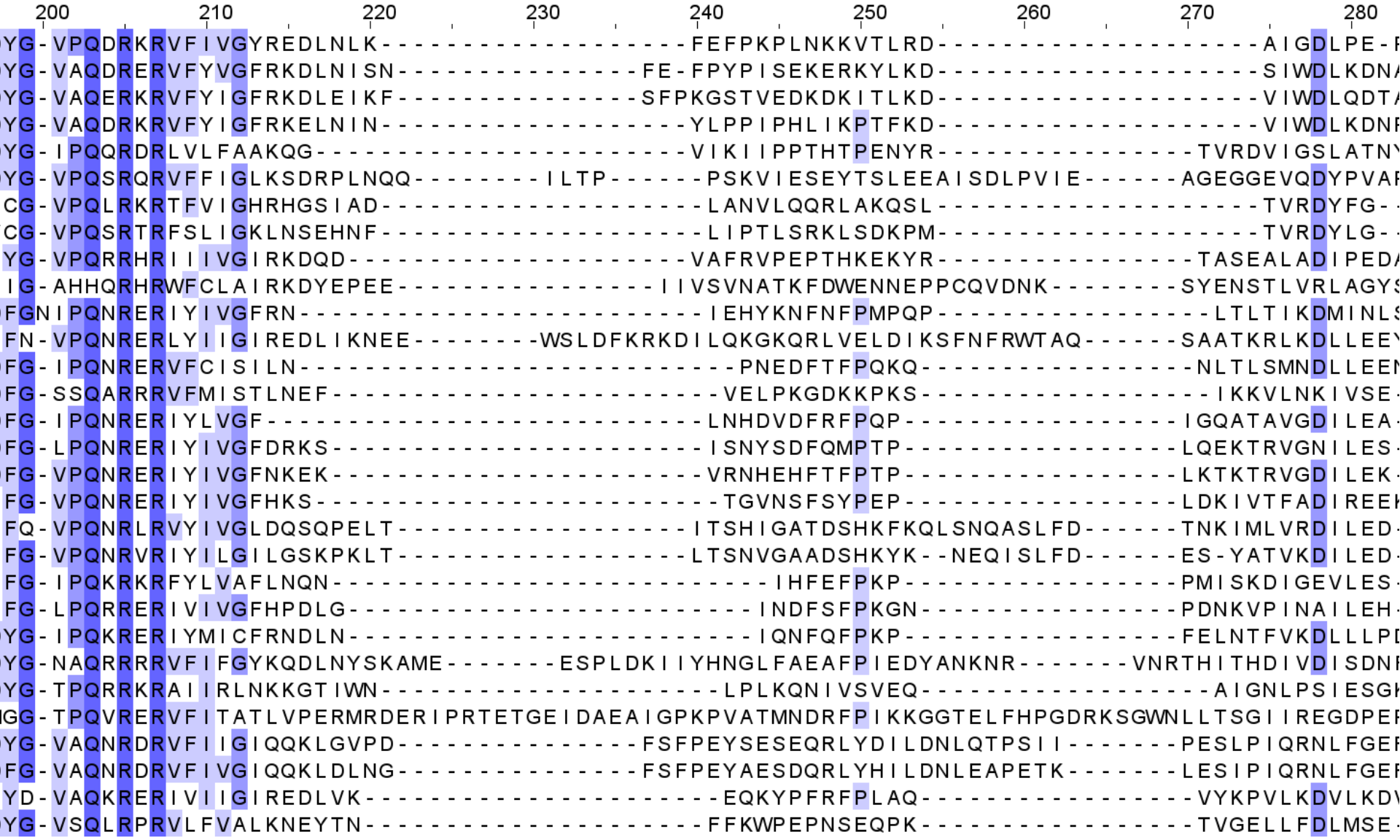
Обсуждаем смысл колонок и гэпов



DNA (cytosine-5-)-methyltransferase

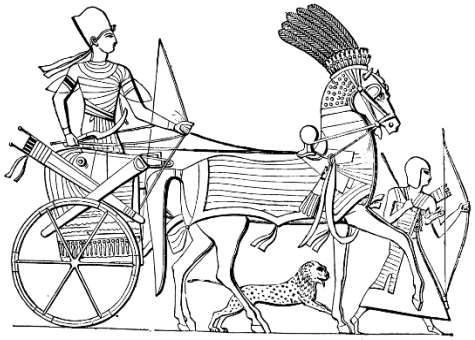
Не всегда так хорошо как на предыдущем слайде

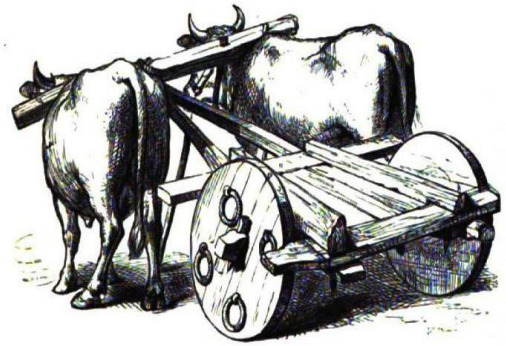
Продолжение того же выравнивания



Консервативное значит важное

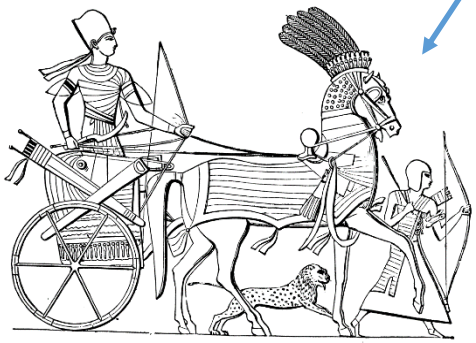
Консервативное - то, что длительно существует в эволюции, с несущественными изменениями

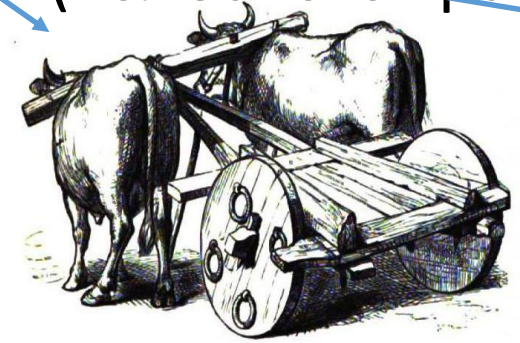




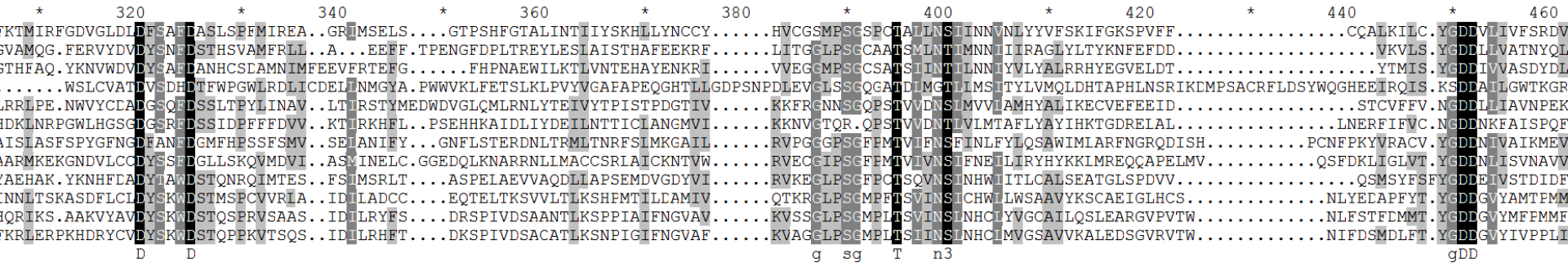
LUCA

(колесного транспорта)





РНК зависимая РНК полимераза (RdRP), консервативные участки



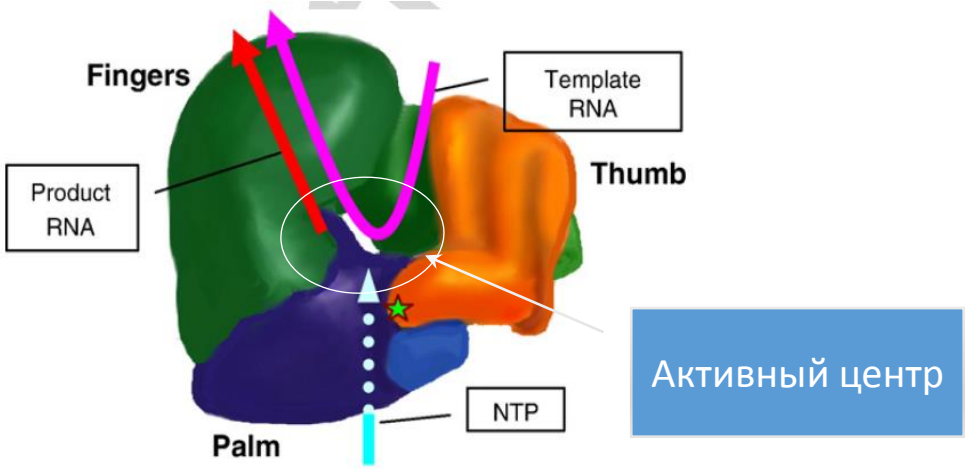


Fig. 1. Schematic architecture of "small" RdRP. The hairpin between the palm and thumb domains is in light blue. The predicted approximate location of MV RdRP Trp460 is marked by star.

На каких участках выравнивание правильное –
совпадает с эволюционным?

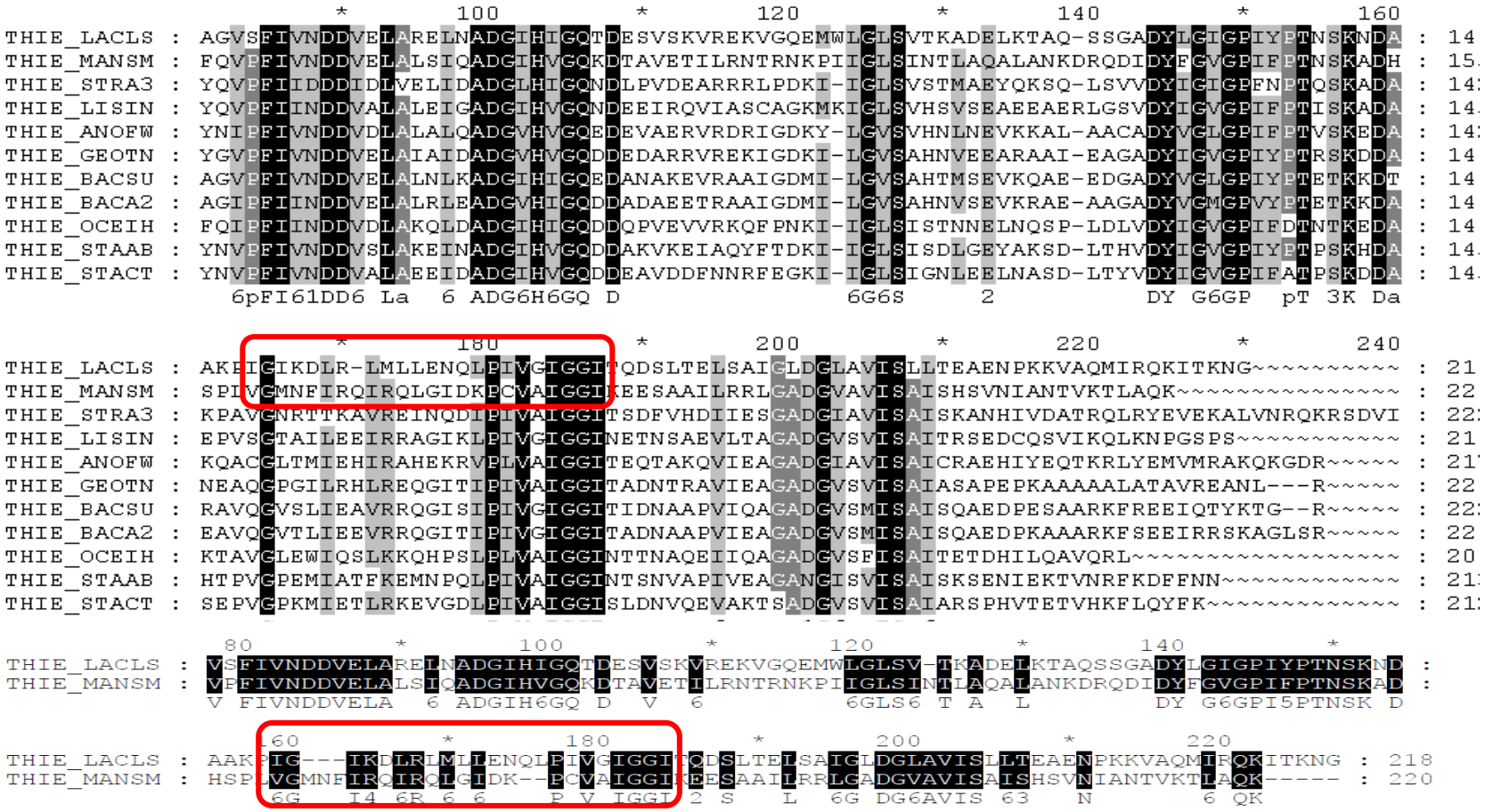
Рисунок мой :) хвастаюсь

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

Смотрим в Pfam и Jalview

- Файл PF00145_seed.fasta

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.

5. Наблюдаемый результат крупных перестроек генома: домены белков

Семейства в БД Pfam (Protein families)

Домены белков

[Длинные] гомологичные участки из разных белков, которые эволюционируют только по типу локальных мутаций,
и максимальной длины, с сохранением этого свойства,
называются

ЭВОЛЮЦИОННЫМИ ДОМЕНАМИ

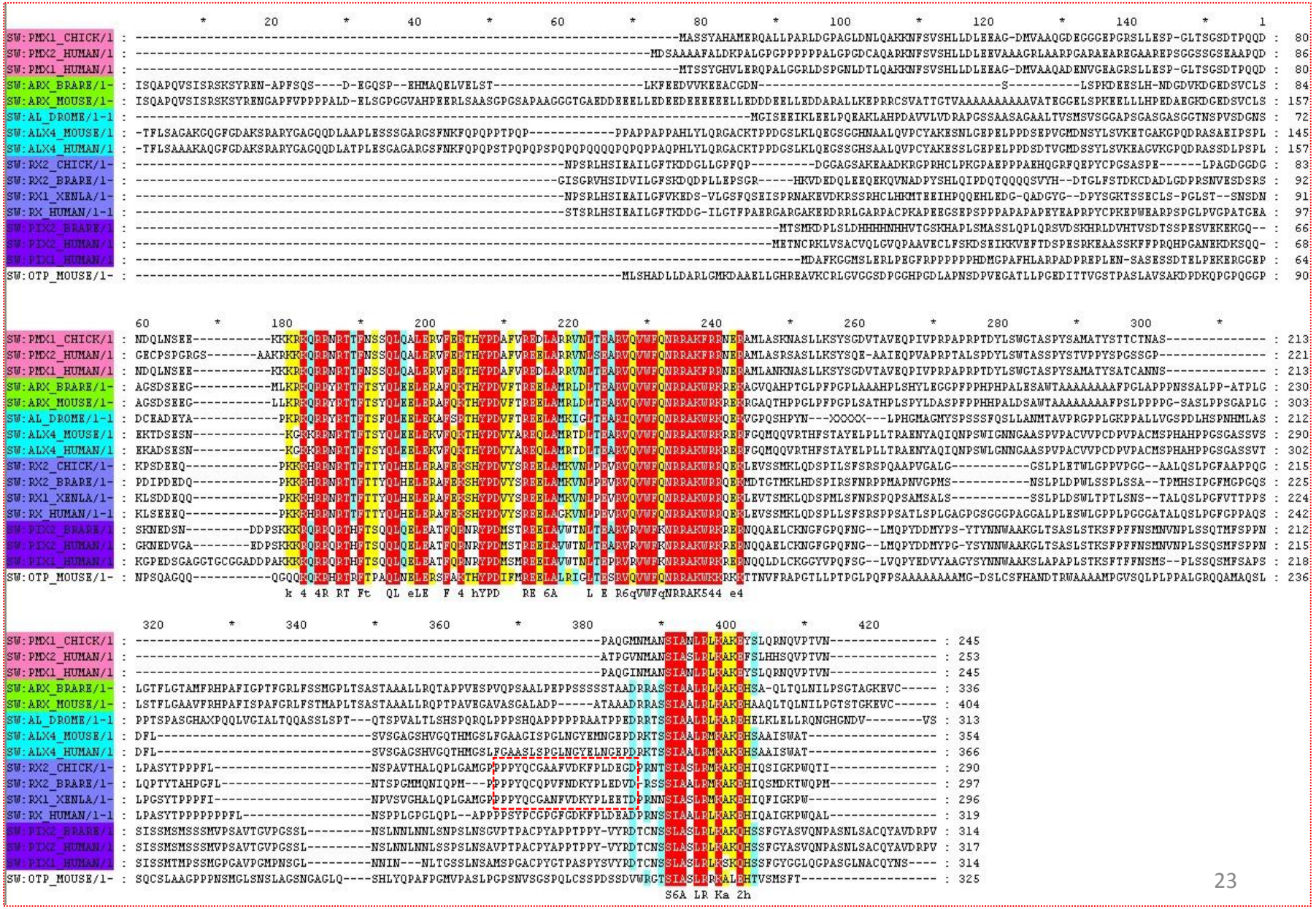
Терминологическая проблема.

ДОМЕН – набор фрагментов последовательностей и их выравнивание. Имеет короткое название. Например, RdRP_1

ДОМЕН белка – фрагмент последовательности, входящий в определенный домен, например, в RdRP_1

ДОМЕННАЯ АРХИТЕКТУРА – последовательность доменов в белке

ДВА ДОМЕНА гомеобелков: гомеодомен и OAR домен



Домены принято изображать так

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

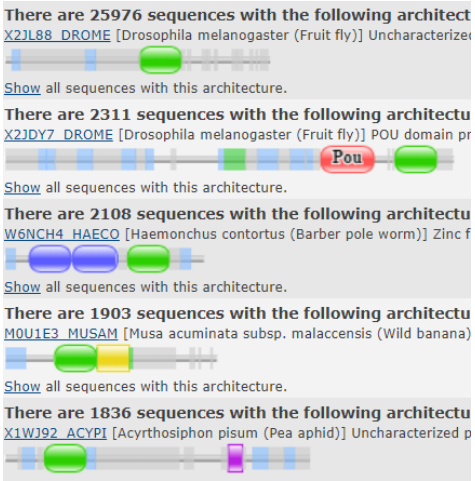
There are 1836 sequences with the following architecture:
Homeodomain, OAR

Гомеодомен является ДОМЕНОМ

Доказательство

- Выравнивание, со сходством, свидетельствующим о гомологичности последовательностей

- Представленность домена в негомологичных белках (обычно, но не обязательно)

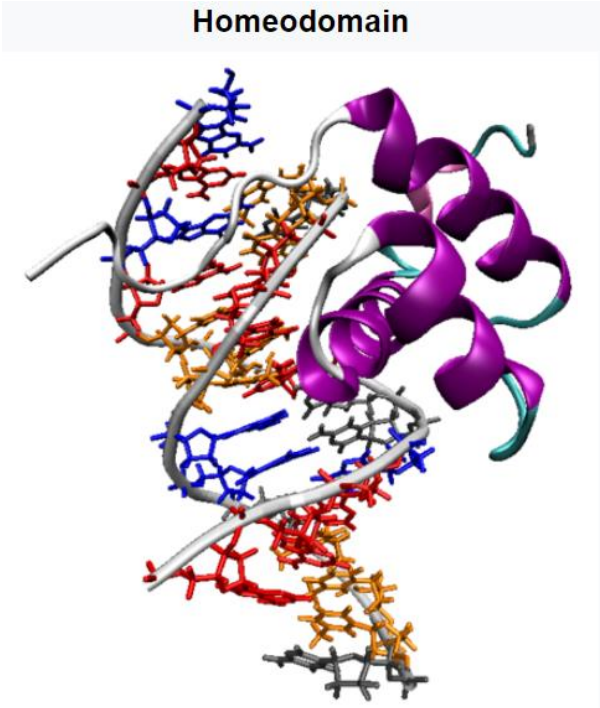


Эволюционные домены

- Имеют определенную функцию (не всегда известна)
  DUF – Domain of Unknown Function
- Часто совпадают со структурными доменами
  (но не всегда)

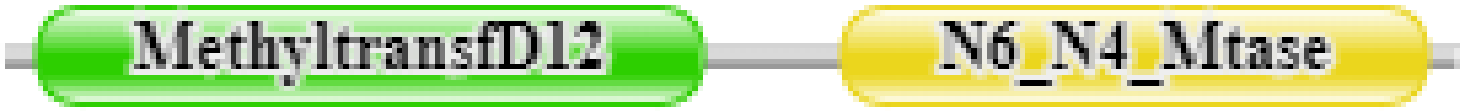
Гомеодомен – ДНК связывающий домен

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

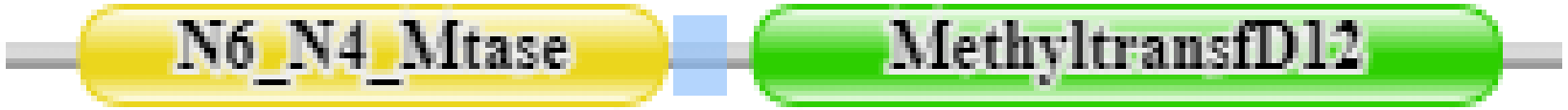


Как выровнять эти две последовательности?

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

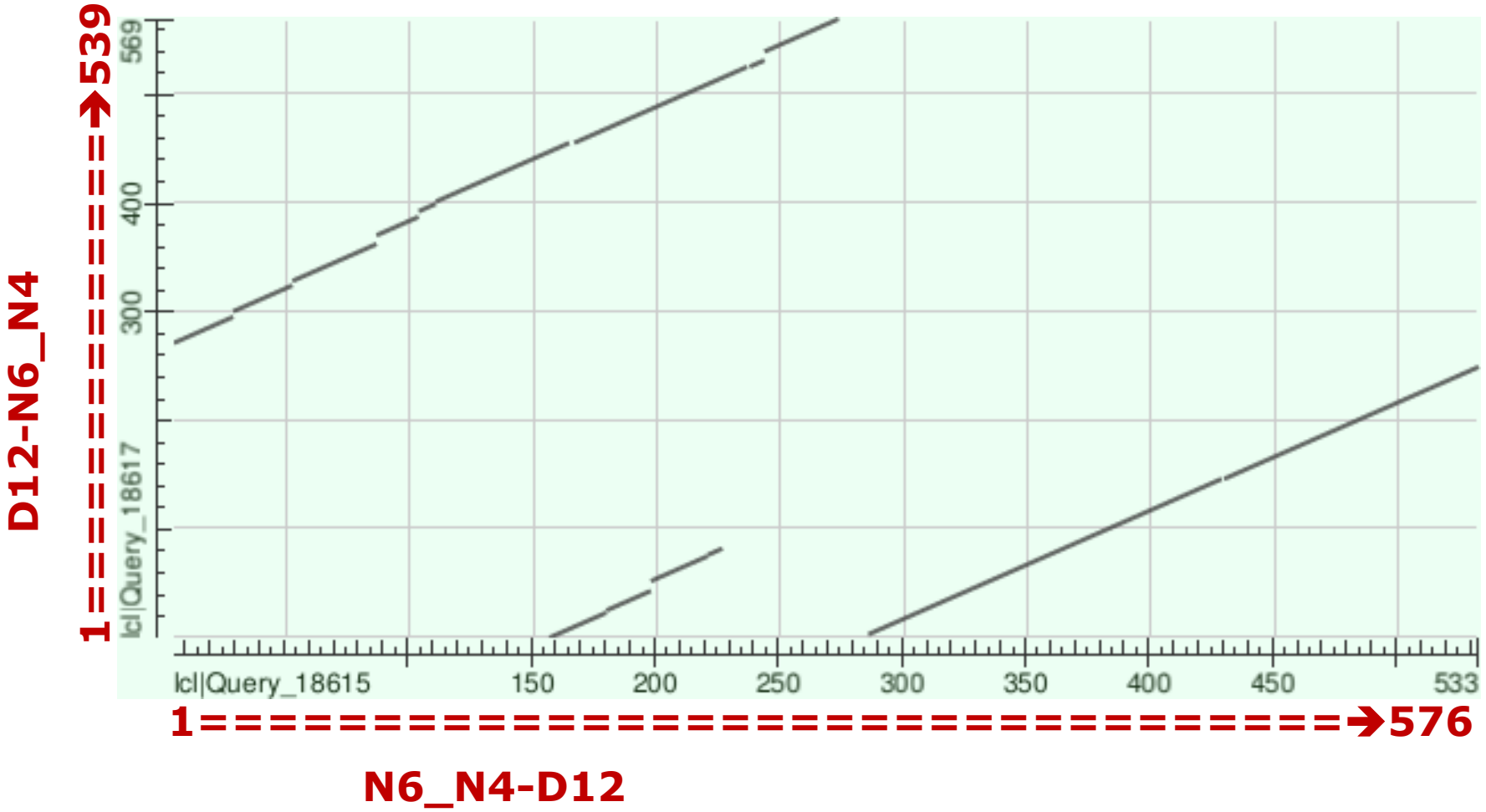


There are 5 sequences with the following architecture:
N6_N4_Mtase, MethyltransfD12
A0A1I7GYG0_9CLOT [**N6_N4-D12**]



Как такое может возникнуть?

Лучшее парное выравнивание: алгоритм множественных локальных выравниваний.

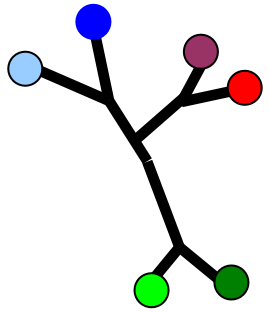


Программа BLASTp. Визуализация Dot Plot

Немного об алгоритмах

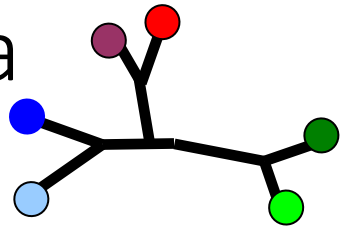
Множественное правильнее парного!

Иерархический алгоритм выравнивания многих последовательностей



- Основная идея: *выравнивание двух выравниваний* с помощью динамического программирования
- Этапы алгоритма
  - Построение направляющего дерева
  - Итерация выравнивания выравниваний
  - “Рафинирование” (refinement) выравнивания
- Результат – ГЛОБАЛЬНОЕ множественное выравнивание

Построение направляющего дерева



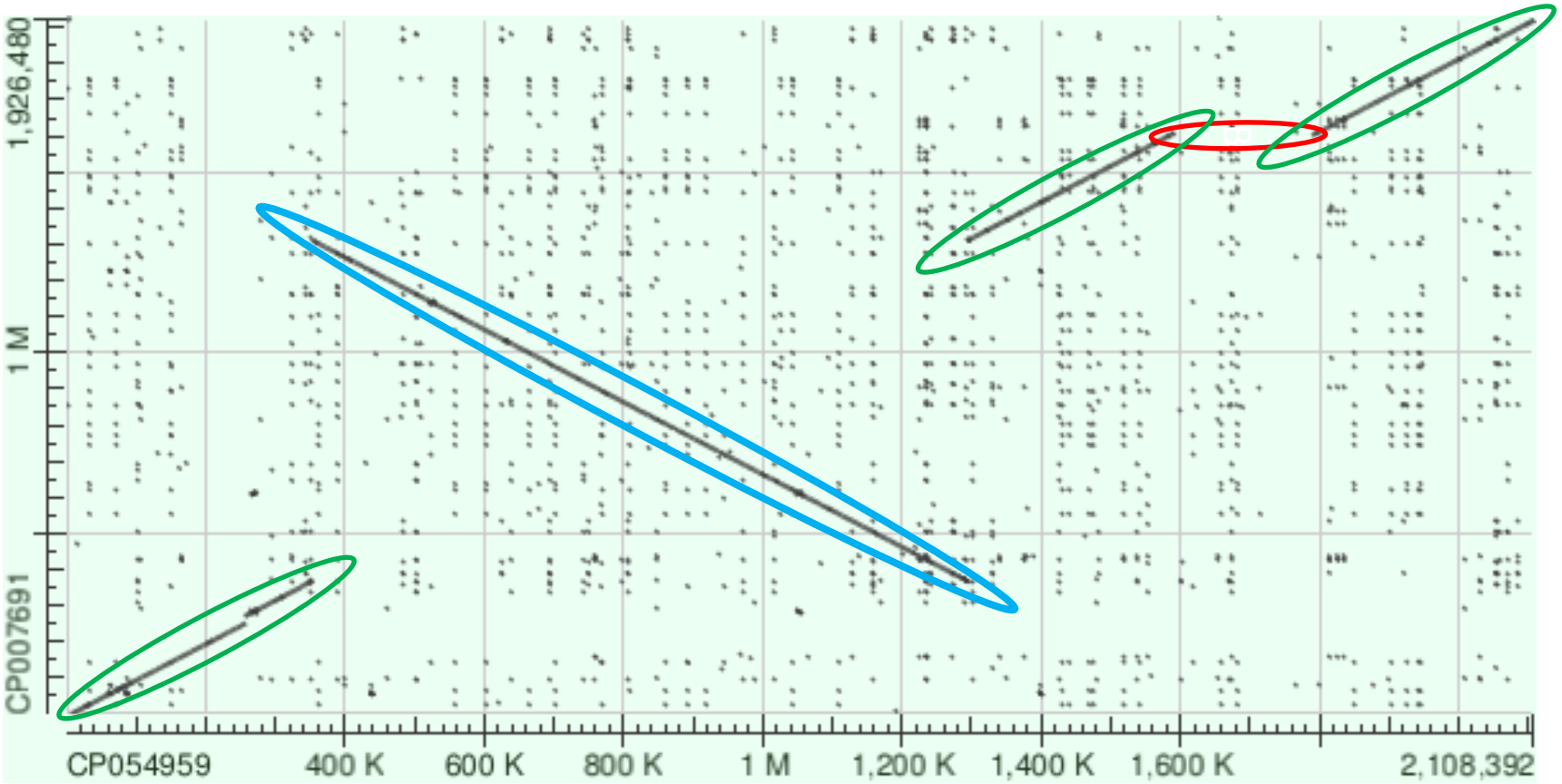
- Для ВСЕХ ПАР последовательностей строится парное выравнивание.
- Вес парного выравнивания пересчитывается в расстояние между последовательностями:
  - чем больше вес, тем меньше расстояние;
  - расстояние между совпадающими последовательностями равно 0.
- Получается матрица расстояний между послед-ми
- Есть алгоритмы, превращающие матрицу попарных расстояний в дерево.
  - Расстояния между листьями по дереву отражают сходство последовательностей

Эволюция геномов бактерий

Крупные - единовременные изменения в геномах:

- Делеция большого участка (многие сотни, тысячи и миллионы пар нуклеотидов)
- Дупликация большого участка
- Горизонтальный перенос, т.е. вставка большого участка из чужого генома
- Инверсия большого участка
- Транслокация большого участка

Карта локального сходства двух геномов родственных штаммов бактерий

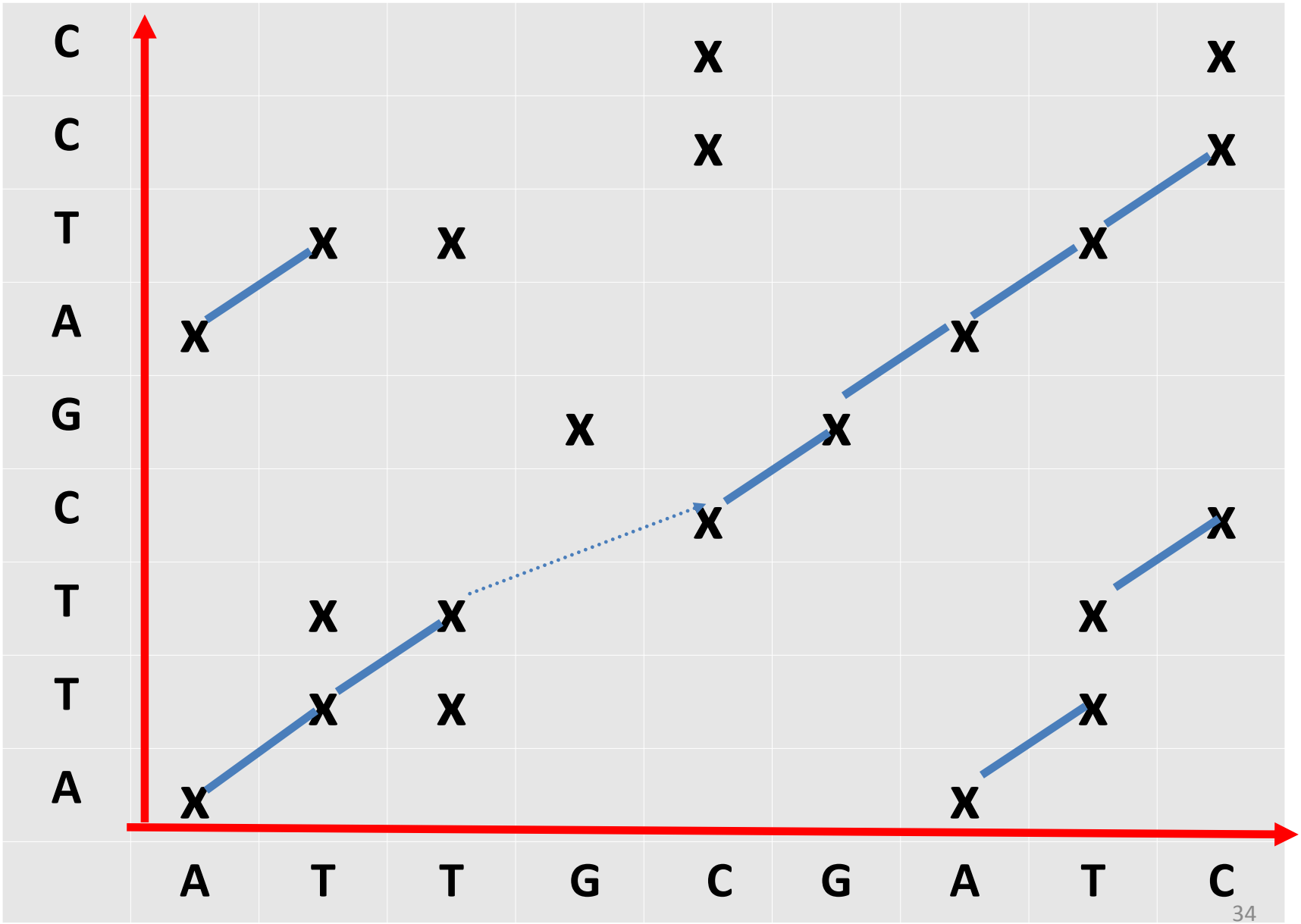


Сходство прямых цепочек

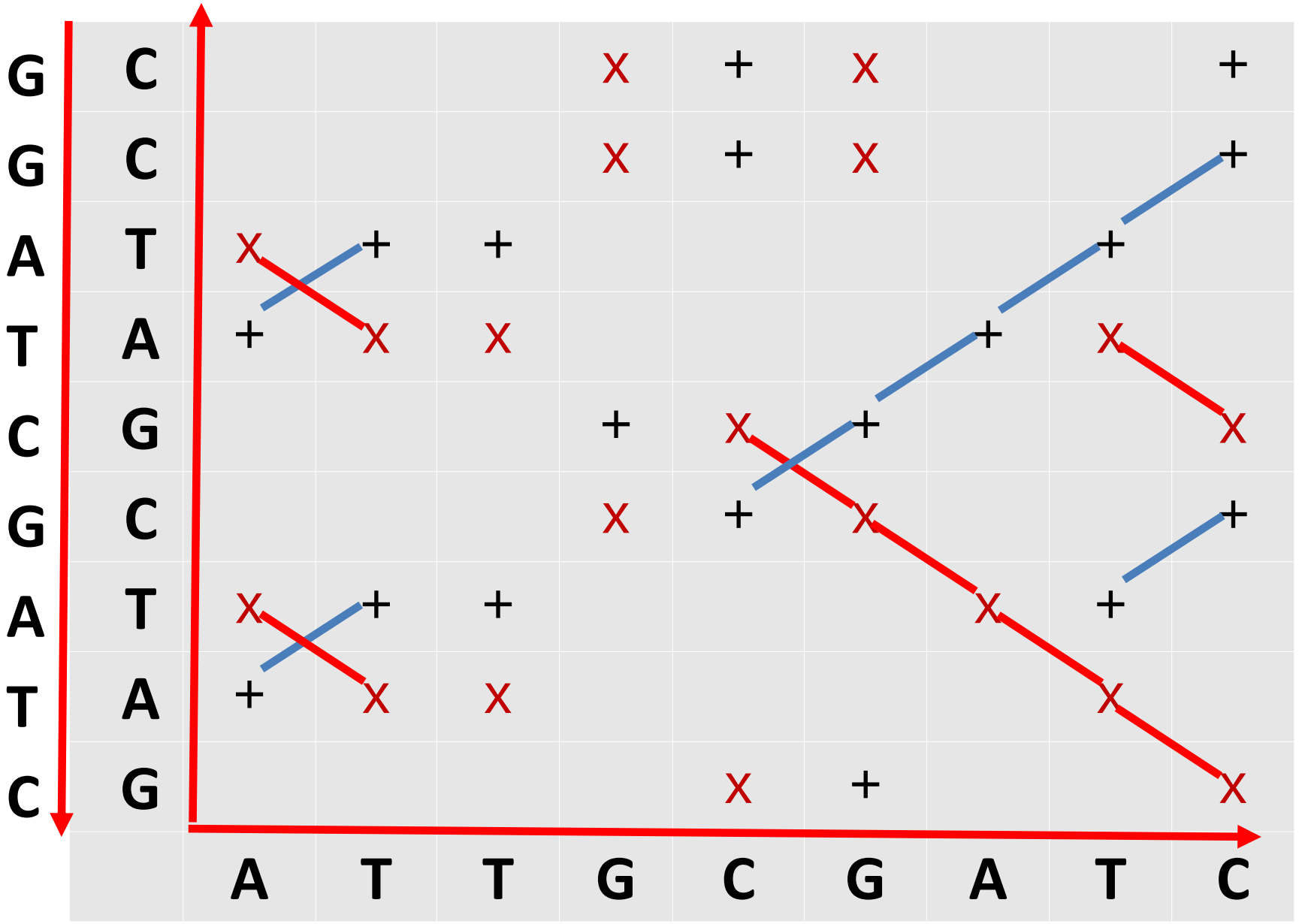
Инверсия - сходство прямой цепочки с комплементарной второго генома

Делеция / вставка

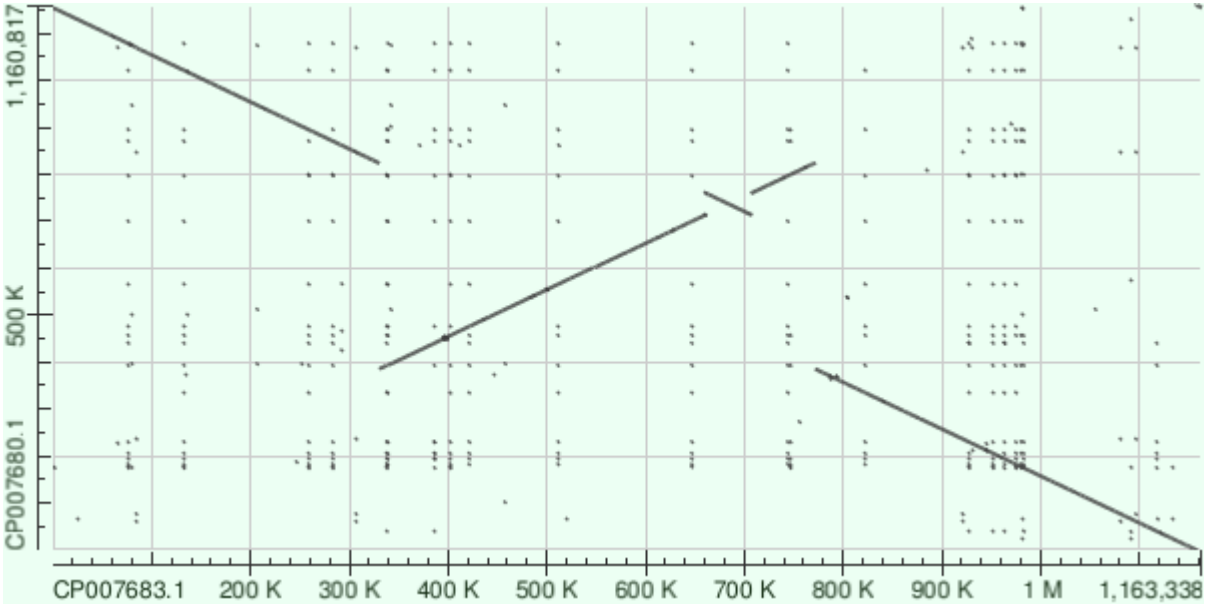
Объяснение Карты локального сходства



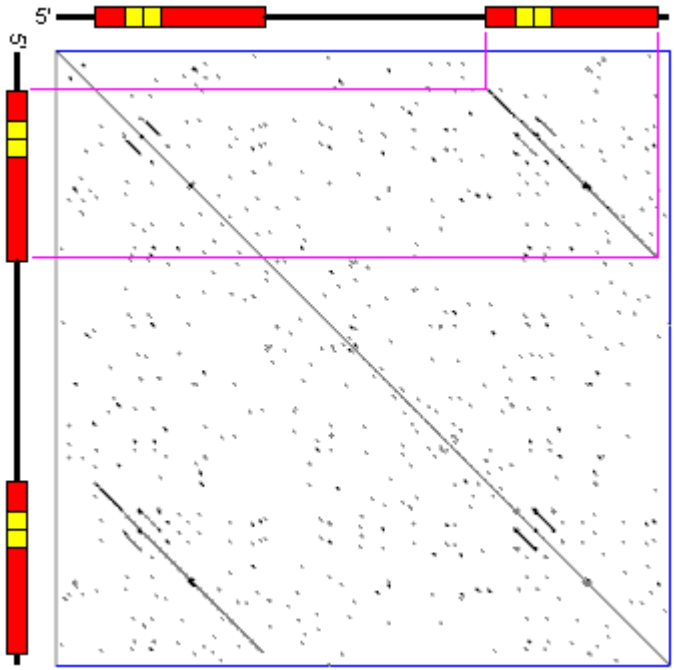
Карта сходства с учетом комплементарной цепочки



Что видим на этой карте?



Дупликация



Пример выполнения задания Семейство доменов ….

КОНЕЦ ПРЕЗЕНТАЦИИ

Какие выравнивания тех же последовательностей совпадают?

```
     1 2 3 4 5 6 7 8 9 10 11 12 13
Seq1 MKFR-SSHYA-S
Seq2 MKYRRS-HYA-S
Seq3 MEFRRRSHYA-R
```

```
     1 2 3 4 5 6 7 8 9 10 11 12
Seq1 MKF-RSSHYAS
Seq2 MKYRRS-HYAS
Seq3 MEFRRRSHYAR
```

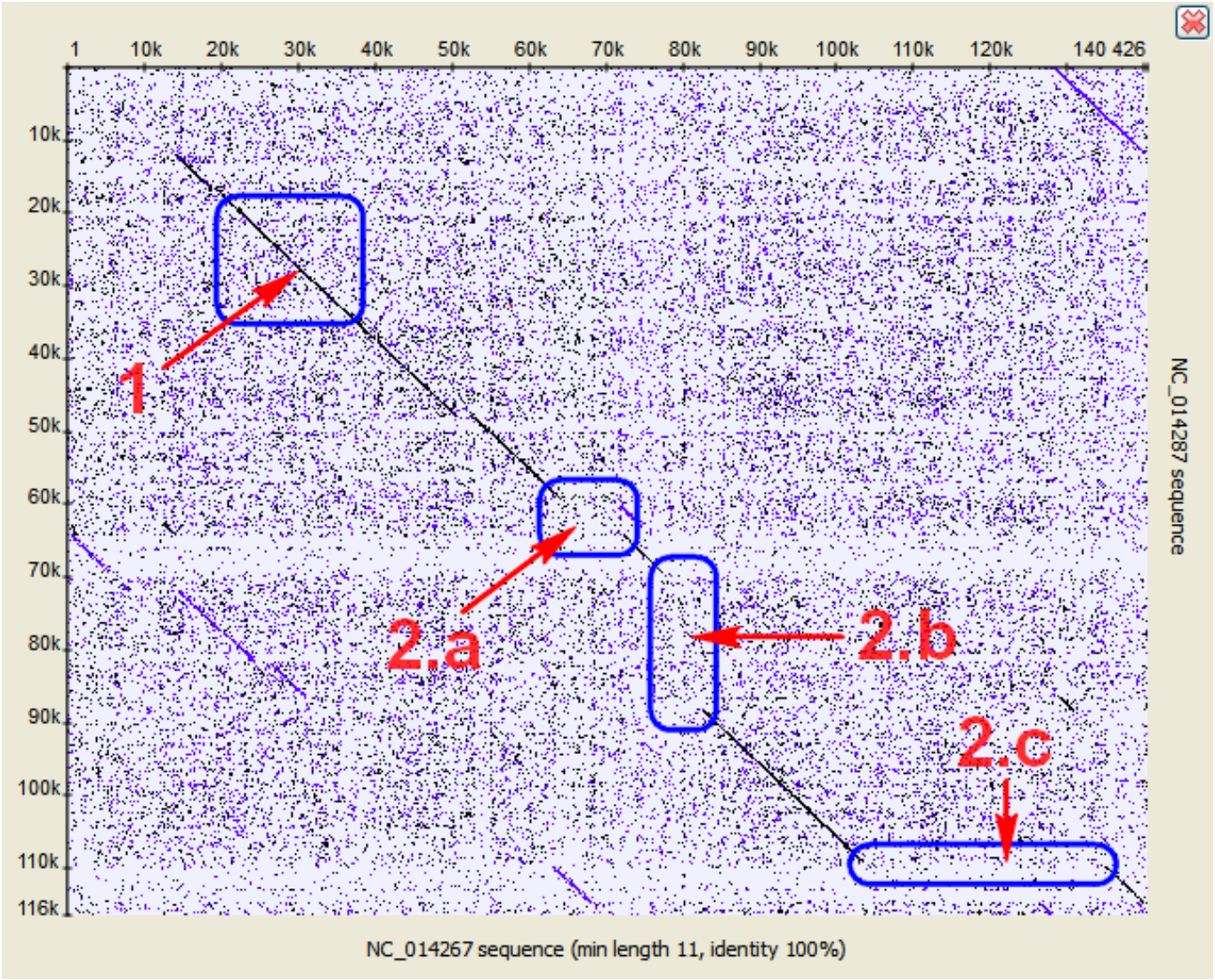
```
     1 2 3 4 5  6 7  8  9 10111213
Seq1 MKFR-SSHYA-S
Seq2 MKYR-RSHYA-S
Seq3 MEFRRRSHYA-R
```

```
     1  2  3  4 5  6 7  8  9 101112
Seq1 -MKFR-SSHYAS
Seq2 -MKYR-RSHYAS
Seq3 -MEFRRRSHYAR
```

Колонка i выравнивания X совпадает с колонкой j выравнивания Y если в них – те же самые остатки; те же самые значит – с теми же номерами, а не с теми же буквами!

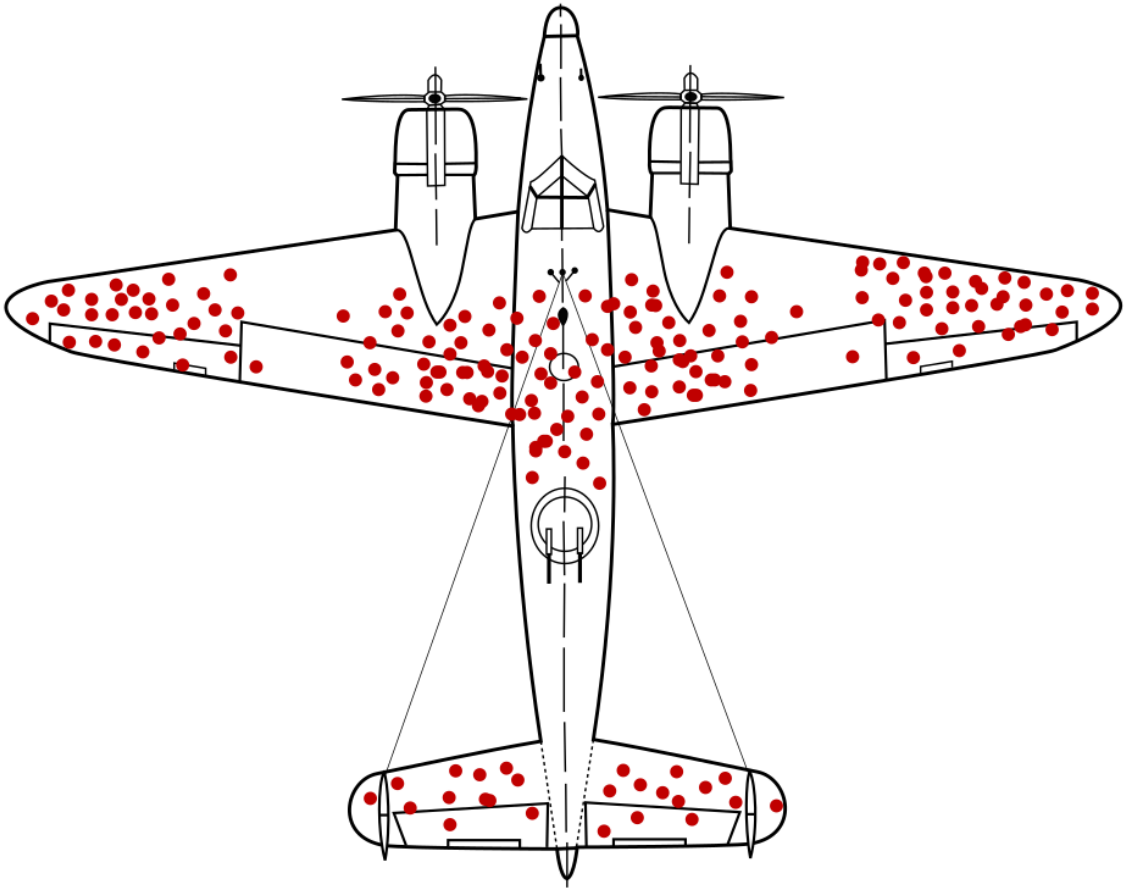
Interpreting Dotplot: Identifying Matches, Mutations, Invertions, etc

Создатель Yuliya Algaer, 2014



Пробоины на вернувшихся американских самолётах во время второй мировой войны.

Какие части укреплять броней?



Слайд заимствовал из презентации МГ

Эволюция белков

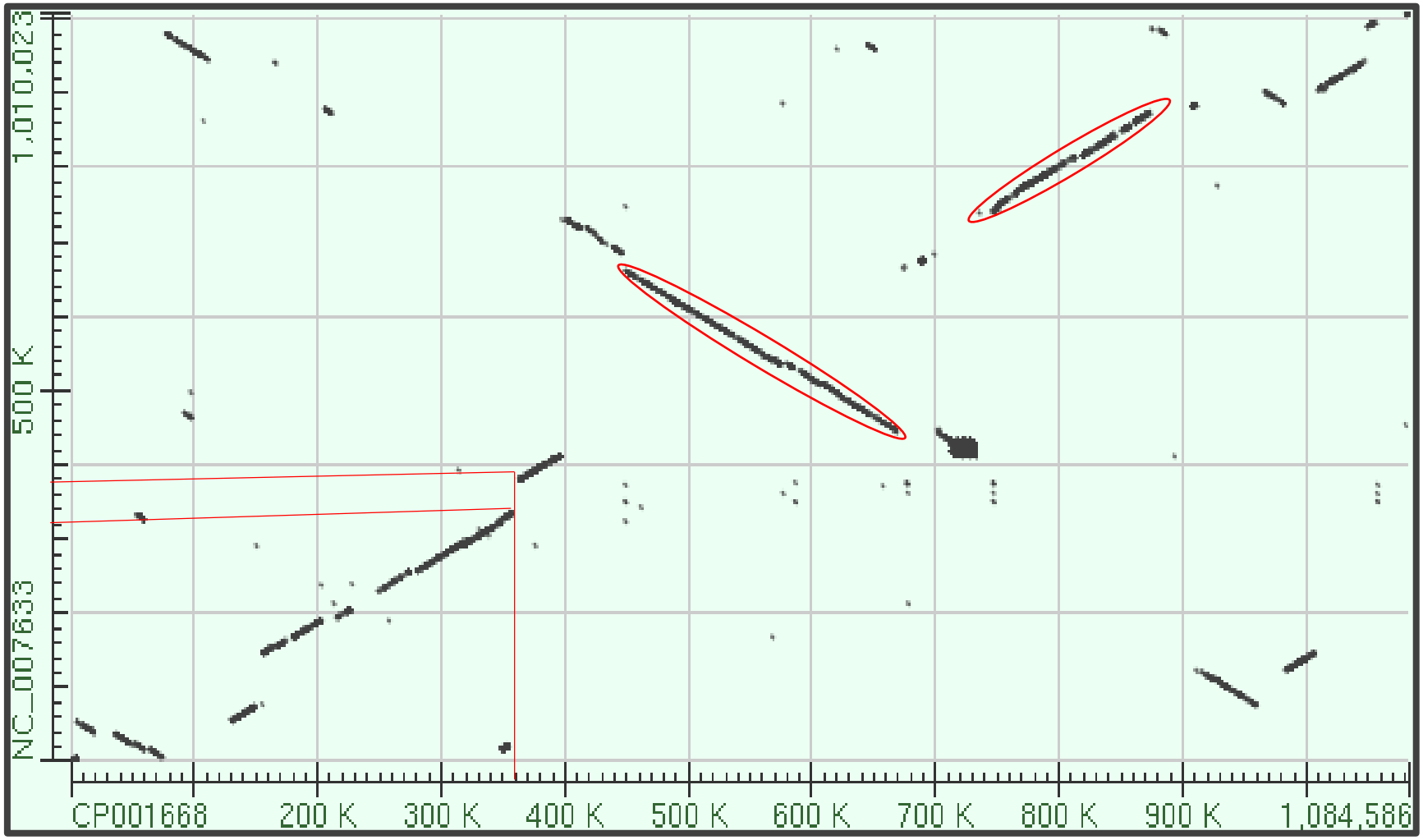
Локальная - небольшие изменения в гене
(Замены а.к. Делеции Вставки)

Большие изменения:

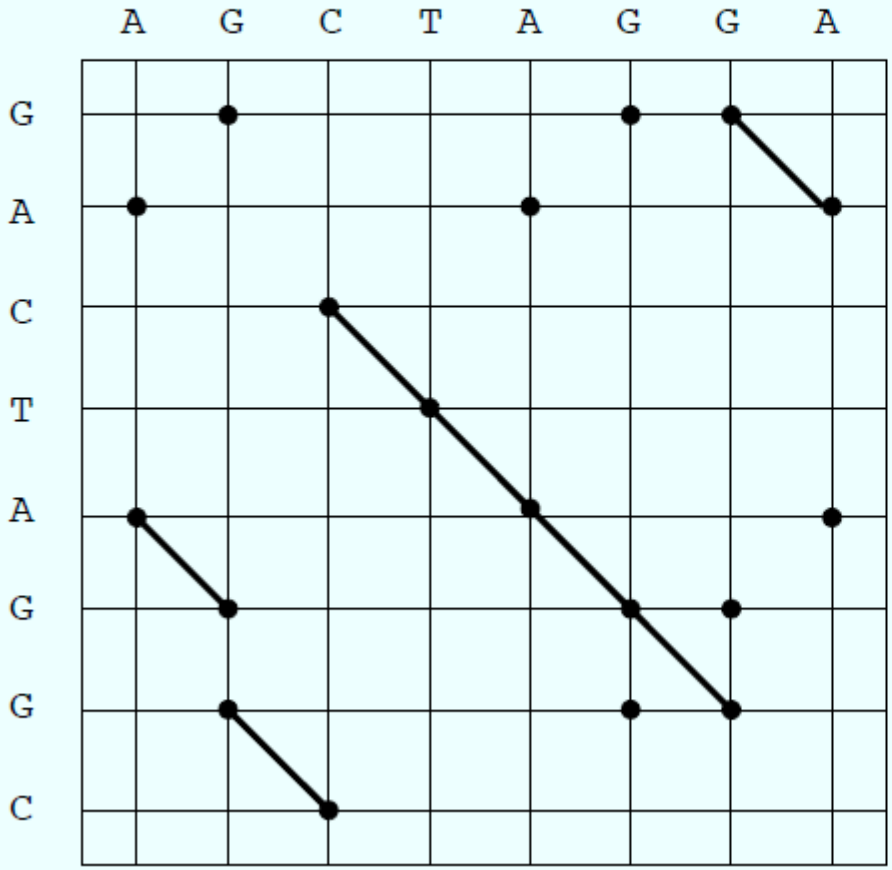
1) Накопленные небольшие изменения
2) Небольшие изменения гена ведущие к большим изменениям белка
    1) Мутация стоп кодона => удлинение последовательности белка
    2) Мутация кодона на стоп кодон
        1) гибель белка = псевдогенизация или
        2) Укорочение последовательности белка
        3)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

2. Парное выравнивание геномов

Карта локального сходства геномов M.capricolum и M.mycoides



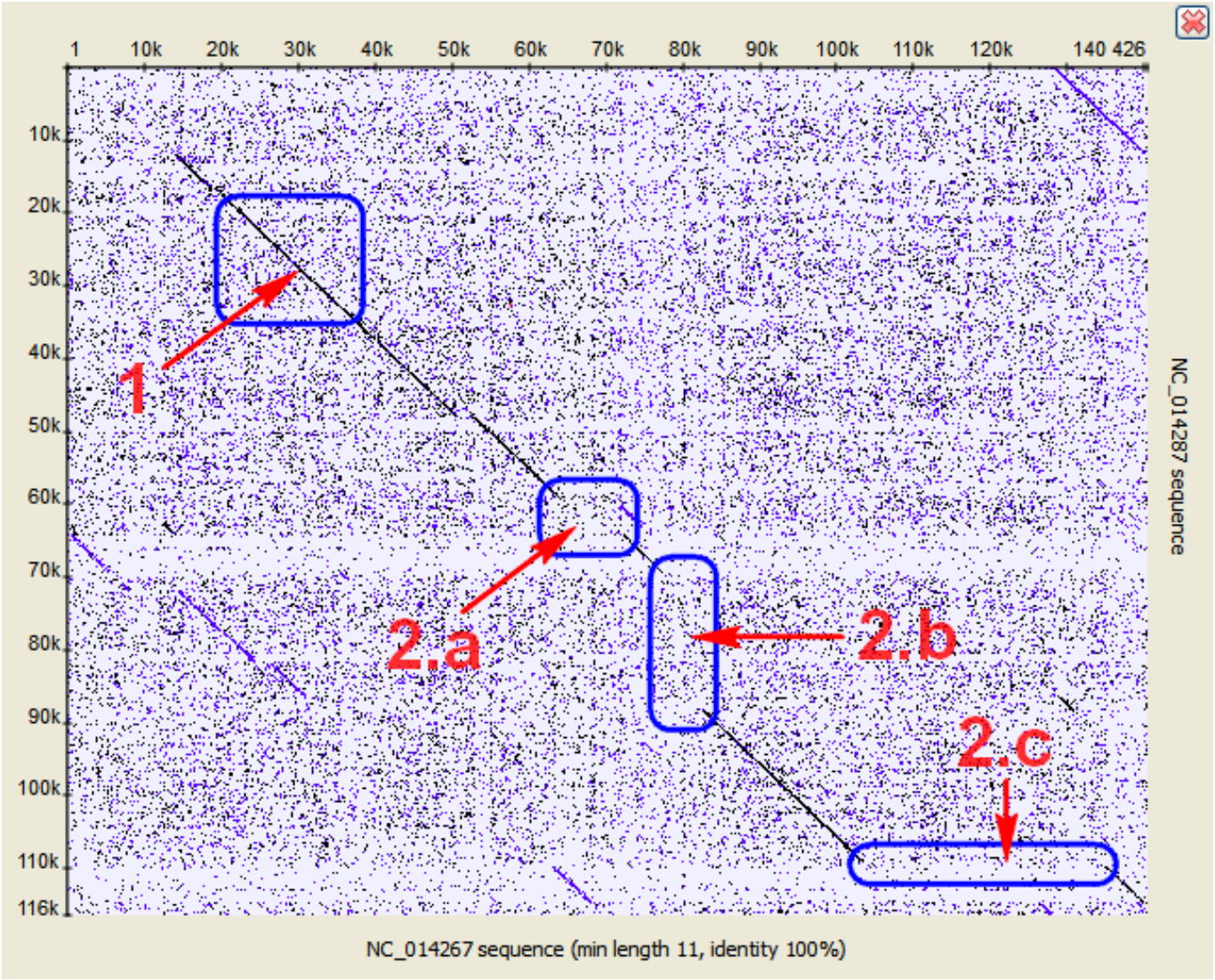
- established in 1970 by A.J. Gibbs and G.A.McIntyre
- method for comparing two amino acid or nucleotide sequences



- each sequence builds one axis of the grid
- one puts a dot, at the intersection of same letters appearing in both sequences
- scan the graph for a series of dots
  - reveals similarity
  - or a string of same characters
- longer sequences can also be compared on a single page, by using smaller dots

Interpreting Dotplot: Identifying Matches, Mutations, Invertions, etc

Создатель Yuliya Algaer, 2014



ВЫРАВНИВАНИЕ