

Гомология

и

выравнивание

Множественное выравнивание последовательностей
гомологичных белков

План

1. Задание
2. Гомология (3 слайда)
 - а. Термин
 - б. Эволюция геномов бактерий
 - с. Гомологию белков и фрагментов геномов определяют по сходству
3. Эволюционное выравнивание (2)
 1. Отражение эволюции с помощью выравнивания
 2. Смысл выравнивания 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
 - а. Эволюция белков
 - б. Эволюционные события
 - с. 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

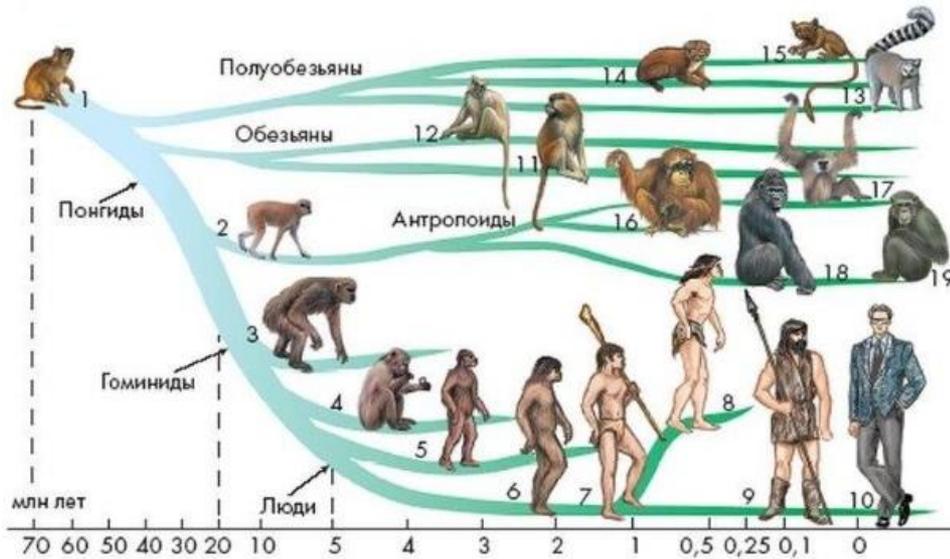
Придумайте эволюционный сценарий наблюдаемого

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 словарь:

* Последний общий предок (LCA)

* Гомология

Последний общий предок ныне живущих обезьян.

Гомоло́гия ([др.-греч. ὁμοιος](#) «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	TATGCGAATGCCCTGAA	
сын	TATG A GAATGCCCTGAA	замена
внук	TATG C GAATG C TCTGAA	замена
правнук	TATG C GAAT C G C TCTGAA	вставка 1 п.н.
праправнуку	TATG A GA A A C G C TCTGAA	замена
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	

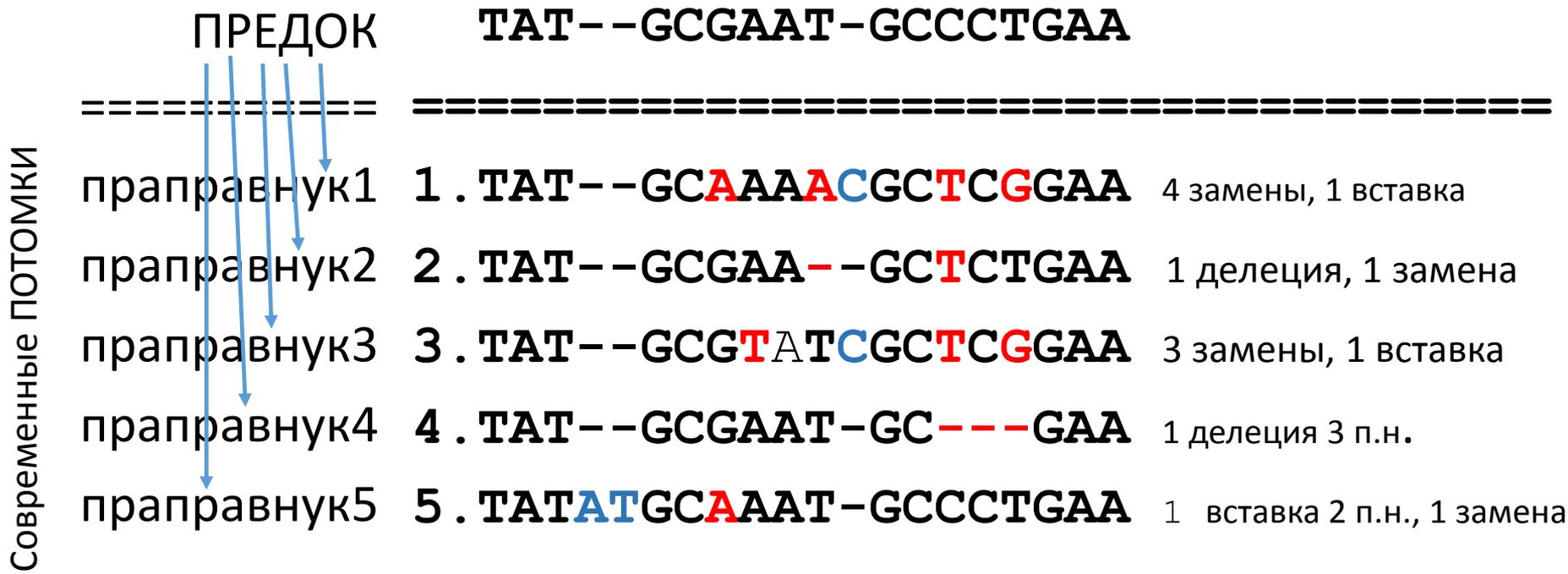
Нукл-ы потомка с номерами как у предка являются гомологами нукл-в предка

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 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 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

Обсуждаем смысл колонок и гэпов

```
          10          20          30          40          50          60
EJL77459.1  GVDLVF GGPPCQGF SQIGMRR- LDDER- NE LYQQYTR I VAKLKP RVFLMENV PNLALMNK GH
RXK67093.1  DLDVVF GGPPCQGY SQIGTRR- LDDER- NE LYLQYAR I VEKQRP RMFLMENV PNMVLLNK GH
OJY44288.1  NVDLVF GGPPCQGY SQIGTRD- LHDPR- NRL FEEFAR VVATLKP KFLMENV PNL LLLNK GH
TRU90449.1  NPEMIV GSPPCQDF SSAGKR NEGLGR- - ANLTLTFAE I VTRVSP QWFMENV D- - -RIE KSK
OXI46696.1  GTDLVF GGPPCQGF SQIGMRR- LDDER- NE LYKQYTR VVSTLR PRVFLMENV PNLALMNK GH
AVZ30243.1  EIDVVF GGPPCQGF SLIGKRS- FEDPR- NSLVFH YIRLVLE LSPKFFV IENVKGM TAGNHQA
AFZ12381.1  DIIGFI GGAPCPDF SVGGKNR GSEGDK- GKLSASY IELICQQ KPDFFLF ENVKGL YKTKKHR
HCQ21462.1  HIIGFI GGPPCPDF SVGGKNR GHLGDN- GKLSASY IELICQN LPDFFLF ENVKGL WRTTKHR
EDN77159.1  SLIGFI GGPPCPDF SIAGKNK GKDGDN- GKLSLSY TNLIIEM KPDFFLF ENVKGL WRTARHR
SOD91684.1  EVSLVV GGAPCQPF SNIGKKL GKNDERN GDLFLEF VRMVKG IQPEAF I FENVVG ITQNK HSD
QCS48280.1  NVVGF I GGPPCPDF SI GGKNRGR QGDH- GKLSSESY IDLI IQHQPDF FIFENVK GLYR TKKHR
SMB95934.1  GLFGII GGPPCPDF SVGGKNR GENGEQ- GRLSKV FVDKI LDLQPV FFLYENVP GLIRTA KHR
RUO38876.1  SPVGF I GGPPCPDF SVGGKNR GENGEN- GRLTRTY VDGII KYAPDF FIFENVK GLWRTK RHR
OIP70538.1  TIDLIC GGPPCQGF STIGTND- KKDHR- NLF FEF LRMVET FKNFI ILENVT GLLAK KNES
AFY60915.1  NLVGFV GGPPCPDF SI GGKNRGQ YGDN- GKLTKVY VDII IENQPDF FVFENVK GLWRTR SHR
CUR30340.1  DLIGFI AGPPCPDF SVGGKNR GKNGDQ- GKLTACY VELICQQ RPDF FVFENVK GLWSTK KHR
TAK03971.1  QAALVV GGAPCQPF SNLGS KRGTAD SR- GTLFQDF IRI VKGVRP KGFIFENVE GLTQDK HKG
AEE51071.1  KVALVV GGAPCQPF SNIGKKE GENDAKNG DLFLEF VRMVKG IQPEAF I FENVAGI IQSK HSK
RTR31666.1  RLVGFV GGPPCPDF SVGGKNR GSEGEN- GKLTRTY IDLIV KDNPDYF I FENVKGL WRTTR HR
PTU64472.1  NIDLVF GGPPCQGF SQIGTRR- LDDER- NE LYKQYTR I VKTLKP RVFLMENV PNLAMMNK GH
```

DNA (cytosine-5-)-methyltransferase

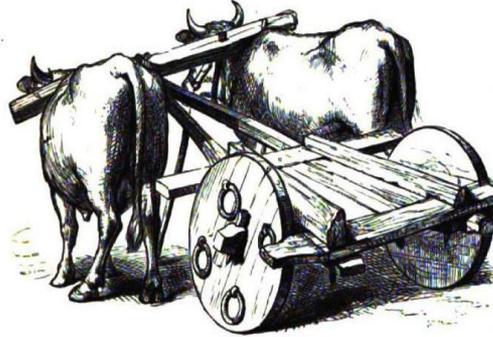
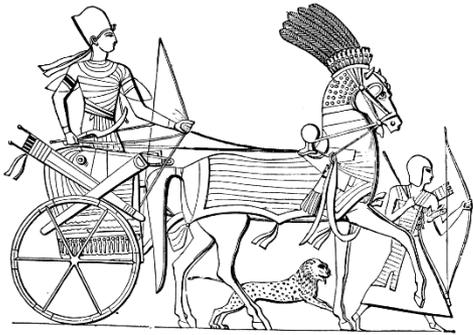
Не всегда так хорошо как на предыдущем слайде

Продолжение того же выравнивания

```
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-VPQSRQRVFFIIGLKSDRPLNQQ-----ILTP-----PSKVI ESEYTSLEEAISDLPVIE-----AGEGGEVQDYPVAE
CG-VPQLRKRTFVIGHRHGS IAD-----LANVLQQR LAKQSL-----TVRDYFG-
CG-VPQSRTRFSLIGKLNSEHNF-----LIPTLSRKLSDKPM-----TVRDYLG-
YG-VPQRRRRI 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 FYLVAFLNQN-----IHFEFPPK-----PMISKDIGEVLES
FG-LPQRRERIVIVGFHPDLG-----INDFSFPKGN-----PDNKVPINAIL E H
YG-IPQKRERIYMICFRNDLN-----IQNFQFPKP-----FELNTFVKDLLLPD
YG-NAQRRRRVFIIFGYKQDLNYSKAME-----ESPLDKI IYHNGLFAEAFP IEDYANKNR-----VNRTHITHDIVDISDNI
YG-TPQRRKRAIIRLNKKGTIWN-----LPLKQNI VSVEQ-----AIGNLPSIESGK
GG-TPQVRERVFITATLVPERMRDERIPRTETGEIDAEAIGPKPVATMNDRFP I KKGTEL FHPGDRKSGWNLLTSGI IREGDPEF
YG-VAQNRDRVFIIGIQQKLGVPD-----FSFPEYSESEQRLYDILDNLQTPSII-----PESLPIQRNLFGEF
FG-VAQNRDRVFI VGIQQKLDLNG-----FSFPEYAESDQRLYHILDNLEAPETK-----LESIP IQRNLFGEF
YD-VAQKRERIVIIGIREDLVK-----EQYPPFRFPLAQ-----VYKPVLKDV LKDY
YG-VSQLRPRVLFVALKNEYTN-----FFKWPEPNSEQPK-----TVGELLFDLMSE
```

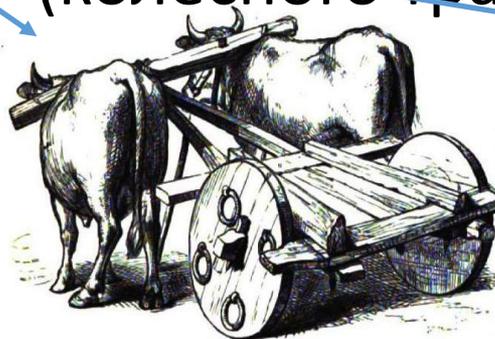
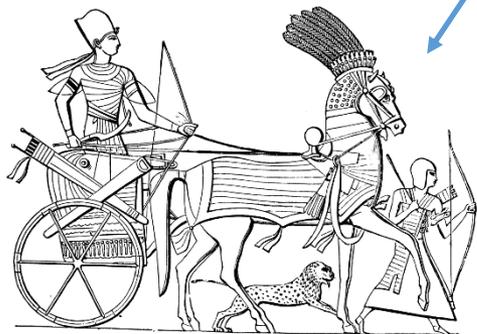
Консервативное значит важное

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



LUCA

(колесного транспорта)

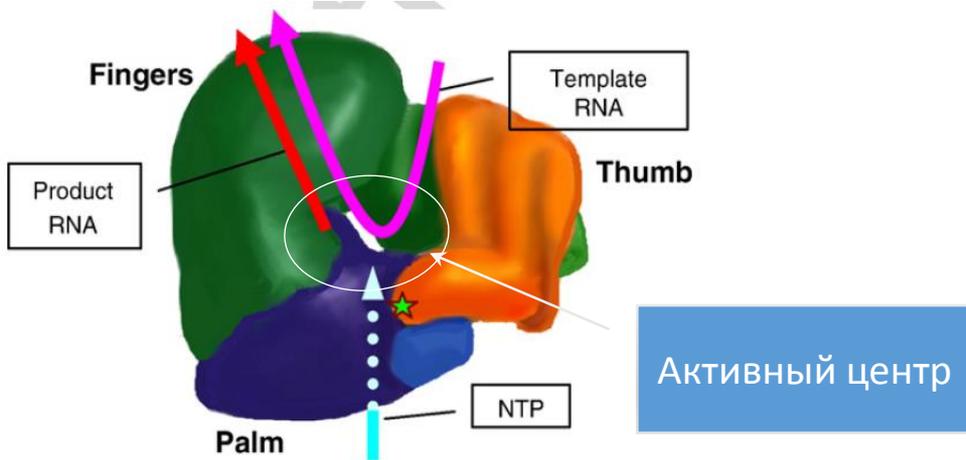


РНК зависимаая РНК полимераза (RdRP), консервативные участки

```

*      320      *      340      *      360      *      380      *      400      *      420      *      440      *      460
FKTMIRFGDVGLDLDFSAFDASLSPPFMIREA..GRIMSELS...GTPSHFGTALINTIIYSKHLIYNCCY.....HVCGSMPSGSECTALNSTINNVNLYYVFSKIFGKSPVFF.....CQALKILC.YGDDVIVFVSRD
EVAMQG.FERVYDVVDSNEDSTHVSAMFRLL..A...EEFF.TPENGFDPLTREYLESIAISTHAFEKRF.....LTGGLPSGCAATSMINTIMNNTIIRAGLYLTYNFEFDD.....VKVLS.YGDDLIVATNYQL
ETHFAQ.YKNVWVDVLYSADANHCSDAMNIMFEEVFERTEFG.....FHPNAEWILKTLVNTTEHAYENKRI.....VVEGCMPSGCSATSIINTILNNTIYVLYALRRHYEGVELDT.....YTMIS.YGDDIVVASYDYD
.....WSLCVATIVSDHDTFWPGWLRDLICDELINMGYA.PWVVKLFETSLKLPVYVGAFAPEQGHLLGDPSNPDLVGLSSGQGATDLMGTLIMSTIYLVMLQDHTAPHLNSRIKDMPSACRFLDSYWQGHEETROIS.KSDDAILGWTKGR
LRLRLE.NWVYCADAGSQEDSSSLTPYLINAV..LTLRSTYMEDWDVGLQMLRNLYTEIVYTPISTPDGTIV.....KKFRGNNSGQPSIVDNLSLMVVIAMHYALIKECVFEFEID.....STCVFFV.NGDDLIAVNPEK
HDKLNRPGLWLGSGDGRDSSIDPFFFDVV..KTKRKHEL..PSEHHRAIDLIIYDEILNTTICLANGMVI.....KKNVGTQR.QPSTVDNTLVMITAFLYAYIHKTDGRELAL.....LNERFIFVC.NGDDNKFAISPQF
AISLASFSFYGFNCFANEDGMFHPSSFSMV..SEIANIFY...GNFLSTERDNLTRMLTNRFSLMKGAIL.....RVPGGSPSGFFMTVFNSEINLFYLSAWIMLARFNGRQDISH.....PCNFPKYVRACV.YGDDNIVAIMEV
AARMKEKGNVDVLCODYSSEFDGLLSKQVMDVI..ASVINELC.GGEDQLKNARRNLLMACCSRIAICKNTVW.....RVECGIPSGFFMTVFNSEINLFYLSAWIMLARFNGRQDISH.....QSFDKLIGLVT.YGDDNLSVNAV
YAEHAK.YKNHFDADYIADSTQNRQIMTES..FSIMSRLT...ASPELAEVVAQDLAPSEMVDG DYVI.....RVKEGLPSGFFPCTSQVNSINHWITLICALSEATGLSPDVV.....QMSYFYSFYGDDIVSTDIDF
NNLTSKASDFLCLDYSKFDSTMSPCVVRIA..IDLADCC...EQTELTKSVVLTILKSHFMTILAMIV.....QTKRGLPSGMPFTSVINSHCHWLLWSAAVYKSCAEIGLHCS.....NLYEDAPFYT.YGDDGVYAMTEMM
IQRIKS.AAKVYAVDYSKWDSTQSPRVSAA..IDLRYFS...DRSPIVDSAANTLKSPPIAIFNGVAV.....KVSGLPSGMPFTSVINSHCHWLLWSAAVYKSCAEIGLHCS.....NLFSTFLMMT.YGDDGVYMFPMF
TKRLERPKHRYCVLYSKWDSTQPPKVTSSQS..IDILRHFT...DKSPIVDSACATLKSNPIGIFNGVAF.....KVAGGLPSGMPFTSVINSHCHWLLWSAAVYKSCAEIGLHCS.....NIFDSMDLFT.YGDDGVYIVPPLI
D      D      g      sg      T      n3      gDD

```



На каких участках выравнивание правильное – совпадает с эволюционным?

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.

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.

```

      *           100           *           120           *           140           *           160
THEIE_LACLS : AGVSEFIVNDDVELARELNADGHIHGQTDSEVSKVREKVGQEMWLGLSVTKADELKTAQ-SSGADYLGIGPIYPTNSKND : 14
THEIE_MANSM : FQVBFIVNDDVELALSIQADGHIHVGQKDTAVETILRNTRNKPIIIGLSINTLAQALANKDRQDIDYFVGVPPIFPTNSKADH : 15
THEIE_STRA3 : YQVBFIIIDDDIDLVELIDADGHIHGQNDLPVDEARRRLPDKI-IGLSVSTMAEYQKSQ-LSVVDYIIGIGPFNPQSKADA : 14:
THEIE_LISIN : YQVBFIIINDDDVALALEIGADGHIHVGQNDDEEIRQVIASCAGKMKIIGLSVHSVSEAEBAERLGSVDYIIGVGPPIFPTISKADA : 14:
THEIE_ANOFW : YNIBFIVNDDVDLALALQADGVHVGQDEVEAERVDRDRIGDKY-LGVSVHNLNEVKKAL-AACADYVGLGPIFPPTVSKEDA : 14:
THEIE_GEOTN : YGVBFIVNDDVELAIAIDADGVHVGQDDEADARRVREKIGDKI-LGVSAHNVVEEARAAI-EAGADYIIGVGPPIYPTRSKDDA : 14
THEIE_BACSU : AGVBFIVNDDVELALNLKADGHIHGQEDANAEREVRAAIGDMI-LGVSAMTSEVVKQAE-EDGADYVGLGPIYPTETTKDDT : 14
THEIE_BACA2 : AGIBFIIINDDDVELALRLEADGVHIGQDDADAEEETRAAIGDMI-LGVSAMTSEVVKQAE-AAGADYVGMGPVYPTETTKDDT : 14
THEIE_OCEIH : FQIBFIIINDDDVDLAKQLDADGHIHGQDDQPVVVRKQFENKI-IGLSISTNNELNQSP-LDLVDYIIGVGPPIFDTNTKEDA : 14
THEIE_STAAB : YNVBFIVNDDVSLAKEINADGHIHVGQDDAKVKEIAQYFTDKI-IGLSISDLGEYAKSD-LTHVDYIIGVGPPIYPTPSKHDA : 14:
THEIE_STACT : YNVBFIVNDDVALAEEIDADGHIHVGQDDEAVDDFNRRFEGKI-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 : SPIVGMNFIRQIRQLGIDKPCVAIGGITKEESAAILRRLGADGVAVISAIHSHSVNIANTVKTLAQK~~~~~ : 22
THEIE_STRA3 : KPAVGNRTTKAVREINQDIPVVAIGGITSDFVHDIIIESGADGLAVISAIISKANHIVDATRQLRYEVEKALVNRQKRSDVI : 22:
THEIE_LISIN : EPVSGTAILEEIRRAGIKLPIVIGGGITNETNSAEVLTAGADGVSVISAITRSEDCQSVIKQLKNPGSPS~~~~~ : 21
THEIE_ANOFW : KQACGLTMEHIRAEKRVPLVAIGGITQETAQKQVIEAGADGLAVISAIKRAEHIYEQTKRLYEMVMRAKQKQKQDR~~~~~ : 21
THEIE_GEOTN : NEAQQPGILRHLRREQGITIPIVVAIGGITADNTRAVIEAGADGVSVISAIASAPEPKAAAAALATAVREANL---R~~~~~ : 22
THEIE_BACSU : RAVQGVSLIEAVRRQGISIPIVIGGGITIDNAAPVIEAGADGVSMISAIQAEDPESAARKFREEIQTYKTG--R~~~~~ : 22:
THEIE_BACA2 : EAVQGVTLIEEVRQGITIPIVIGGGITADNAAPVIEAGADGVSMISAIQAEDPKAAARKFSEEIRRSKAGLSR~~~~~ : 22
THEIE_OCEIH : KTAVGLEWIIQSLKKQHPSLPIVVAIGGITNTNAQEIIQAGADGVSVISAITETDHIILQAVQRL~~~~~ : 20
THEIE_STAAB : HTPVGPMEIATFKEMNPQLPIVVAIGGITSNVAPIVEAGANGISVISAIKXSENIIEKTVNRFKDFFN~~~~~ : 21:
THEIE_STACT : SEPVGPKMIETLRKEVGDLEIPIVVAIGGISLDNVQEVAKTSADGVSVISAIARSPHVTETVHKFLQYFK~~~~~ : 21:

      80           *           100           *           120           *           140           *
THEIE_LACLS : VSEFIVNDDVELARELNADGHIHGQTDSEVSKVREKVGQEMWLGLSV-TKADELKTAQSSGADYLGIGPIYPTNSKND :
THEIE_MANSM : VEFIVNDDVELALSIQADGHIHVGQKDTAVETILRNTRNKPIIIGLSINTLAQALANKDRQDIDYFVGVPPIFPTNSKAD :
      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--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.

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

Домены принято изображать так

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



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

Эволюционные домены

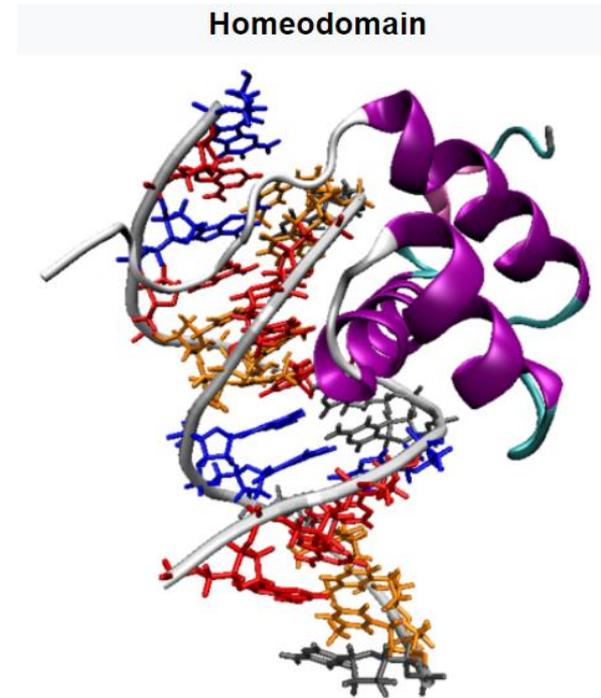
- Имеют определенную функцию (не всегда известна)

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]

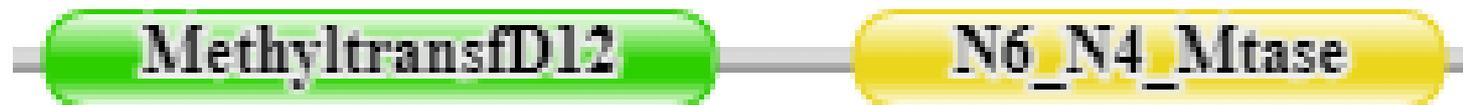


Как выровнять эти две последовательности?

There are 9 sequences with the following architecture:

MethyltransfD12, N6_N4_Mtase

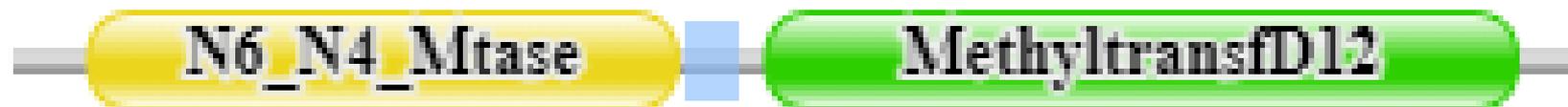
[A0A2Z5QVW5](#) [9MICC](#) [**D12-N6_N4**]



There are 5 sequences with the following architecture:

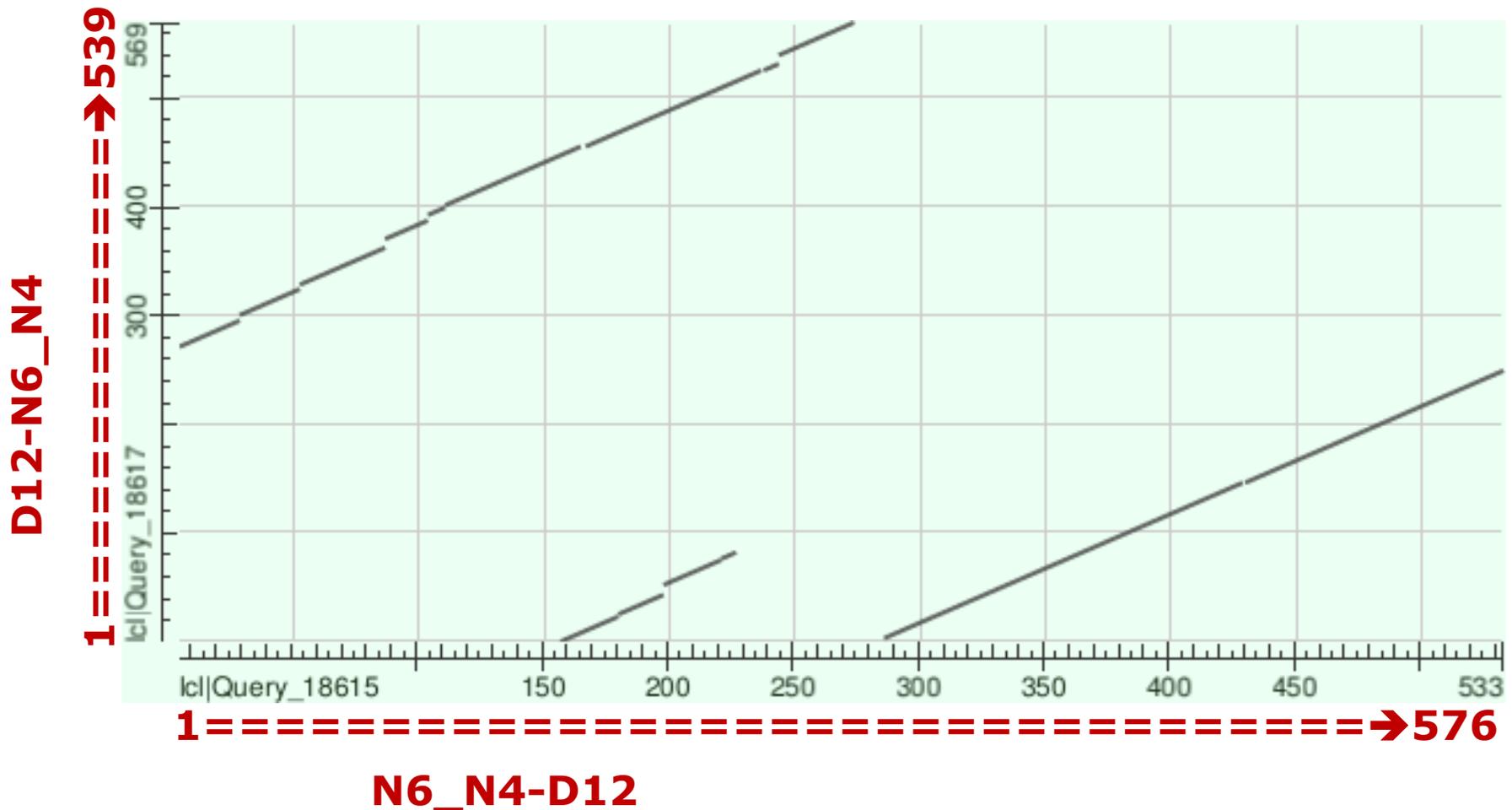
N6_N4_Mtase, MethyltransfD12

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



Как такое может возникнуть?

Лучшее парное выравнивание:
алгоритм множественных локальных
выравниваний.

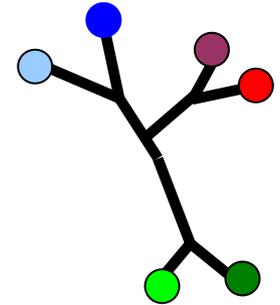


Программа BLASTp. Визуализация Dot Plot

Немного об алгоритмах

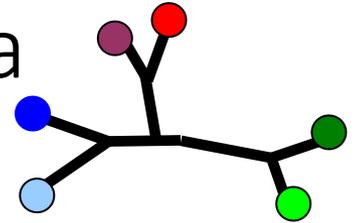
Множественное правильнее парного!

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



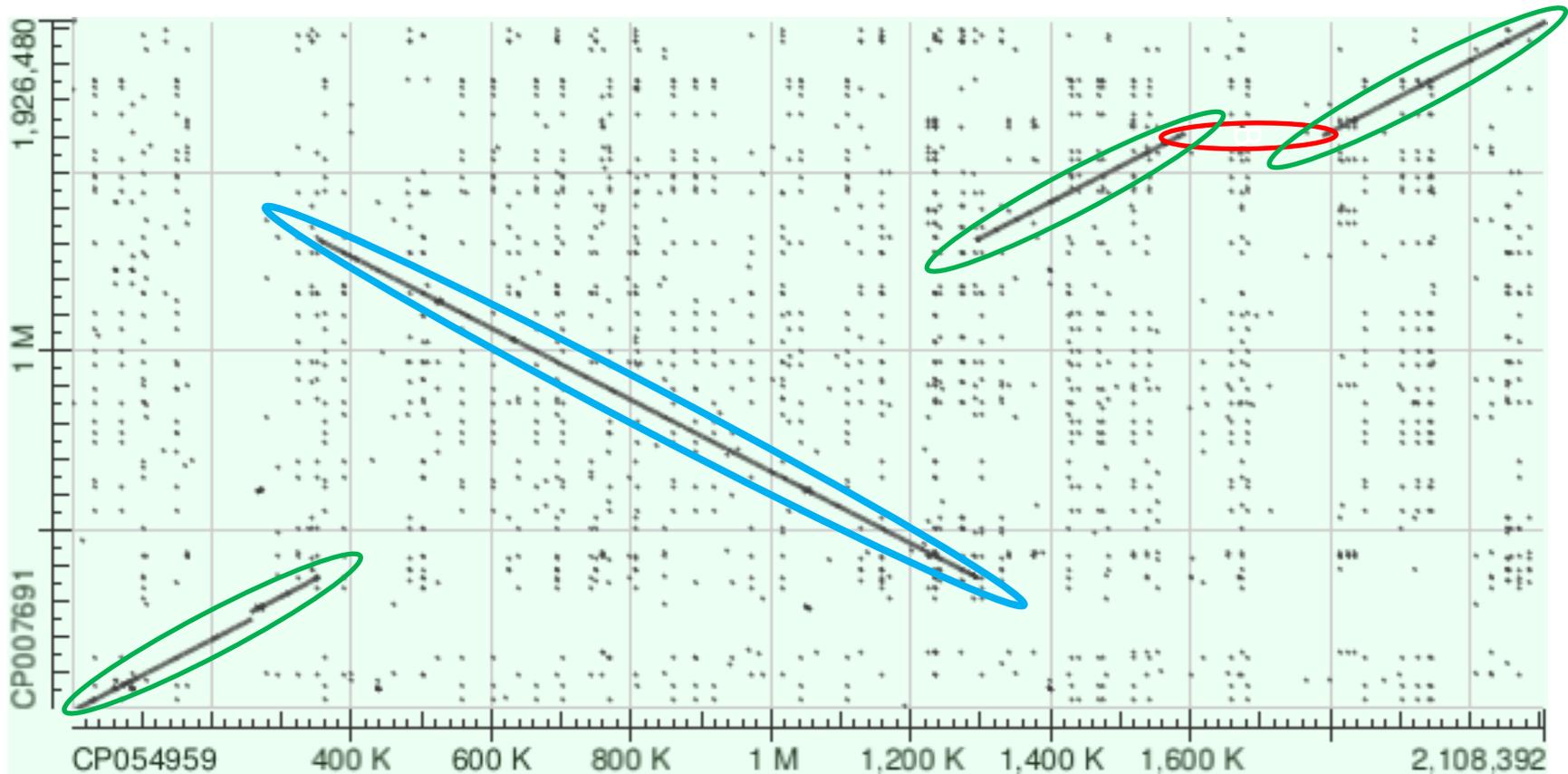
- Для ВСЕХ ПАР последовательностей строится парное выравнивание.
- Вес парного выравнивания пересчитывается в расстояние между последовательностями:
 - чем больше вес, тем меньше расстояние;
 - расстояние между совпадающими последовательностями равно 0.
- Получается матрица расстояний между послед-ми
- 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

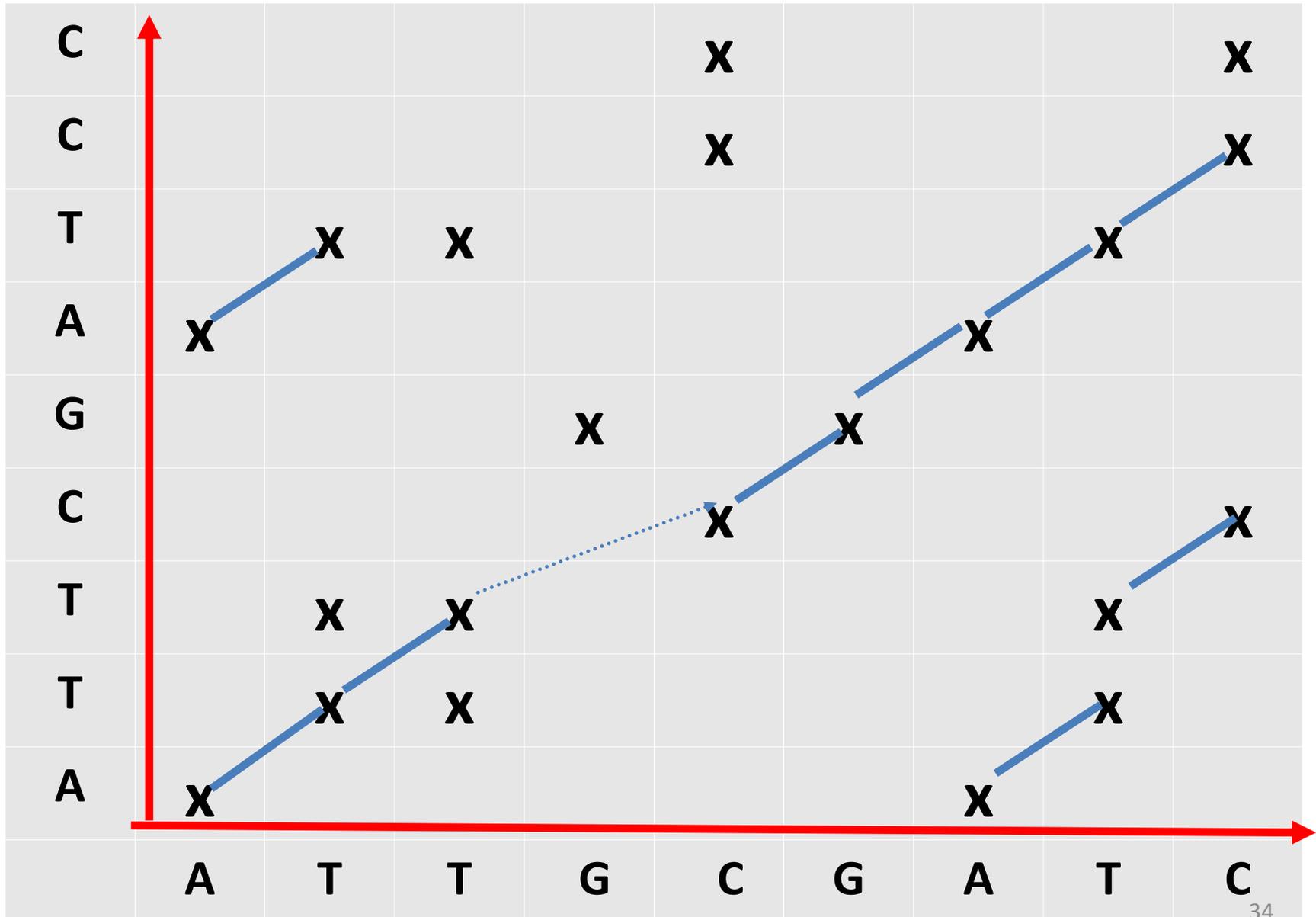


Сходство прямых цепочек

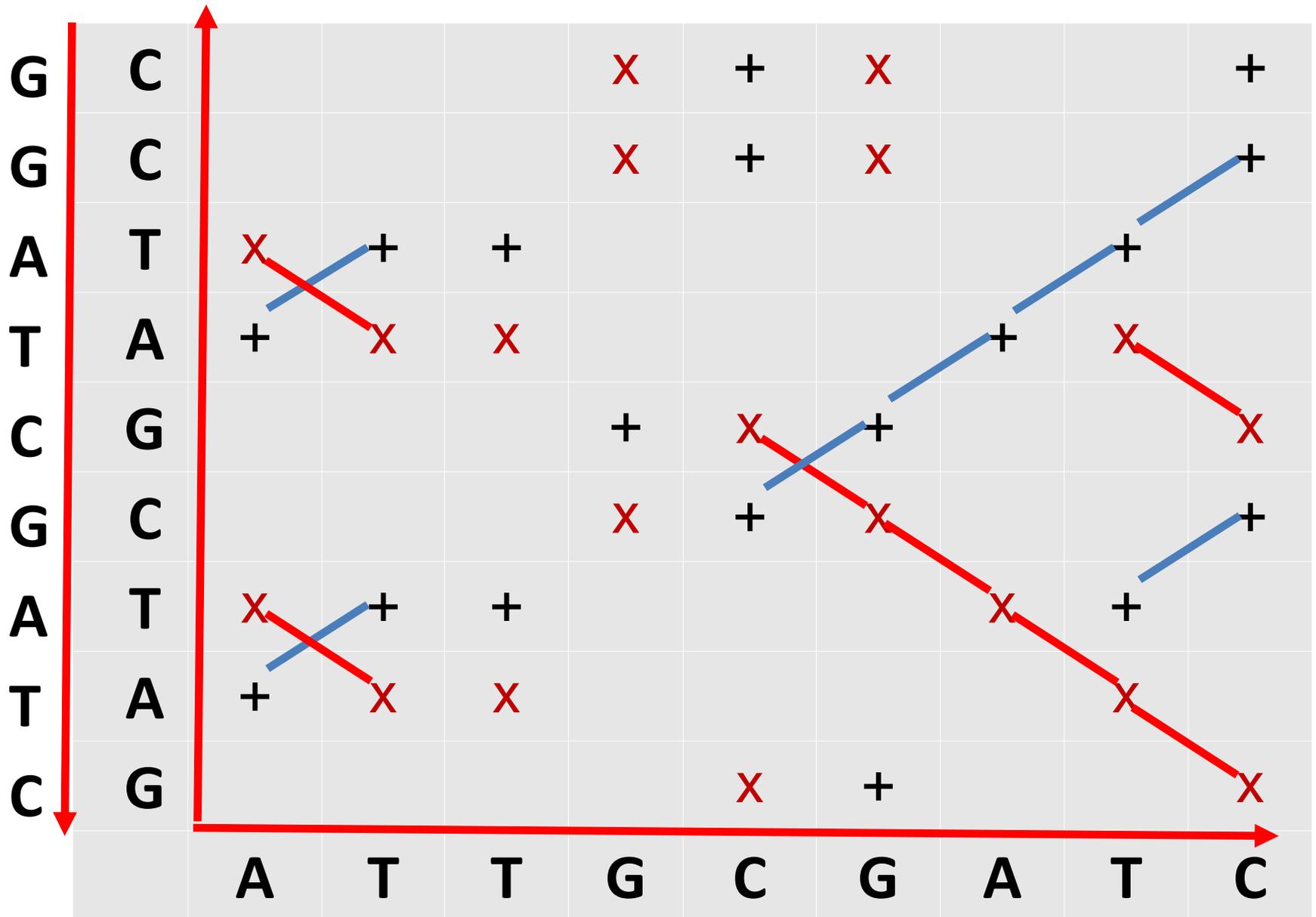
Инверсия - сходство прямой цепочки
с комплементарной второго генома

Делеция / вставка

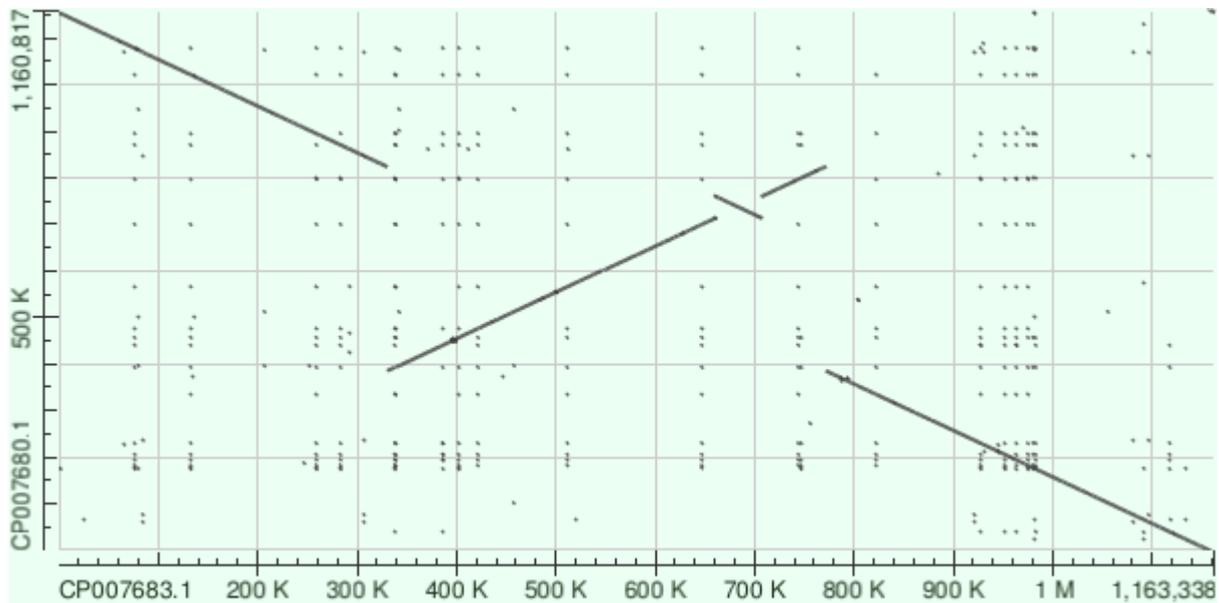
Объяснение Карты локального сходства



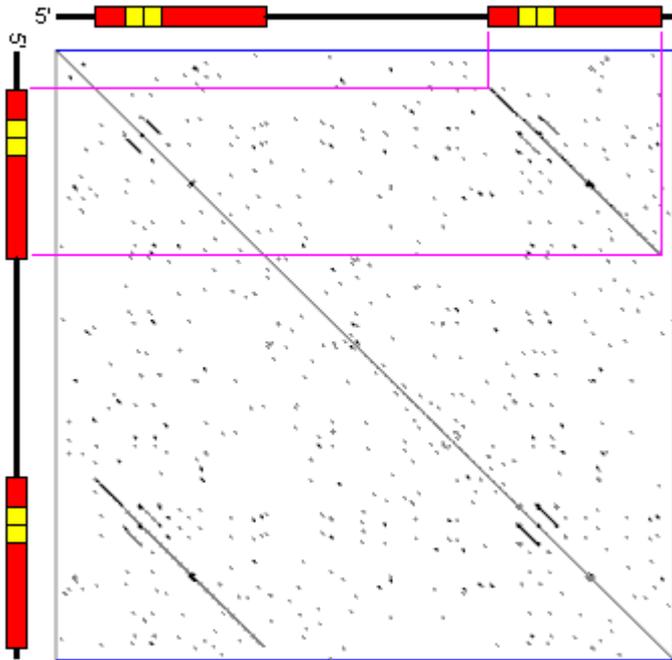
Карта сходства с учетом комплементарной цепочки



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	M	K	F	R	-	S	S	H	Y	A	-	S	
Seq2	M	K	Y	R	R	R	-	H	Y	A	-	S	
Seq3	M	E	F	R	R	R	R	S	H	Y	A	-	R

	1	2	3	4	5	6	7	8	9	10	11	12
Seq1	M	K	F	-	R	S	S	H	Y	A	S	
Seq2	M	K	Y	R	R	R	-	H	Y	A	S	
Seq3	M	E	F	R	R	R	R	S	H	Y	A	R

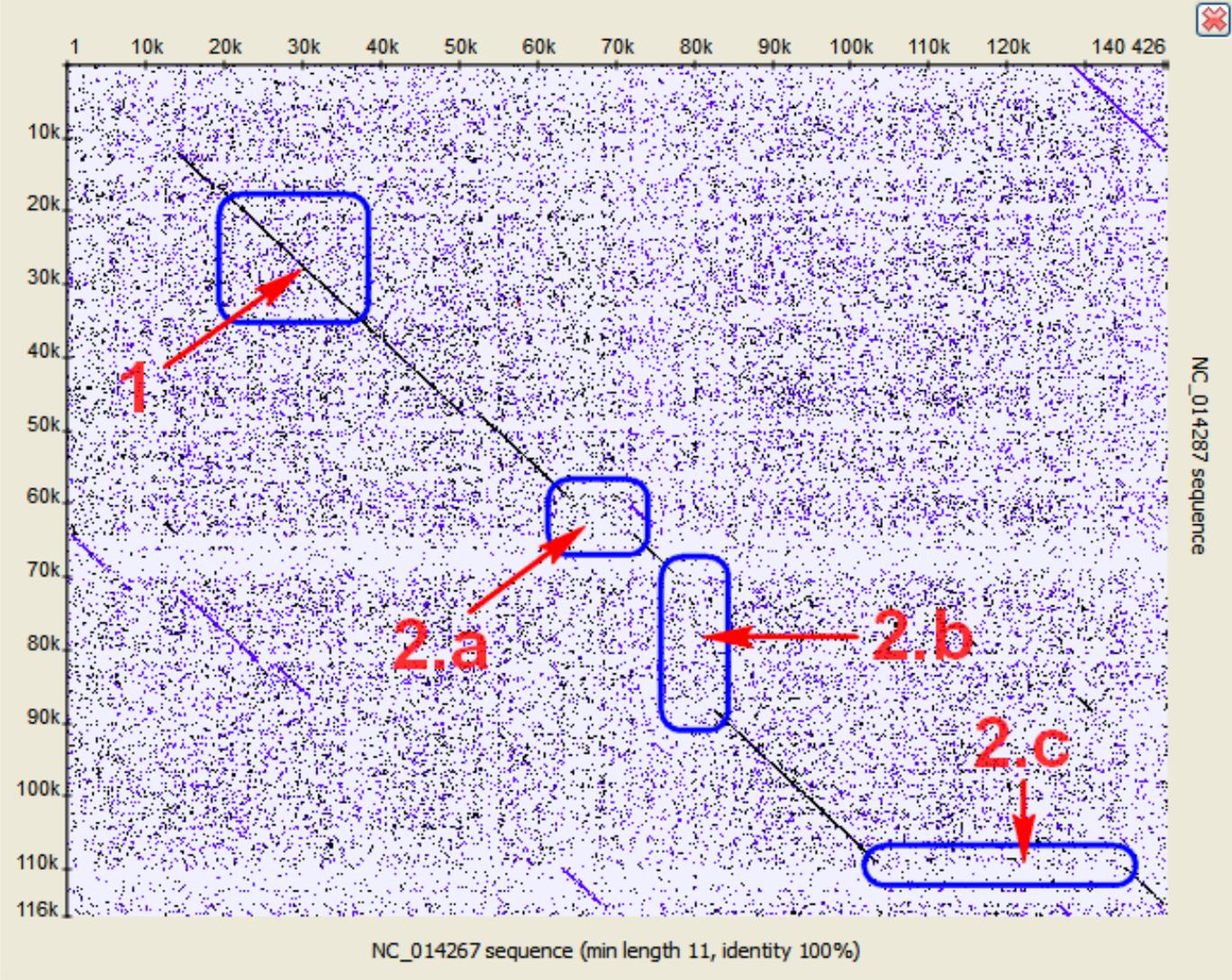
	1	2	3	4	5	6	7	8	9	10	11	12	13
Seq1	M	K	F	R	-	S	S	H	Y	A	-	S	
Seq2	M	K	Y	R	-	R	S	H	Y	A	-	S	
Seq3	M	E	F	R	R	R	R	S	H	Y	A	-	R

	1	2	3	4	5	6	7	8	9	10	11	12	
Seq1	-	M	K	F	R	-	S	S	H	Y	A	S	
Seq2	-	M	K	Y	R	-	R	S	H	Y	A	S	
Seq3	-	M	E	F	R	R	R	R	S	H	Y	A	R

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!

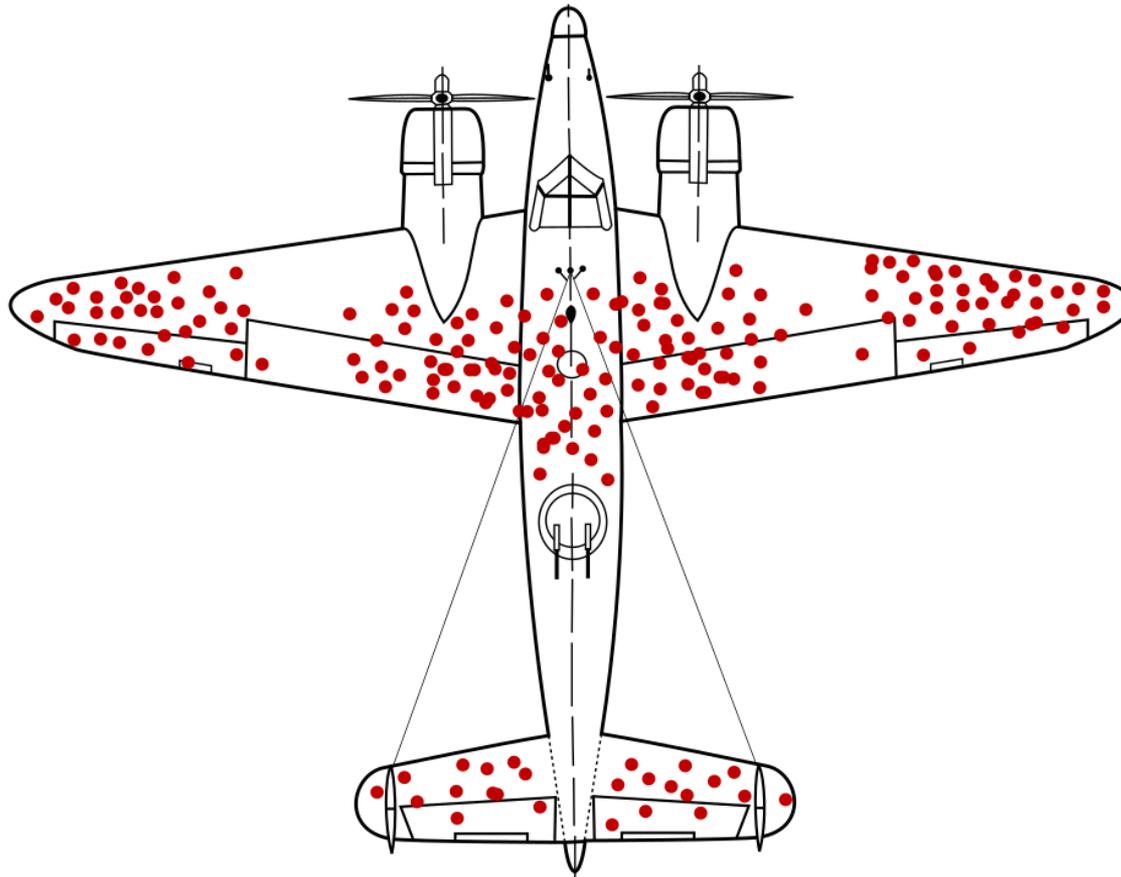
Interpreting Dotplot: Identifying Matches, Mutations, Inversions, etc

Создатель Yuliya Algaer, 2014



Пробоины на вернувшихся американских самолётах во время второй мировой войны.

Какие части укреплять броней?



Эволюция белков

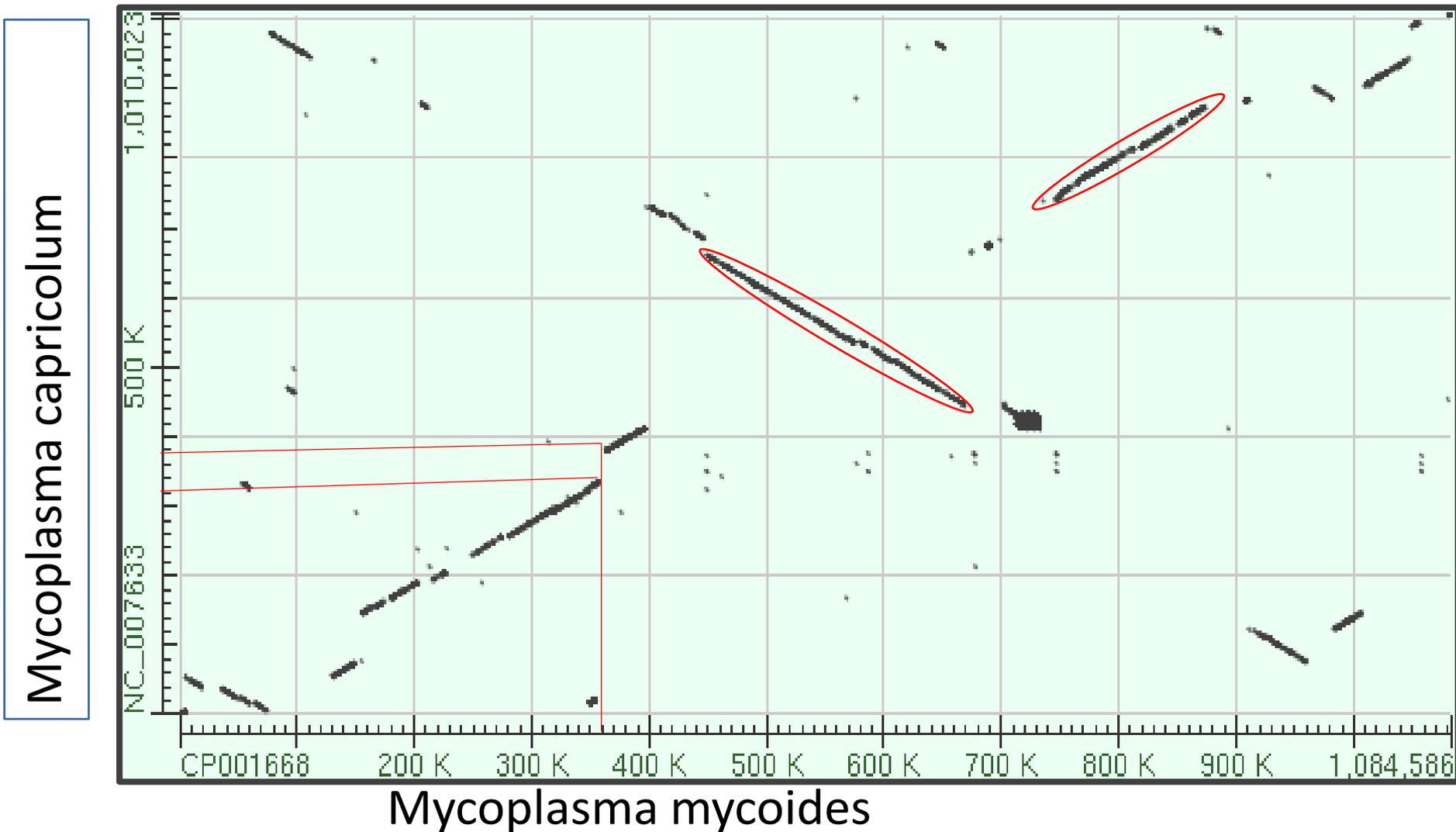
Локальная - небольшие изменения в гене
(Замены а.к. Делеции Вставки)

Большие изменения:

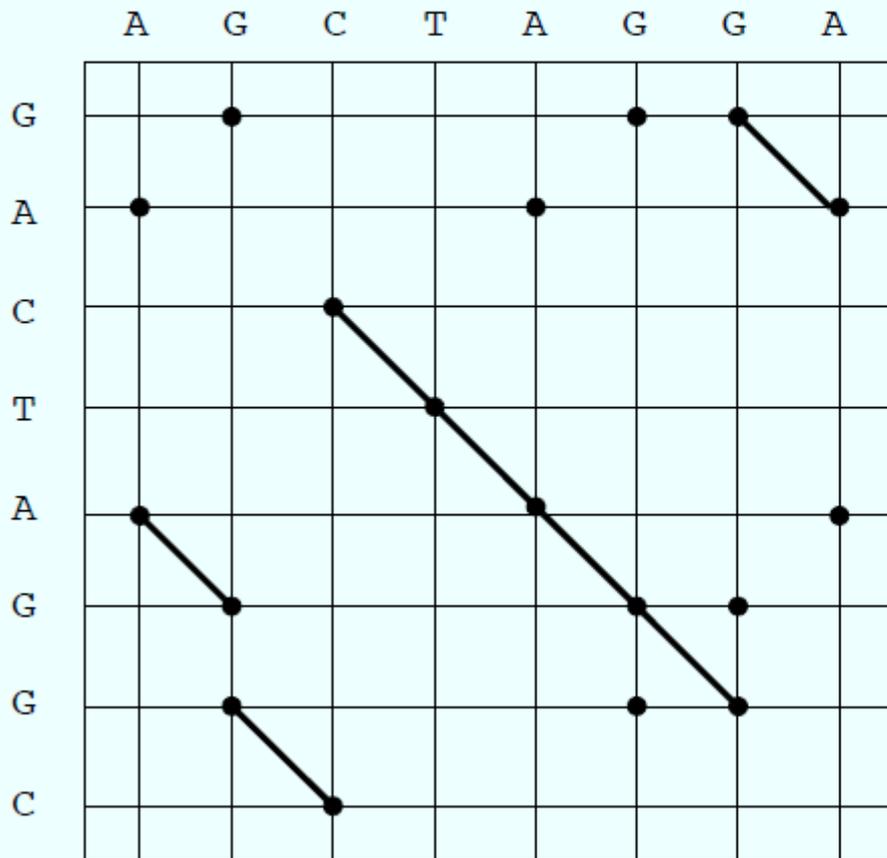
- 1) Накопленные небольшие изменения
- 2) ~~Небольшие изменения гена ведущие к большим~~
изменениям белка
 - 1) Мутация стоп кодона => удлинение последовательности белка
 - 2) Мутация кодона на стоп кодон
 - 1) гибель белка = псевдогенизация или
 - 2) Укорочение последовательности белка
 - 3) Программируемый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

2. Парное выравнивание геномов

Карта локального сходства геномов M. capricolum и M. mycoides



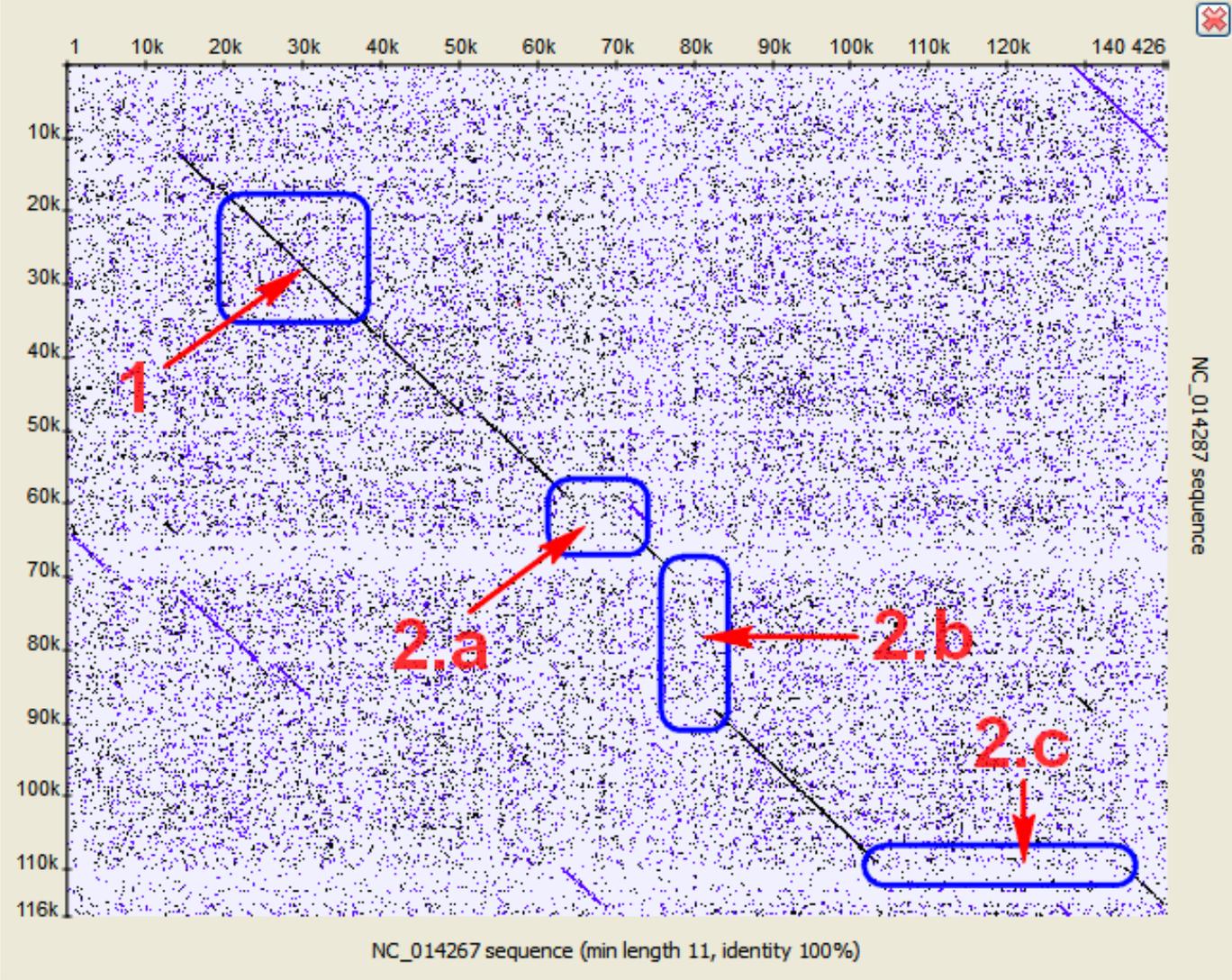
- established in 1970 by A.J. Gibbs and G.A.McIntyre
- method for comparing two amino acid or nucleotide sequences



- 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, Inversions, etc

Создатель Yuliya Algaer, 2014



ВЫРАВНИВАНИЕ

DNAK_THEAC 82 KFKVFDKEFTPQQISAFILQKIKKDA-EAFLGEPVNEAVITVPAYFNDNQRR 131
 DNAK_PICTO 82 KYKIFGKEYTPQQISAFILQKIKRDA-EAFLGEPVTDAVITVPAYFNDNQRR 131
 HSCA_ACIF2 116 RLRTVAGEKSPVEVSAEILRVLKERAVETLGGEPEGAVITVPAYFDEAQR 165
 HSCA_ACIF5 116 RLRTVAGEKSPVEVSAEILRVLKERAVETLGGEPEGAVITVPAYFDEAQR 165

DNAK_THEAC 132 QATKDAGTIAGFDVKRIINEPTAAALAYGVDKSGKSEKILVFDLGGGTLDV 182
 DNAK_PICTO 132 QATKDAGAIAGLNVRRRIINEPTAACLAYGIDKLNQTLKIVIYDLGGGTLDV 182
 HSCA_ACIF2 166 QATKDAARLAGLNVLLAEPTAAAVAYGLDKGSEGI-FAIYDLGGGTFDI 215
 HSCA_ACIF5 166 QATKDAARLAGLNVLLAEPTAAAVAYGLDKGSEGI-FAIYDLGGGTFDI 215

DNAK_THEAC 183 TIMDFGDGVFQVLSSTSGDTRLGGTDMDEAIVNYIADDFQKKEGIDLKDRS 233
 DNAK_PICTO 183 TIMDFGQGVFQVLSSTSGDTHLGGTDMDEAIVNFLADNFQRENGIDLKDHHS 233
 HSCA_ACIF2 216 SILRLQAGVFEVLATAGDSALGGDDMDHALAEWLMQE-EGGDASDPLW 262
 HSCA_ACIF5 216 SILRLQAGVFEVLATAGDSALGGDDMDHALAEWLMQE-EGGDASDPLW 262

DNAK_THEAC 234 AYIRLRDAAEKAKIELSTTLSTDIDLPIYITVTNSGPKHKIKMTLTRAKLEEL 284
 DNAK_PICTO 234 AYIRLRDAAEKAKIELSTVLETEINLPIYITATQDGPKHLQYTLTRAKFEEL 284
 HSCA_ACIF2 263 RRQVLQQ-ARTAKEALSVAEET-MIVLTPSGRAAREIKLSRGRLES 307
 HSCA_ACIF5 263 RRQVLQQ-ARTAKEALSVAEET-MIVLTPSGRAAREIKLSRGRLES 307

DNAK_THEAC 285 ISPIVERVKGPIDKALEGAKLKKTEITKLLFVGGPTRIPYVRKYVEDYLG 335
 DNAK_PICTO 285 IAPIVDRSKVPLDTALEGAKLKKGDIDKIILIGGPTRIPYVRKYVEDYFGR 335
 HSCA_ACIF2 308 IQPVIQRSLPACRRALRDAGLKLDEIEGVVLVGGATRVPVAVRAMVEEFFRQ 358
 HSCA_ACIF5 308 IQPVIQRSLPACRRALRDAGLKLDEIEGVVLVGGATRVPVAVRAMVEEFFRQ 358