

Программы множественного выравнивания

На материале последовательностей гомологичных белков.

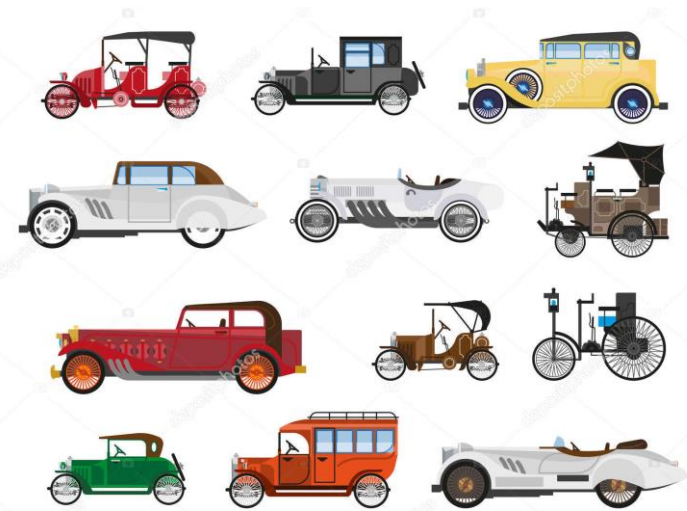
Задания на дом

1. (*) Написать программу для сравнения двух выравниваний одних и тех же последовательностей. *Продемонстрируйте чему вас научили в программировании!!! (*) потому, что не уверен, что все справятся.*
2. Сравните выравнивания двух пар программ. *Вручную или с использованием программ, можно, написанных коллегами (со ссылкой)*
3. Постройте выравнивание по совмещению полипептидных цепей трёх доменов из одного семейства Pfam. Сравните с выравниванием построенным программой выравнивания последовательностей
4. Коротко опишите одну программу множественного выравнивания. *Литературное задание. Google и Pubmed вам в руки*

1. Выравнивание последовательностей гомологичных белков отражает *непрерывную эволюцию гена белка*

Непрерывная эволюция – это небольшие изменения последовательности белка, вызванные локальными небольшими локальными изменениями в последовательности гена белка.

А потом бац
электоавто – крупная перестройка
Самокаты – крупная перестройка
путем делеции двух колес



© depositphotos

Image ID: 31937606 www.depositphotos.com

Много доменные белки в эволюции образуются в результате последовательных единовременных крупных перестроек в генах белков

Домен белка – единица непрерывной эволюции белков

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

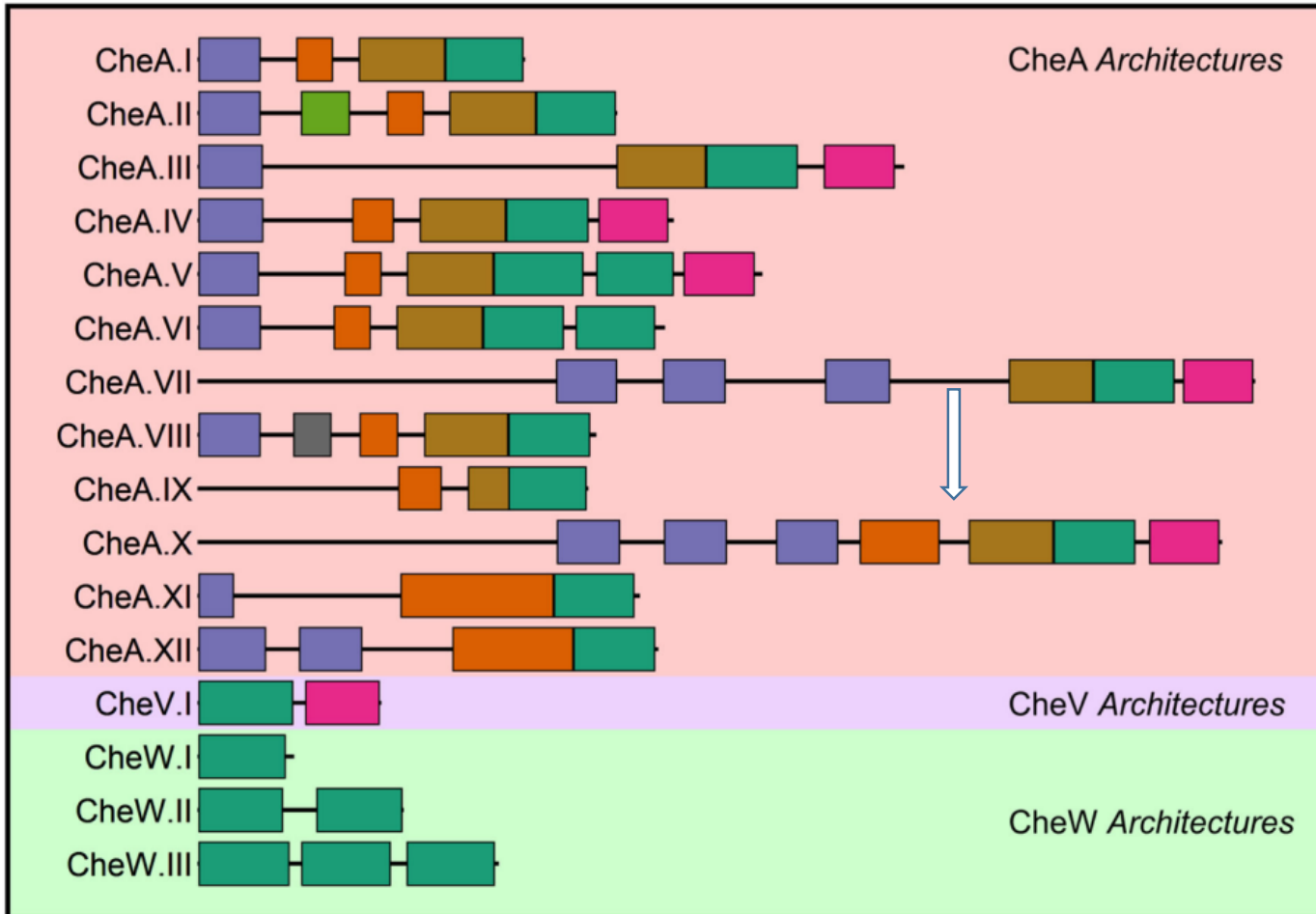
ДОМЕННАЯ АРХИТЕКТУРА – последовательность доменов в белке

Эволюционные домены в белке изображают так

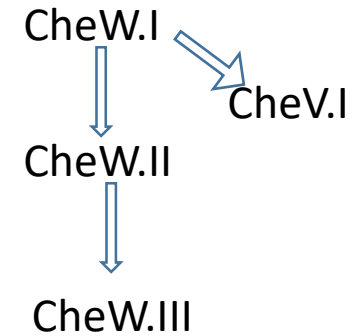
Che белки участвуют в хемотаксисе бактерий

Pfam Domain

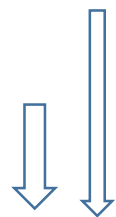
- CheW-like
- H-kinase dimerization
- HPT
- Response regulatory
- CheY-binding
- HATPase_c
- P2



Крупные перестройки



Далее предложите варианты



Доменные архитектуры белков,
содержащих домен CheW-like

2. Программы множественного выравнивания последовательностей (MSA = Multiple Sequence Alignment)

Их много с разными алгоритмами, использующих
разные эвристики

Список популярных программ MSA

Программа	Год	в Jalview	на kodomo	Website
ClustalO		yes		https://www.ebi.ac.uk/jdispatcher/msa/clustalo
ClustalW	1994	yes		http://www.clustal.org/
MAFFT	2002	yes	yes	https://mafft.cbrc.jp/alignment/software/
MSAProbs	2010	yes		http://msaprobs.sourceforge.net/
MUSCLE	2004	yes	yes	http://www.drive5.com/muscle/
ProbCons	2005	yes		http://probcons.stanford.edu/
T-Coffee	2000	yes		http://tcoffee.crg.cat/
Glprobs	2015	yes		
DiAlign	1998			http://dialign.gobics.de/
FAME	2020			http://github.com/naznoosh/msa
Kalign	2005			https://www.ebi.ac.uk/Tools/msa/kalign/
NX4	2019			https://www.nx.io
PRANK	2008			http://wasabiapp.org/software/prank/
Probalign	2006			http://probalign.njit.edu/standalone.html

Цель множественного выравнивания последовательностей гомологичных доменов или белков:

- Реконструкция эволюции белков от общего предка.
А именно, в колонке выравнивания стоят аминокислотные остатки (ако), унаследованные от ако общего предка всех этих белков.
Наследуются ДНК гена, т.е. кодоны, это подразумевается.
- Основа для установления гомологии – сходство последовательностей. Оно неравномерно на разных участках (слайд +1)
- Достигнута ли цель практически - не проверяемо (не считая парочки долговременных экспериментов (Ленский – E.coli, Кондрашов - Schizophyllum commune)
- На практике стремятся к этой цели, но следует понимать, что она недостижима. Разные программы множественного выравнивания используют разные алгоритмы и разные эвристики.

Как сравнивать программы?

- Нужны БД “идеальных” выравниваний (Alignment Benchmarks)
- Наиболее признанной является BAlIBASE (Benchmark Alignment dataBASE)
- Есть и другие Alignment Benchmarks:
 - SABMark(SequenceAlignmentBenchMark)
 - OXBench(OXfordBenchmark)
 - SMART(SimpleModularArchitectureResearchTool)
 - PREFAB(ProteinREferenceAlignmentBenchmark)
 - QuanTest

[1] Zhang Y, et al. A survey on the algorithm and development of multiple sequence alignment. Brief Bioinform. 2022 May 13;23(3):bbac069

Результаты сравнения программ на BaliBase [1]

Program	R1-1	R1-2	R2	R3	R4	R5	AvgScore	Tottime(s)
MSAProbs	0.441	0.865	0.464	0.607	0.622	0.608	0.607	12382
Probalign	0.453	0.862	0.439	0.566	0.603	0.549	0.589	10095
MAFFT	0.439	0.831	0.450	0.581	0.605	0.591	0.588	1475
Probcons	0.417	0.855	0.406	0.544	0.532	0.573	0.558	13086
ClustalOmega	0.358	0.789	0.450	0.575	0.579	0.533	0.554	539
T-Coffee	0.410	0.848	0.402	0.491	0.545	0.587	0.551	81041
Kalign	0.365	0.790	0.360	0.476	0.504	0.435	0.501	21
MUSCLE	0.318	0.804	0.350	0.409	0.450	0.460	0.475	789
FSA	0.270	0.818	0.187	0.259	0.474	0.398	0.419	53648
DiAlign	0.265	0.696	0.292	0.312	0.441	0.425	0.415	3977
PRANK	0.223	0.680	0.257	0.321	0.360	0.356	0.376	128355
ClustalW	0.227	0.712	0.220	0.272	0.396	0.308	0.374	766

Колонки 2-7 – разные наборы тестовых выравниваний. Указан ТС вес сравнения с референсным выравниванием. Значения от 0 до 1, чем больше – тем более похоже выравнивание на референсное. Последняя колонка – время работы программы

Le et al., 2016

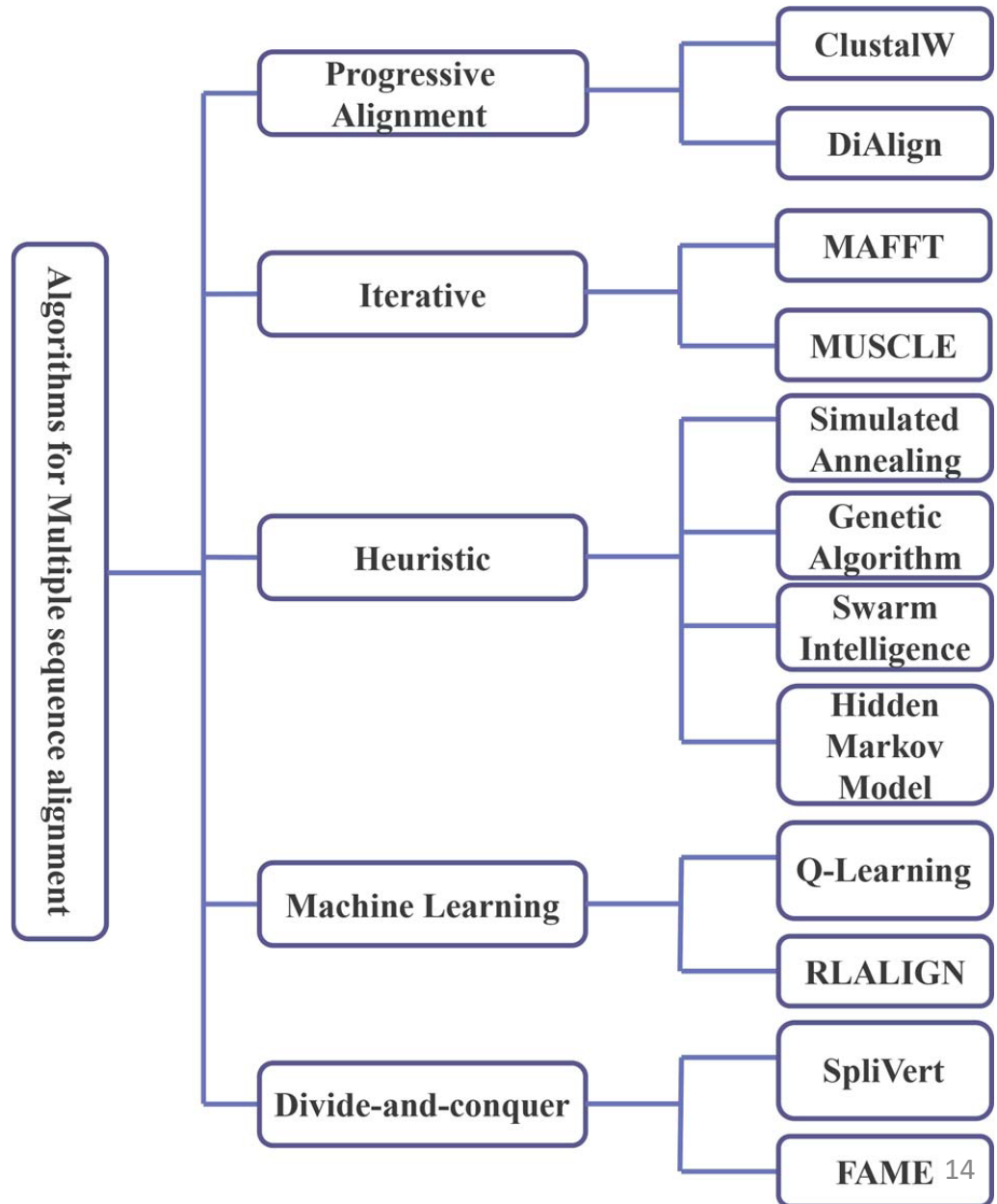
Table 1. The prediction accuracy for alignments of 200 sequences for 238 Pfam families. Aligner settings Prediction Accuracy (in %)

MAFFT L-INS-i	78.94 *
MAFFT—Default	78.19
MAFFT—Fast Mode	77.53 *
Clustal Omega—2 iter	78.36 *
Clustal Omega—1 iter	78.56 *
Clustal Omega—Default	78.63
MUSCLE—2 iter	78.17
MUSCLE—Default	78.13
MUSCLE—1 iter	77.29 *
PASTA—Default	78.70
T-Coffee—Default	78.45
Kalign 2—Default	77.93
Clustal W2—Default	77.13
HMMER—Default	77.86

For aligner settings from the same aligner, the sign (*) signifies that the score is significantly different (higher or lower) from the default score with $P < 0.01$ using the Wilcoxon signed rank test.

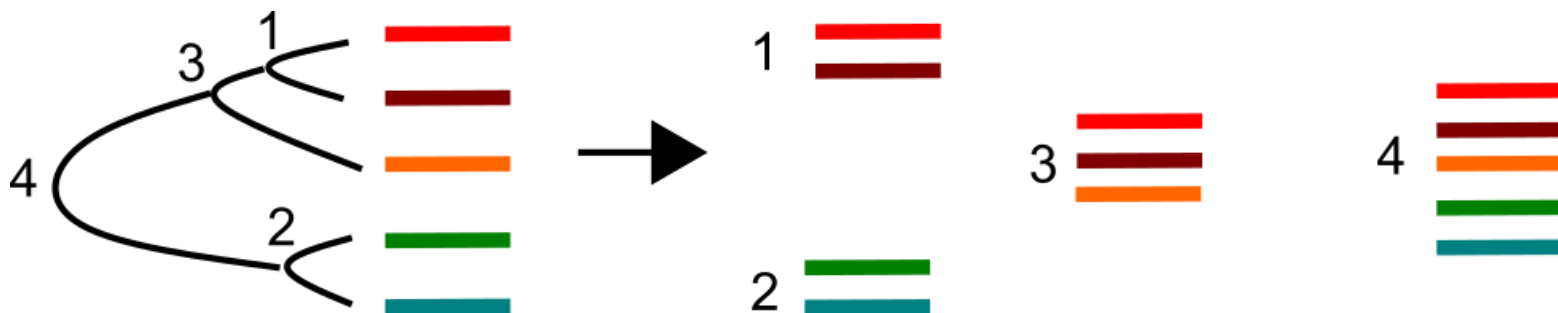
Классификация алгоритмов из того же обзора [1]

Пока не будем углубляться в алгоритмы – чтобы не утонуть)



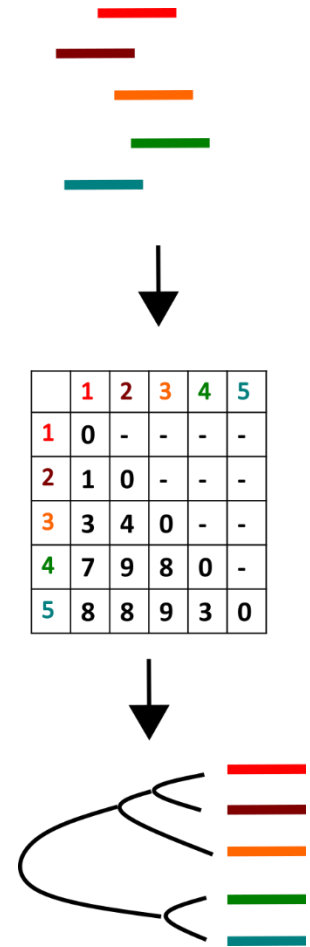
Прогрессивное выравнивание

- Основная идея: выравнивание двух выравниваний с помощью динамического программирования
- Этапы:
 - Построить дерево родственности всех последовательностей – направляющее дерево
 - Выравниваем стопки выровненных последовательностей от листьев до корня

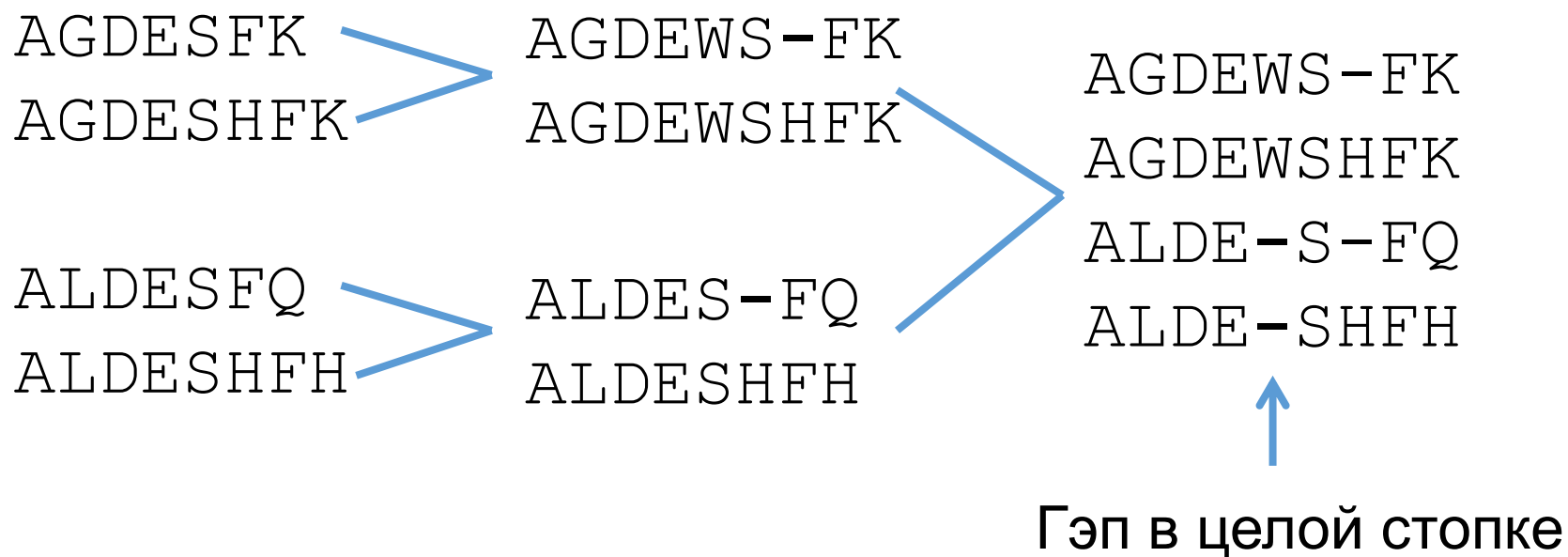


Построение направляющего дерева

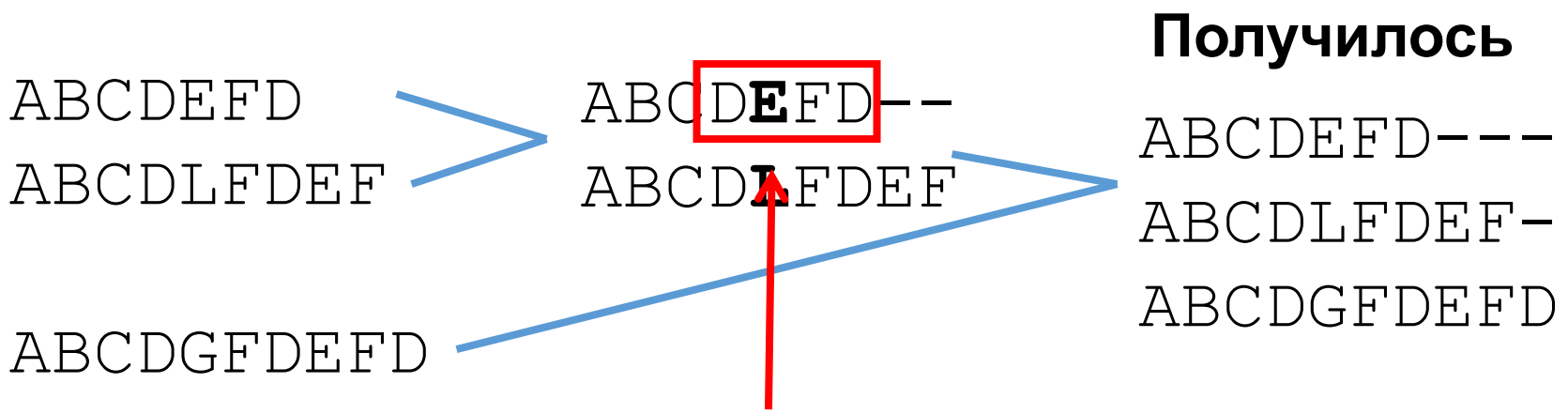
- Для ВСЕХ ПАР последовательностей строится парное выравнивание.
- Вес парного выравнивания пересчитывается в расстояние между последовательностями:
 - чем больше вес, тем меньше расстояние;
 - расстояние между совпадающими последовательностями равно 0.
- Получается матрица расстояний между послед-ми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



Пример прогрессивного выравнивания



Проблема: если ошиблись, то ошибка никуда не денется



Получилось

ABCDEFD---
ABCDLFDEF--
ABCDGFDEFD

Но уже поздно
передвигать этот блок

А хотелось бы так:

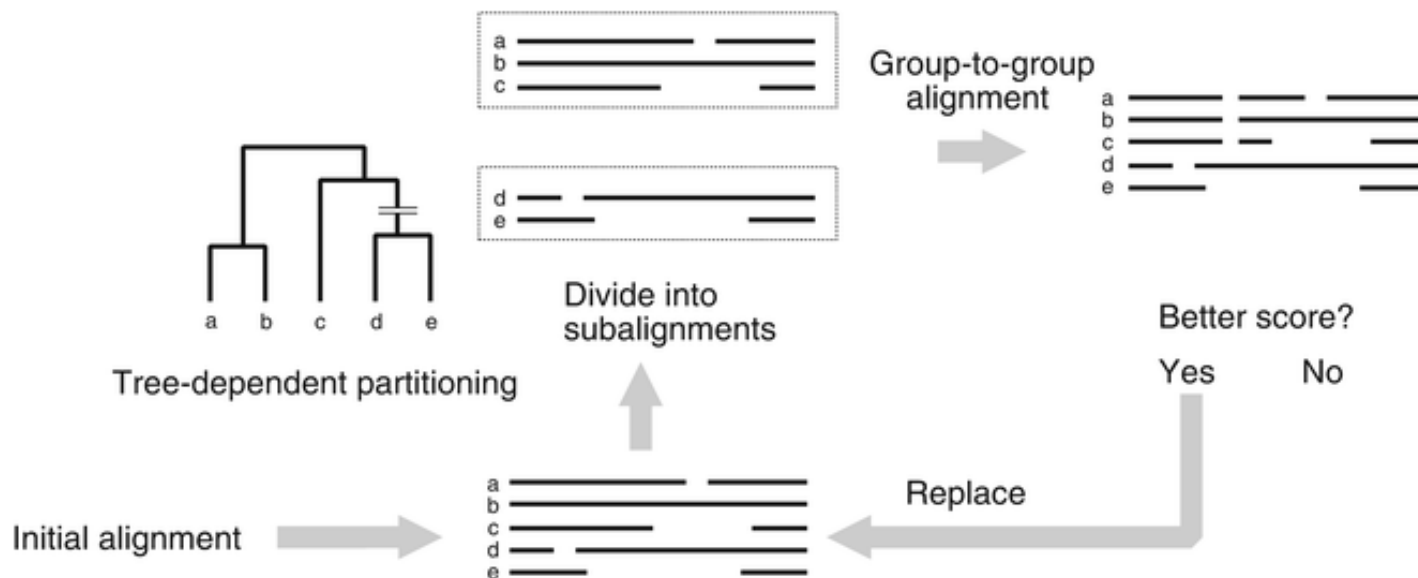
ABC---DEFD
ABCDLFDEF--
ABCDGFDEFD

Проблема прогрессивных выравниваний – гэп, появившийся в начальных выравниваниях (например, парных) сохраняется до конца

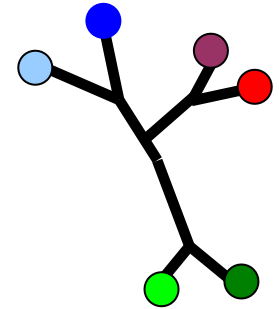
Из-за этого необходим последний этап – улучшение выравнивания (refinement)

Итеративное рафинирование выравнивания

- i. Построить множественное выравнивание
- ii. Разделить его на две группы
- iii. Перевыровнять две группы. Повторить ii

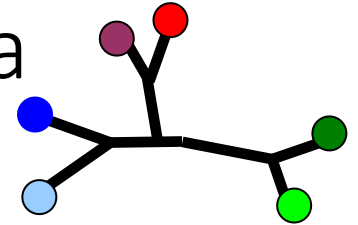


Прогрессивные = иерархические алгоритмы выравнивания многих последовательностей



- 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

Книга

Multiple Sequence Alignment Methods and Protocols

Edited by Kazutaka Katoh

Research Institute for Microbial Disease, Osaka University,
Osaka, Japan

https://link.springer.com/protocol/10.1007/978-1-0716-1036-7_17

Free

Слайды +1, +2, +3 – выдержки из книги

Алгоритмы и программы MSA

Progressive Method

A reasonable and widely used heuristic is the progressive method [2–4].

In this strategy, a tentative tree, called a “guide tree,” is built based on an all-to-all approximate comparison. Then, the sequences are aligned from the leaves to the root on the tree, in a group-to-group manner.

When the calculation reaches the root, the full MSA is obtained.

Many MSA programs use the progressive method as a part of the calculation. Among them, PRANK (Chapter 2) has a notable point that it rigorously considers insertions and deletions on the guide tree.

Алгоритмы и программы MSA

Iterative Refinement

The progressive method has a well-known problem that errors can occur in early steps (i.e., close to a leaf) of the guide tree, and those errors remain in the final step (the root of the guide tree).

One effective solution is to correct this type of mistake is iterative refinement [5–7]. The procedure is:

- (i) construct an initial MSA;
 - (ii) divide the MSA into two groups; (iii) re-align the two groups;
- repeat (ii) and (iii). This technique is used in Prrn5 (Chapter 5), Clustal Omega (Chapter 1), MAFFT (Chapter 11), and MSAProbs (Chapter 3).

Алгоритмы и программы MSA

Consistency

Another important idea to overcome the limitations of the progressive method is consistency [8, 9]. In the tree-based consistency transformation technique, proposed first in Notredame et al. [10], when aligning two sequences (A and B), other sequences (e.g., C) in the dataset are also used.

That is, in addition to direct alignment between A and B, an alternative alignment between A and B is synthesized by using alignments AC and BC.

Alignment AB is recalculated by considering such alternative alignments and used in the progressive alignment step. As a result, alignment errors in early steps are efficiently suppressed.

This method was further elaborated to use probabilistic pairwise alignments by a pair hidden Markov model in ProbCons [11]. MSAProbs (Chapter 3) represents this type of MSA method and gives highly accurate MSAs.

3. Сравнение разных выравниваний тех же последовательностей

Построенных разными программами

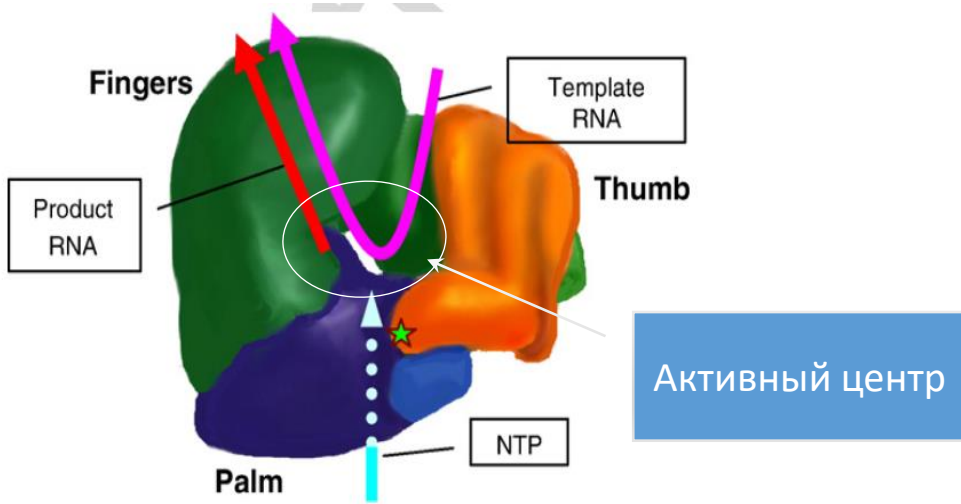
Цель множественного выравнивания последовательностей гомологичных доменов или белков:

- Реконструкция эволюции белков от общего предка.
А именно, в колонке выравнивания стоят аминокислотные остатки (ако), унаследованные от ако общего предка всех этих белков.
Наследуются ДНК гена, т.е. кодоны, это подразумевается.
- Основа для установления гомологии – сходство последовательностей. Оно неравномерно на разных участках (слайд +1)
- Достигнута ли цель практически - не проверяемо (не считая парочки долговременных экспериментов (Ленский – E.coli, Кондрашов - Schizophyllum commune)
- На практике стремятся к этой цели, но следует понимать, что она недостижима. Разные программы множественного выравнивания используют разные алгоритмы и разные эвристики.

РНК зависимая РНК полимераза (RdRP), консервативные участки

```

*      320      *      340      *      360      *      380      *      400      *      420      *      440      *      460
FKTMIRFGDVGDLDDFFSADASLSPFMIREA..GRIMSELS...GTPSHFGTALINTIIYSKHLIYNCCY.....HVCGSMPSGSPCTALNSTINNVLYYVFSKIFGKSPVFF.....CQALKILC.YGDDVIVFSDRV
EVAMQG.FERVYDVDYSNEDSTHSMVAFRL..A...EEFF.TPENGFDPLTREYLESLAISTHAFEEKRF.....LTGGLPSGCAATSMNTIMNIIIRAGLYLTYKNFEFDD.....VKVLS.YGDDLIVATNYQL
ETHFAQ.YKNVWDVLYSADANHCSDAMNMFEEVFERTEFG.....FHPNAEWILKTLVNTTEHAYENKRI.....VVEGCMPSGCSATSIINTILNNTYVLYALRRHYEGVELDT.....YTMIS.YGDDIVASDYDL
.....WSLCVATIVSDHDTFWPGWLRDLICDELINMGYA.PWVVKLFETSLKLPVYVGAFAPEQGHTLLGDPSNPDLVGLSSGQGATDLMGTLIMSTIYLVMLQDHTAPHLNSRIKDMPSACRFLDSYWQGHEETROIS.KSDDAILGWTKGR
LRLRLE.NWVYCDADGSOEDSSSLTPYLINAV..LTLRSTYMEDWDVGLQMLRNLYTEIVYTPISTPDGTIV.....KKFRGNNSGQPSIVDNLSLMVVIAMHYALIKECFEVEEID.....STCVFFV.NGDDLIAVNPEK
HDKLNRPGLWLGSGDGRDSSIDPFFFDVV..KTKRKHFL..PSEHHRAIDLIIYDEILNTPICLANGMVI.....KKNVGTQR.QPSTVDNTLVMITAFLYAYIHKTDGRELAL.....LNERFIFVC.NGDDNKFAISPQF
AISLASFSYPYGFNGDFANEDGMFHPSSFSMV..SELANIFY...GNFLSTERDNLTRMLTNRFSIMKGAIL.....RVPGGSPSGFFMTVFNSEINLFYLQSAWIMLARFNGRQDISH.....PCNFPKYVRACV.YGDDNIVAIMEV
AARMKEKGNVDVLCODYSSEFDGLLSKQVMDVI..ASVINELC.GGEDQLKNARRNLMACCSRIAICKNTVW.....RVECGIPSGFFMTVFNSEINLFYLRHYHKIMREQQAPELMV.....QSFDKLIIGLVT.YGDDNLSVNAVV
YAEHAK.YKNHFDADYIANDSTQNRQIMTES..FSIMSRLT...ASPELAEVVAQDLLAPSEMDVGDYVI.....RVKEGLPSGFFPCTSQVNSINHWITLICALSEATGLSPDVV.....QMSYFYSFYGDDIVSTDIDF
NNLTSKASDFLCLDYSKFDSTMSPCVVRIA..IDLADCC...EQTELTKSVVLTILKSHFMTILAMIV.....QTKRGLPSGMPFTSVINSICHWLLWSAAVYKSCAEIGLHCS.....NLYEDAPFYT.YGDDGVYAMTEMM
IQRIKS.AAKVYAVDYSKWDSTQSPRVSAA..IDLRYFS...DRSPIVDSAANTLKSPPIAIFNGVAV.....KVSSGLPSGMPFTSVINSINHCLYVGCAILQSLEARGVPVTW.....NLFSTFLMMT.YGDDGVYMFPMFM
TKRLERPKHDRYCVLYSKWDSTQPPKVTSSQS..IDLRHFT...DKSPIVDSACATLKSNIPIGIFNGVAF.....KVAGGLPSGMPFTSVINSINHCLMVGSAVVKALEDSGVRVTW.....NIFDSMDLFT.YGDDGVYIVPPLI
D      D      g      sg      T      n3      gDD
    
```



На каких участках выравнивание правильное – совпадает с эволюционным?

Точное сравнение двух выравниваний.

- Два выравнивания I и II тех же последовательностей совпадают если в каждой колонке i ($i = 1, 2, \dots, N$) выравниваний I и II стоят те же самые буквы. Буквы не в смысле а.к.о., а в смысле номера буквы в последовательности (рис. 1.)
- Различие выравниваний I и II определяется числом колонок, для которых это не так и их расположением

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

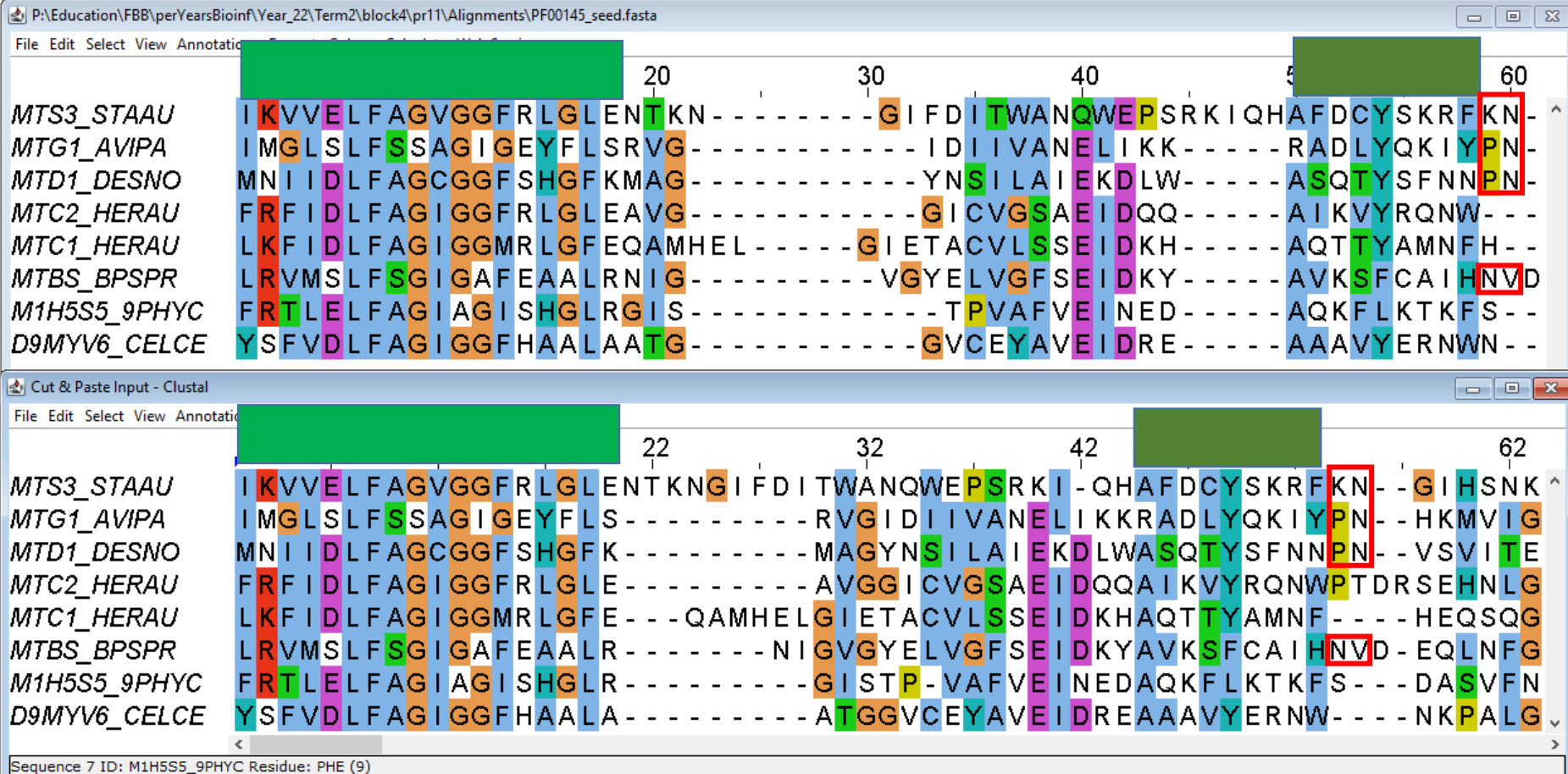
	1	2	3	4	5	6	7	8	9	10	11	12	13
Seq1	M	K	F	R	-	S	S	H	Y	A	-	S	
Seq2	M	K	Y	R	-	R	S	H	Y	A	-	S	
Seq3	M	E	F	R	R	R	R	S	H	Y	A	-	R

	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!

ПРИМЕР. На рис. выравнивания совпадают на двух выделенных зеленым участках

Ещё есть совпадения не на всех, а на подмножествах последовательностей. НАЙДИТЕ!



Продолжение. Разметьте самостоятельно!

	70	80	90	100	110
<i>MTS3_STAAU</i>	---GIHSNKDIAQVSDE---	EMAN	TEADMIVGGFPCQDYSVAR	---	---
<i>MTG1_AVIPA</i>	---HKMVI	DIRDQRIFNKVLNIALTN	-QVDFLIASPPCQGM	SVAG	KNRDV
<i>MTD1_DESNO</i>	---VSVITEDITTLDPG	DLKISVSD	-VDGII	GGPPCQGFSL	SG
<i>MTC2_HERAU</i>	---PTDRSEHNLGDIT	TLQQLPA	---	HDLVVGGVPCQPWSI	AGK
<i>MTC1_HERAU</i>	---EQSQGDITQIQ	---	DFP	-S	FDLLAGFP
<i>MTBS_BPSPR</i>	D--EQLNFGDVS	SKIDKK	---	KLP	-E
<i>M1H5S5_9PHYC</i>	---DASVFNDVTKFTKS	---	DFPED	---	IDMITAGFPCTGFSI
<i>D9MYV6_CELCE</i>	---KPALGDITDDAND	EGVTL	LRGYDGP	IDVLTGGFPCQPFSK	SGA

	62	72	82	92	102
<i>MTS3_STAAU</i>	N--GIHSNKDIAQVS	DEEMANT	EA	---	DMIVGGFPCQDYSVARSLNGE
<i>MTG1_AVIPA</i>	N--HKMVI	DIRDQRIFNKVLNIALTN	-QVDFLIASPPCQGM	SVA	---
<i>MTD1_DESNO</i>	N--VSVIT	EDITTLDPGDLKISVSDV	---	DGII	GGPPCQGFSL
<i>MTC2_HERAU</i>	TDRSEHNLGDITTLQ	---	QLPAH	---	DLVVGGVPCQPWSIA
<i>MTC1_HERAU</i>	---HEQSQGDITQIQ	---	DFPSF	---	DFLLAGFP
<i>MTBS_BPSPR</i>	VD-EQLNFGDVS	SKID	---	KKKL	PEF
<i>M1H5S5_9PHYC</i>	---DASVFNDVTKFT	---	KSDF	---	PED
<i>D9MYV6_CELCE</i>	---NKPALGDITDDA	NDE	GVTL	LRGYDGP	IDVLTGGFPCQPFSK

Take home message

To compare alignments I and II

1. Sort sequences by ID in both alignments.
2. Если в двух блоках без гэпов есть хотя бы одна одинаково выровненная позиция, то блоки выровнены одинаково

3. Алгоритм сравнения

1) Для каждого выравнивания идём по колонкам $N = 1, 2, 3, \dots$ и составляем вектор S :

$N: S_i(N) = s_1, s_2, \dots, s_n$

s_1 – номер буквы в первой последовательности

s_2 – номер буквы во второй последовательности

.....

Если в последовательности i стоит гэп, то $s_i = "-"$

2) Если $S_i(N) = S_{ii}(N')$, то колонка N выравнивания I выровнена одинаково с колонкой N' выравнивания II

3) Последовательно идущие в обоих выравниваниях одинаково выровненные колонки объявляются блоком одинаково выровненных фрагментов

Для ручного сравнения 2х
выравниваний тех же
последовательностей удобно
использовать сервис

[VerAlign: a multiple sequence alignment
assessment tool](#)

VerAlign раскрашивает 2e выравнивание цветами первого.
 Это облегчает поиск неодинаково выровненных столбцов

```

GVLAGLIDNDPS ----CKYAYEQNNK----TRFLEKSISEVDGKELNALYFPNNQ--HKILLVGCAPCQDFSSQYTK-
EIVAAAVDNWRP ----AINTYQQNF----THPIHELDLAQIDA AVSLIKTHS--PELITGGPPCQDFSSAG--
NSILAEIKDLW----ASQTYSENNPN--VSVITEDITTLDPG--DLKISVSD--VDGIIGGPPCQGFSLSG--
ICVGSAEIDQQ----AIKVYRONW----PTDRSEHNLGDIT--TLQQLP A---HDLV VGGVPCQPWSIAGK-
ACVLSSEIDKH--AQT---TYAMNFH----EQSQGDITQIQ-----DFP--S--FDFL LAGFPCQPF SYAGK-
ELVGFSEIDKY----AVKSFCAIHNV D--EQLNFGDVSKIDKK----KLP--E--FDLLVGGSPCQSFSVAGH-
KCVFSSEWDKY--AAQ---TYEANYG----EKPHGDITKINEN----DIP--D--QDVL LAGFPCQPF SNIGK-
ECVLSSEIDKK--ACE---TYALNFK----EEPQGDITHEIT-----SFP--E--FDFL LAGFPCQPF SYAGK-
ETVWANEYDKN----AAITYQSNFKN----KLIIDDIRNIKVE----DVP--D--FDVL LSGFPCTSFSVAGY-
VCVASAEIDQQ----AIKVYRONW----PTDGVVDHNLGDIT--AIQQLP A---HDVL VGGVPCQPWSIAGK-
TPVAFVEINED----AQKFLKTKFS----DASVFNDVTKFTKS----DFPED--IDMI TAGFPC TGFSIAGS-
ECVYSNEWDKY--AQE---VYEMNFG----EKPEGDITQVNEK----TIP--D--HDIL CAGFPCQA FSI SGK-
VCEYAVEIDRE----AAAVYERNWN----KPALGDITDDAND--EGVTLRGYDGP IDVL TGGFPCQPF SKSGA-
NVVFSSEWDKF--AQK---TYHANYG----DFPDGDITKIDEK----DIP--D--HEILVGGFPCVA F SQAGL-
EHAWANDIDEW----ACETFRTNICPDRPD SVVCGD VRELDIKSLGEEKFG----EIDAF TFGFPCNDYSIVGE-
    
```

```

FEI--PAANEYD----K----TIWATFKANHPKT----HLIEGDIRKI-----KE-----E-D----F---
PVSNGYWKRRKGD----D----ELKIIYNAIKLSEK--EGNIFDIRDL-----YK-----R-T----L---
----FENRADKLGQ----KLKDMYIANKLNK----NFGDIRSI-----DP-----K-K---L---
VKS--VFSSEID----K----FAIKTYKANFGD----E-PHGDITKI-----DE-----K-D---I---
LST--YGAVEID----K----NAAETLRINRPKW--KVIENDIEFI-----AD-----NLDEFI----D---
FDI--TWANQWE----PSRKIQHAFDCYSKRFKNGI--H-SNKDIAQV-----SD-----E-E---M---
VRC--VFSSEID----K----YAVQTYQANHGGE--T-VCGDITQT-----DV-----A-D---I---
FSH--VALIEIE----P----SACQTLRLNRPDW--NVIEGDVRLF-----QG-----E-G---Y---
GKC--VFSSEID----P----FAKFTYNTNEGV--V-PFGDITKV-----EA-----T-T---I---
FNI--VFANDNW----K----GCWKTFEKNHGI--KINKKPIEWL-----KP-----S-E---I---
FEI--CAAFENW----E----KAIEIYKNNFSH--PIYNIDL RNE-----KE-----AV-E-K--IKK--
FRI--ICANEYD----K----SIWKTYESNHS A--KLIKGDISKI-----SS-----D-E---F---
    
```

3. Выравнивание последовательностей по пространственной структуре

Для этого, как минимум, пространственные
структуры сравниваемых последовательностей
должны быть решены

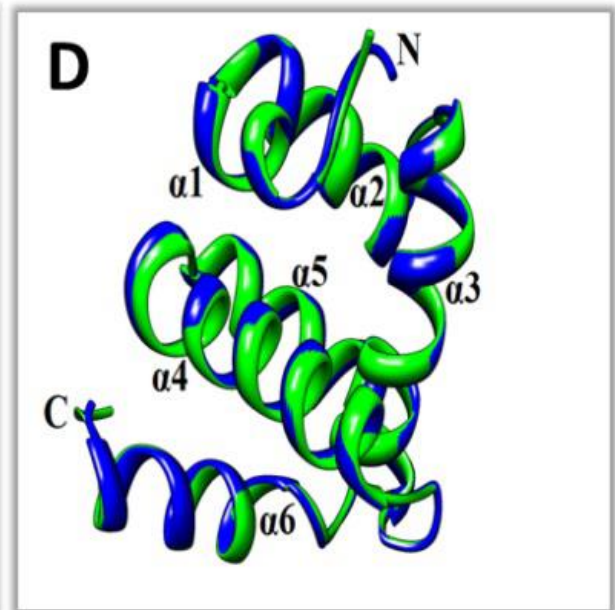
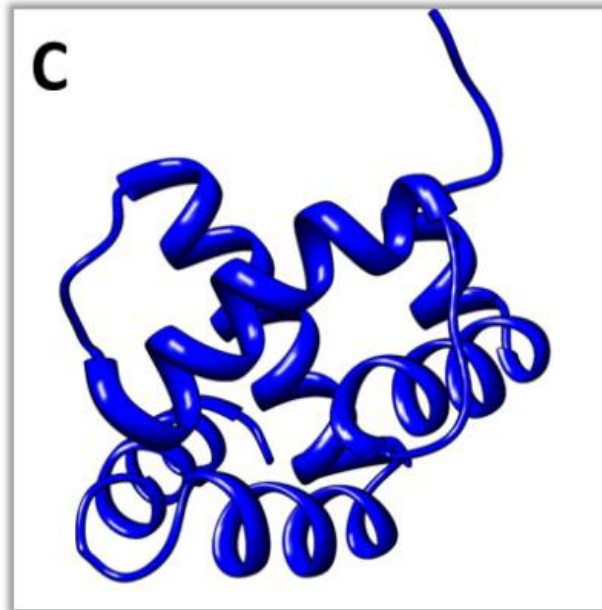
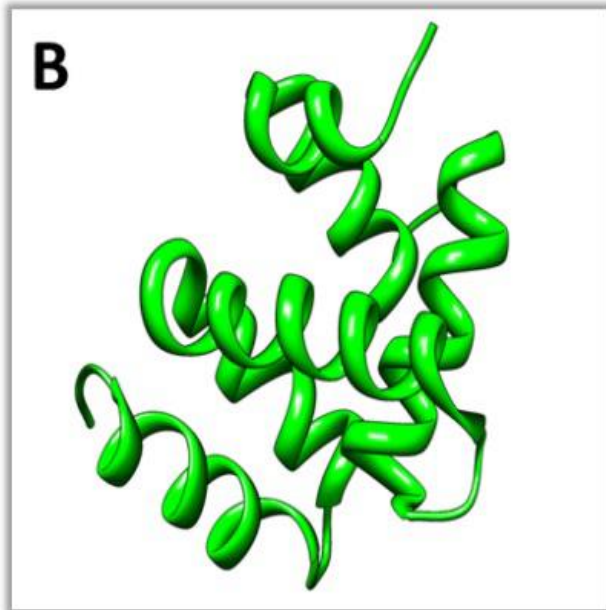
Выравнивание, построенное по совмещению полипептидных цепей

A

```

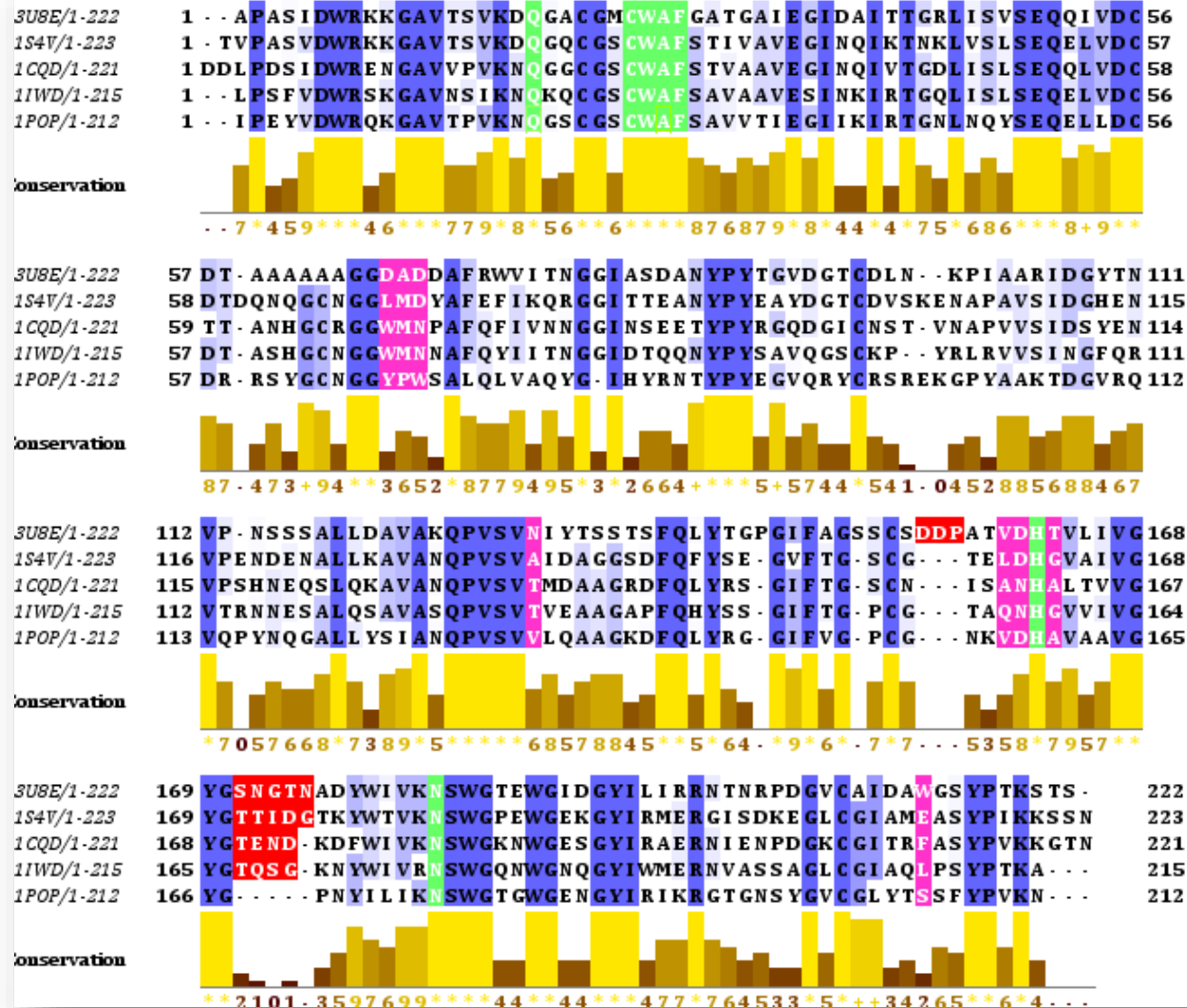
      *           20           *           40           *           60           *
MsepCSP5_ : MNSFTVLCFLFAIVALAVARPDGKYTD RYDSV NLDQILSNRRLIVPYIKMLDQ GKCTPDGKELKFT HIREA : 70
MbraCSFA6 : -----EDKYTDKYDNINLDEILANKRLIVAYVNCVMER GKCSPEGKELKEHLQDA : 50
              KYTD YD  NLD IL N  RLIV Y  C    GKC P GKELK H  A

      80           *           100          *           120           *
MsepCSP5_ : LEQDCAKCTKAQRDGT RQVMGHLINHEVDYWNELKAKYDFK NLYSTHHEQELRKLKQ----- : 127
MbraCSFA6 : IENGCKKCTENQEK GAYRVIEHLIKNEIEIWR ELTAKYDF TGNWRKHYEDRAKAAGIV IPEE : 112
              E  C KCT  Q  G  V  HLI  E  W  EL  AKYDF  K  E
    
```



Younas et al., Int J Biol Sci. 2018 A chemosensory protein MsepCSP5 involved in chemoreception of oriental armyworm *Mythimna separata*

Multiple Sequence Alignment of CsCP against Plant CPs



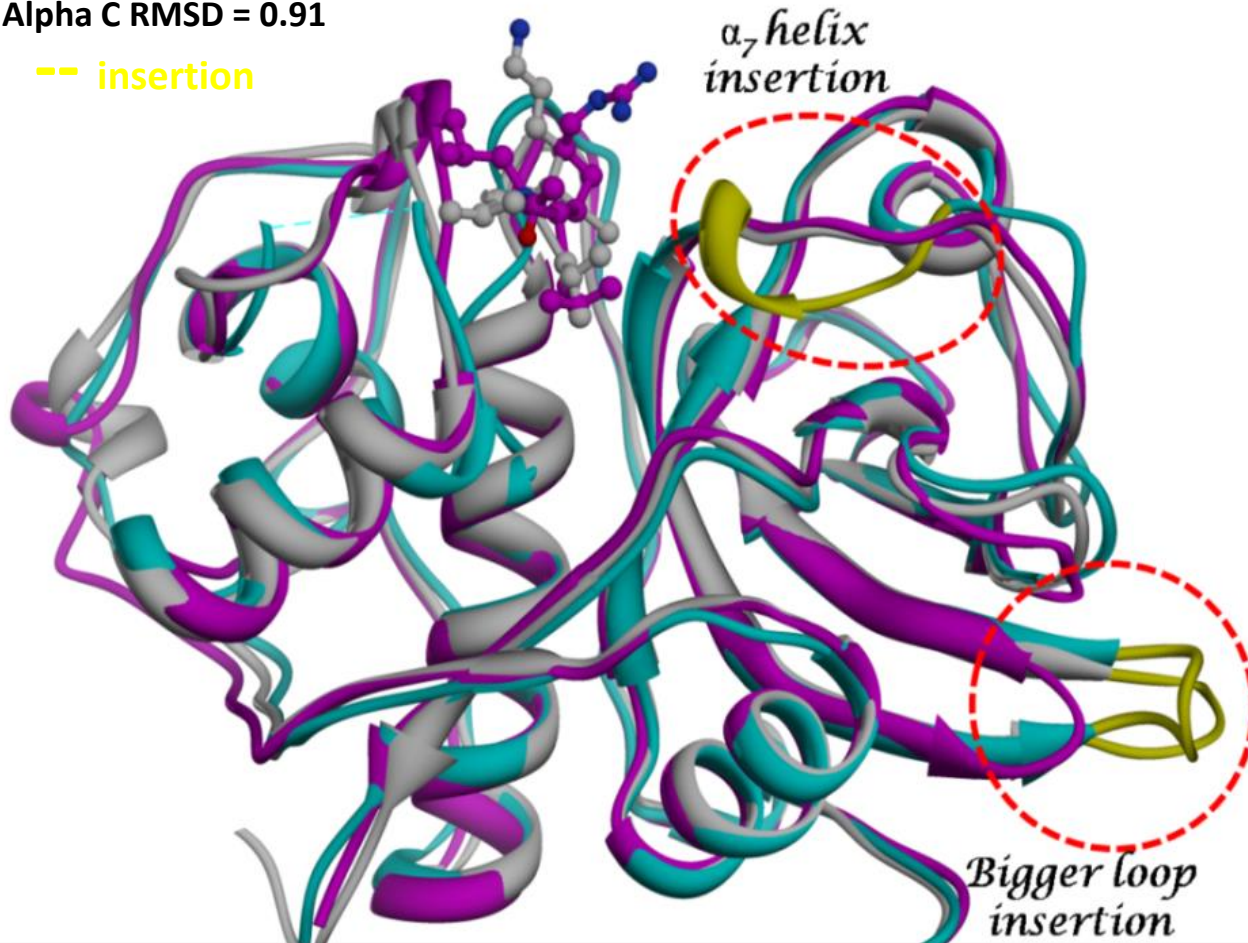
-- Conserved
 -- Catalytic
 -- Active site

-- 3U8E (*Crocus sativus*)
 -- 1S4V (*Ricinus communis*)
 -- 1CQD (*Zingiber officinale*)
 -- 1IWD (*Ervatamin B*)
 -- 1POP (*Carica papaya*)

Conserved Structure of CsCP Superimposition with Plant CPs

Alpha C RMSD = 0.91

-- insertion

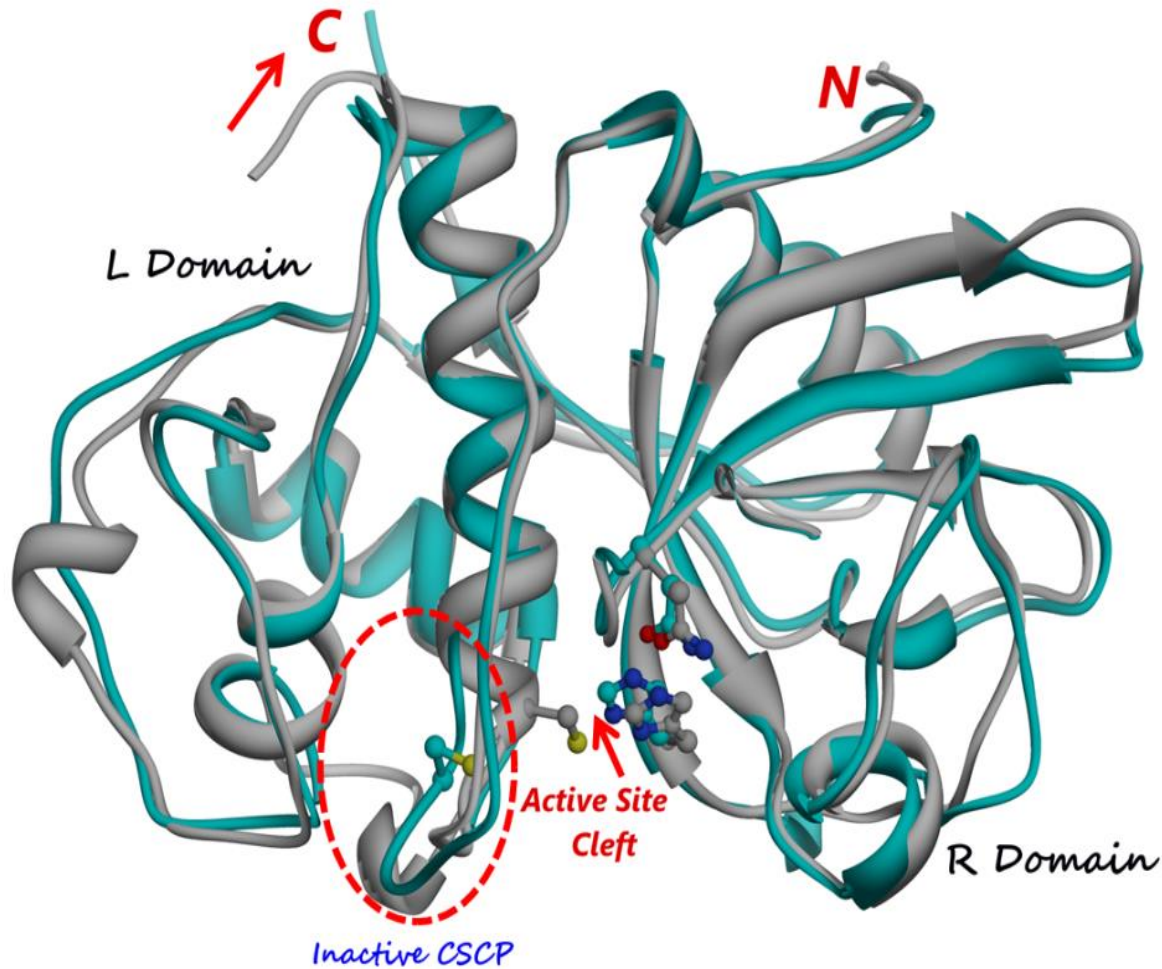


-- 3U8E (*Crocus sativus*)

-- 1S4V (*Ricinus communis*)

-- 1POP (*Carica papaya*)

Overall Topology of CsCP



-- Inactive Form

-- Active Form

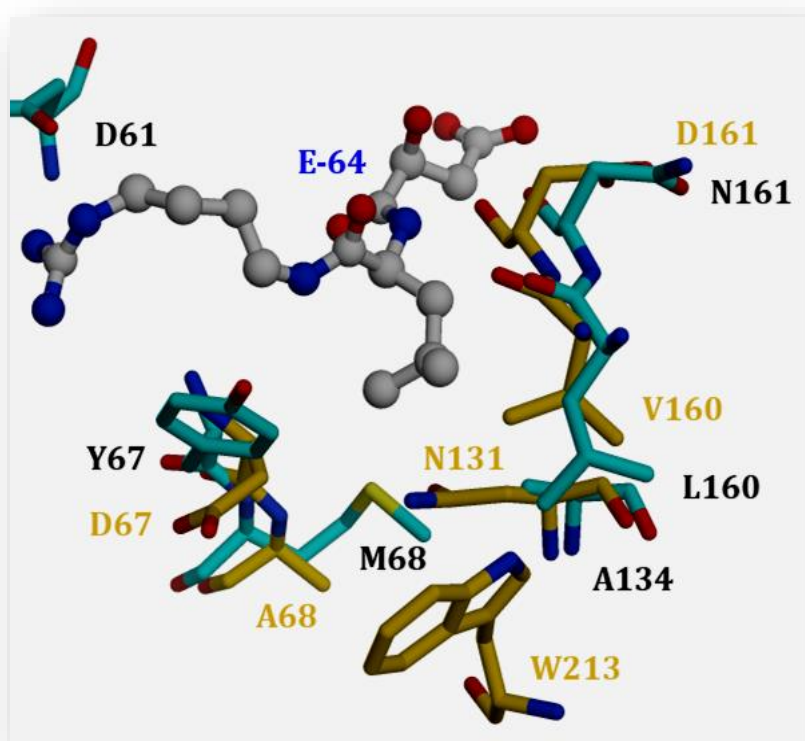
Catalytic Triad:
C25, H162, and N183
(Active Site Cleft)

Inactive Form:
C25 bridged with C22

Active Form:
C22 bridged with C64
C25 is free

KEY Findings Identification of S₂ Pocket Residues of CSCP

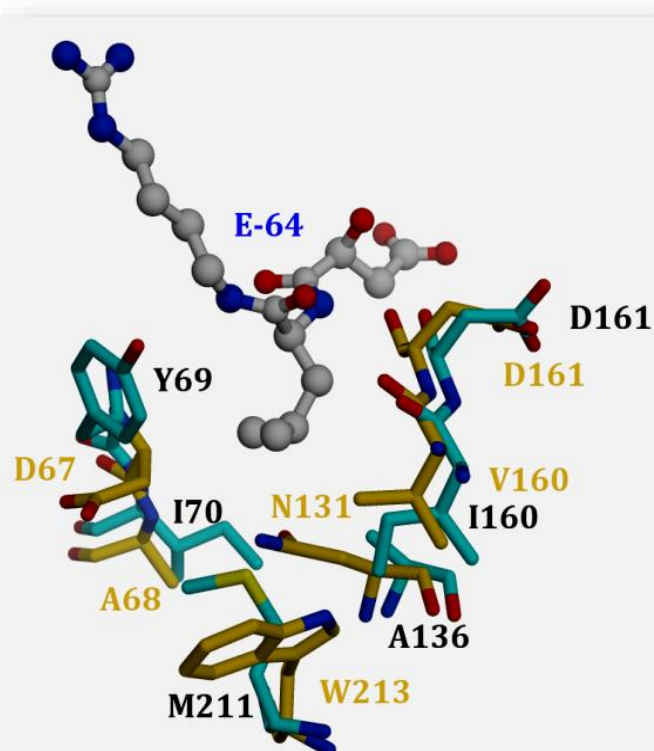
Cathepsin K-E64



-- Others

-- CsCP

Actinidin-E64



Конец презентации

Задание для работы в классе.

Дано два выравнивания pf00145_seed.fasta и pf00145_seed-tcoffee.fasta одних и тех же 31 последовательностей

Найти

- (i) блок одинаково выровненных колонок в двух выравниваниях
- (ii) участок, на котором выравнивания полностью различны
- (iii) Результаты сохранить в форме

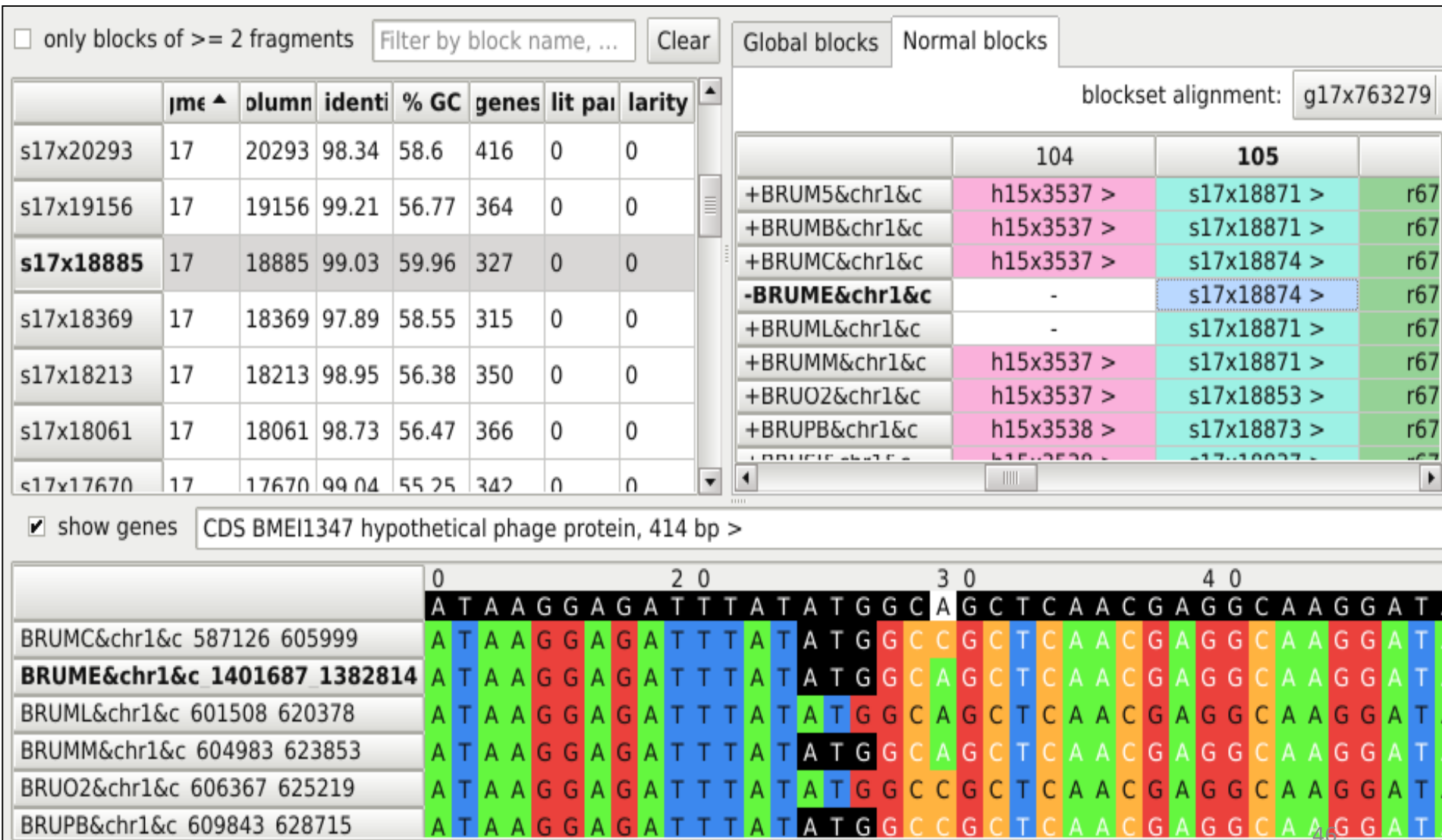
4. Проект блочного MSA

- Определение. MSA = набор блоков достоверного выравнивания во всех или части последовательностей. Аал
- Этап I: во входном выравнивании найти достоверные блоки специфичные для ветвей входного филогенетического дерева
- Этап II: во входном выравнивании найти все блоки достоверного выравнивания.
- Этап III: найти все блоки достоверного выравнивания во входном множестве последовательностей
- Этап IV: создать сервис для блочного MSA

Пример: блочное выравнивание геномов

Борис Нагаев NPGexplorer

Визуализация выравнивания 17 геномов бруцелл (рис. из работы К.Худяковой)



Каждая ДНК представляется последовательностью блоков

Global blocks Normal blocks

blockset alignment:

	35	36	37	38	39
+BRUA1&chr1&c	s17x21286 >	-	s17x10111 >	-	s17x7319 >
+BRUA2&chr1&c	s17x21286 >	-	s17x10111 >	-	s17x7319 >
+BRUA8&chr1&c	s17x21286 >	-	s17x10111 >	-	s17x7319 >
-BRUAO&chr1&c	s17x21286 >	-	s17x10111 >	-	s17x7319 >
+BRUC2&chr1&c	s17x21286 >	h13x107 >	s17x10111 >	h6x2653 >	s17x7319 >
-BRUCA&chr1&c	s17x21286 >	h13x107 >	s17x10111 >	h6x2653 >	s17x7319 >
+BRUM5&chr1&c	s17x21286 >	h13x107 >	s17x10111 >	-	s17x7319 >
+BRUMB&chr1&c	s17x21286 >	h13x107 >	s17x10111 >	-	s17x7319 >
+BRUMC&chr1&c	s17x21286 >	h13x107 >	s17x10111 >	h6x2653 >	s17x7319 >
-BRUME&chr1&c	s17x21286 >	h13x107 >	s17x10111 >	-	s17x7319 >
+BRUML&chr1&c	s17x21286 >	h13x107 >	s17x10111 >	-	s17x7319 >
+BRUMM&chr1&c	s17x21286 >	h13x107 >	s17x10111 >	-	s17x7319 >
+BRUO2&chr1&c	s17x21286 >	h13x107 >	s17x10111 >	-	s17x7319 >
+BRUPB&chr1&c	s17x21286 >	h13x107 >	s17x10111 >	-	s17x7319 >
+BRUSI&chr1&c	s17x21286 >	h13x107 >	s17x10111 >	h6x2653 >	s17x7319 >
+BRUSS&chr1&c	s17x21286 >	h13x107 >	s17x10111 >	h6x2653 >	s17x7319 >
+BRUSU&chr1&c	s17x21286 >	h13x107 >	s17x10111 >	h6x2653 >	s17x7319 >

Про типы блоков следующий слайд

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 : YQVFFIINDDDVALALEIGADGHIHVGQNDDEEIRQVIASCAGKMKIIGLSVHVSVEAEEAERLGSVDYIIGVGPPIFPTISKADA : 14:
THEIE_ANOFW : YNIFFIVNDDVDLALALQADGVHVGQDEVAERVRDRIGDKY-LGVSVHNLNEVKKAL-AACADYVGLGPIFPTVSKEDA : 14:
THEIE_GEOTN : YGVFFIVNDDVELAIAIDADGVHVGQDDEADARRVREKIGDKI-LGVSAHNVVEEARAAI-EAGADYIIGVGPPIYPTRSKDDA : 14
THEIE_BACSU : AGVFFIVNDDVELALNLKADGHIHGQEDANAEREVRAAIGDMI-LGVSAMTSEVVKQAE-EDGADYVGLGPIYPTETKKDT : 14
THEIE_BACA2 : AGIFFIINDDDVELALRLEADGVHIGQDDADAEEETRAAIGDMI-LGVSAMTSEVVKRAE-AAGADYVGMGPVYPTETKKDA : 14
THEIE_OCEIH : FQIFFIINDDDVDLAKQLDADGHIHGQDDQPVVVRKQFENKI-IGLSISTNNELNQSP-LDLVDYIIGVGPPIFDTNTKEDA : 14
THEIE_STAAB : YNVFFIVNDDVSLAKEINADGHIHVGQDDAKVKEIAQYFTDKI-IGLSISDLGEYAKSD-LTHVDYIIGVGPPIYPTPSKHDA : 14:
THEIE_STACT : YNVFFIVNDDVALAEEIDADGHIHVGQDDEAVDDFNRRFEGKI-IGLSIGNLEELNASD-LTYVDYIIGVGPPIFATPSKDDA : 14:
      6pFI61DD6 La 6 ADG6H6GQ D 6G68 2 DY G6GP pT 3K Da
```

```
      *           180           *           200           *           220           *           240
THEIE_LACLS : AKETGKDLR-LMLLENQLPIVGIIGGITQDSLTELSAIGLDGLAVISLLTEAENPKKVAQMIRQKITKNG~~~~~ : 21
THEIE_MANSM : SPIVGMNFIRQIRQLGIDKPCVAIGGITKEESAAILRRLGADGVAVISAIHSHSVNIANTVKTLAQK~~~~~ : 22
THEIE_STRA3 : KPAVGNRTTKAVREINQDIPVVAIGGITSDVFVDIIESEGADGLAVISAIISKANHIVDATRQLRYEVEKALVNRQKRSDVI : 22:
THEIE_LISIN : EPVSGTAILEIRRAGIKLPIVGIIGGITNETNSAEVLTAGADGVSVISAITRSEDCQSVIKQLKNPGSPS~~~~~ : 21
THEIE_ANOFW : KQACGLTMEHIRAEKRVPLVAIGGITETQAKQVIEAGADGLAVISAIICRAEHIYEQTKRLYEMVMRAKQKQKDR~~~~~ : 21
THEIE_GEOTN : NEAQQPGILRHLRREQGITIPIVVAIGGITADNTRAVIEAGADGVSVISAIASAPEPKAAAAALATAVREANL---R~~~~~ : 22
THEIE_BACSU : RAVQGVSLIEAVRRQGISIPIVGIIGGITIDNAAPVIEAGADGVSMISAIISQAEDPESAARKFREEIQTYKTG--R~~~~~ : 22:
THEIE_BACA2 : EAVQGVTLIEEVRQGITIPIVGIIGGITADNAAPVIEAGADGVSMISAIISQAEDPKAAARKFSEEIRRSKAGLSR~~~~~ : 22
THEIE_OCEIH : KTAVGLEWISLKKQHPSLPLVVAIGGITNTNAQEIIQAGADGVSVISAITETDHIHQAVQRL~~~~~ : 20
THEIE_STAAB : HTPVGPMEIATFKEMNPQLPIVVAIGGITSNVAPIVEAGANGISVISAIISKSENIIEKTVNRFKDFFN~~~~~ : 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.

ДВА ДОМЕНА гомеобелков: гомеодомен и OAR домен

SW: PMX1_CHICK/1	-----MASSYAHAMERQALLPARLDGPACLDNLQAKNFVSVSHLLDLEEAG-DMVAAQDCEGGGPRGRSLLLESP-GLTSGSDTPQQD	: 80
SW: PMX2_HUMAN/1	-----MDSAAAAFALDKPALGCPGPPPPALGCPGDCQAQRNFVSVSHLLDLEEVAAAGRLAARPGARABAREGAAREPSCGSSGSEAAFPD	: 86
SW: PMX1_HUMAN/1	-----MTSSYGHVLEQPALGCRRLDSPGLMDTLQAKNFVSVSHLLDLEEAG-DMVAAQADENVGACGRSLLLESP-GLTSGSDTPQQD	: 80
SW: ARX_BRARE/1	ISQAPQVVISRSKSYREN-APFSQS---D-EGQSP--EHAQELVELST-----LKFEEDEVVKEEACQDN-----S-----	: 84
SW: ARX_MOUSE/1	ISQAPQVVISRSKSYRENGCAPVPPPPALD-ELSGPCGVHPEERLSAASGPGSAPAAGCGTCAEDDEEELLEDEEEDEREELLEDDDEELLEDDARALLKEPERRCVATTCTVAIAAAAAAAAAAVATEGGELSPKRELLLHPEDARCKDCGDSVCLS	: 157
SW: AL_DROME/1-1	-----MGISEEIKLEELPQAKLAHPDAVVLVDRAPGSSAASAGAAALTVSMVSYSGGAPSCASGASGCTNSPVSDGNS	: 72
SW: ALX4_MOUSE/1	-TFLSAGAKQCPCDAKSRARYGACQDLAAPLESSSGARGSPNKFQPPPTQP-----PPAPPAPPAHLYLQRCACKTPDPCSLKLBQEGSSGCHNAALQVPCYAKRESNLCEPELPPDSFVPCVMDNSYLSVKRTGARCPQDASAEIIPSL	: 145
SW: ALX4_HUMAN/1	-TFLSAAAKAQCPCDAKSRARYGACQDLATPLESSGARGSPNKFQPPPTQPQPSPQPQQQPPQPQPPAHPHLYLQRCACKTPDPCSLKLBQEGSSGCHNAALQVPCYAKRESNLCEPELPPDSFVPCVMDNSYLSVKREAGVRCQDASSDLPSP	: 157
SW: RX2_CHICK/1	-----NPSRLHSIEAILGFTKDDGLLGFQFP-----DGCAGSAKEAADKRGPRHCLPKGPAEPPPAEHQGRFQEPYPCGASAPF-----LPAGCGGC	: 83
SW: RX2_BRARE/1	-----GISCRVHSIDVILGFSKDDPPLLEPSGR-----HKVDLEDQLEEQEKQVADPYSHLQIPDQIQQQQSVYH---DTGLFSTDKCADLGDPRSINVEDSRS	: 92
SW: RX1_XENLA/1	-----NPSRLHSIEAILGFKEDS-VLGSFQSEIISPRNAKEVDKRSRHLHMTTEIHPQEHLEDG-QADCYG--DPYSGRTSSECLP-PCLS--SNSDN	: 91
SW: RX_HUMAN/1-1	-----STSRLHSIEAILGFTKDDG-ILGTFPAERGARGAKEERRLGARCPACPKAPEGSEEPSPPAPAPAPYEAPRPPYCPKPEWRAPSPGLPVGPAATGEA	: 97
SW: PIX2_BRARE/1	-----MTSHKDLPLSLDHHHHHHVTCGKHAFLSMASLLQLRQSVDSKHLRDLVHTVSDTSSPEVKEKRCQ--	: 66
SW: PIX2_HUMAN/1	-----METNCRKLVSAVGLQVPAEVECLFSKDSSEIKKVFETDPSRKRKAASAKFPFPHQPCANEKRSKQ--	: 68
SW: PIX1_HUMAN/1	-----MDAFKCGMSLERLPEGFPPPPPHDMGPAFHLLARPADPREPLEN-SASESSDTLPEKREKGGEP	: 64
SW: OTP_MOUSE/1	-----MLSHADLLDARLCHKDAEALLGHREAVKRLGVGCSDPGCHPCDLAPNSDPVEGATLLPREDITTVGSPASLAVSAKDPDKQPGPQCGP	: 90
SW: PMX1_CHICK/1	NDQLNSEE-----KKKRRRRRRTFTTSSQLQALERVPERTHYPDAFVRBDLARRVNLTEARVQVVFQMRRAKFRRNEEAMLSKMASLLKSYSGDVTAVEQIIVPRPAPRPTDYLWGTASPYSAMATYSTTCTMAS-----	: 213
SW: PMX2_HUMAN/1	GCPCSPGRCG-----AAKRRKRRRRTFTTSSQLQALERVPERTHYPDAFVRBDELARRVNLTEARVQVVFQMRRAKFRRNEEAMLSRSASLLKSYSGE-AAIEQVPAPRPTALSPDYLWGTASSPYSTVPPYSPGSSGP-----	: 221
SW: PMX1_HUMAN/1	NDQLNSEE-----KKKRRRRRRTFTTSSQLQALERVPERTHYPDAFVRBDLARRVNLTEARVQVVFQMRRAKFRRNEEAMLANKMASLLKSYSGDVTAVEQIIVPRPAPRPTDYLWGTASPYSAMATYSATCANNS-----	: 213
SW: ARX_BRARE/1	AGSDSEEG-----MLKRRRRTFTTFTSYQLEELERAFQTHYPPDVFTRBELAMRLDTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPPSCAPLG	: 230
SW: ARX_MOUSE/1	AGSDSEEG-----LLKRRRRTFTTFTSYQLEELERAFQTHYPPDVFTRBELAMRLDTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPPSCAPLG	: 303
SW: AL_DROME/1-1	EKADSEY-----PKRRRRTFTTFTSYQLEELERAFQTHYPPDVFTRBELAMRLDTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPPSCAPLG	: 212
SW: ALX4_MOUSE/1	EKTDSESN-----KCKRRRRTFTTFTSYQLEELERAFQTHYPPDVFTRBELAMRLDTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPPSCAPLG	: 290
SW: ALX4_HUMAN/1	EKADSESN-----KCKRRRRTFTTFTSYQLEELERAFQTHYPPDVFTRBELAMRLDTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPPSCAPLG	: 302
SW: RX2_CHICK/1	KPSDEEQ-----PKRRRRTFTTFTSYQLEELERAFQTHYPPDVFTRBELAMRLDTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPPSCAPLG	: 215
SW: RX2_BRARE/1	PDIPDEDQ-----PKRRRRTFTTFTSYQLEELERAFQTHYPPDVFTRBELAMRLDTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPPSCAPLG	: 225
SW: RX1_XENLA/1	KLSDDEEQ-----PKRRRRTFTTFTSYQLEELERAFQTHYPPDVFTRBELAMRLDTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPPSCAPLG	: 224
SW: RX_HUMAN/1-1	KLSEEEQ-----PKRRRRTFTTFTSYQLEELERAFQTHYPPDVFTRBELAMRLDTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPPSCAPLG	: 242
SW: PIX2_BRARE/1	SKNEDSW-----DDPSKRRRRTFTTFTSYQLEELERAFQTHYPPDVFTRBELAMRLDTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPPSCAPLG	: 212
SW: PIX2_HUMAN/1	GRNEDVGA-----EDPSKRRRRTFTTFTSYQLEELERAFQTHYPPDVFTRBELAMRLDTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPPSCAPLG	: 215
SW: PIX1_HUMAN/1	KCPEDSCAGCTGCCGADDPAKRRRRTFTTFTSYQLEELERAFQTHYPPDVFTRBELAMRLDTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPPSCAPLG	: 218
SW: OTP_MOUSE/1	NPSQACQQ-----CQQQKRRRRTFTTFTSYQLEELERAFQTHYPPDVFTRBELAMRLDTEARVQVVFQMRRAKWRREERCAQTHPPGLPFPFPCPLSATHPLSPYLDASFPFPHHPALDSAWTAAAAAAAFPSLPPPPG-SASLPPSCAPLG	: 236
	k 4 4R RT Ft QL EL E R F 4 HYPD RE 6A L E R6qVWFQNRRAK54 e4	
SW: PMX1_CHICK/1	-----PAQGMNMANSLALPLAKRHSYSLQRNQVPTVN-----	: 245
SW: PMX2_HUMAN/1	-----ATPCVMNMANSLALPLAKRHSYSLQRNQVPTVN-----	: 253
SW: PMX1_HUMAN/1	-----PAQGINMANSLALPLAKRHSYSLQRNQVPTVN-----	: 245
SW: ARX_BRARE/1	LGTFLGTAAMFRHFAFIPTFCRLFFSSMCLPTSASTAAALLRQTAPPVPSVPQSAALPEPPSSSSSTAADRRASSIAALPLAKRHSYA-QLTQLNLIPSGTACKRVC-----	: 336
SW: ARX_MOUSE/1	LSTFLGAAVFRHFAFISPAFCRLFFSTMAPLTSASTAAALLRQTAPPVGAASGALADP-----ATAAADRRASSIAALPLAKRHSYAQLTQLNLIPGCTCKRVC-----	: 404
SW: AL_DROME/1-1	PPTSPASGHAXPQVLVGIATLQQAASSLPT---QTSFVALTSHSPQRQLPPSHQAPPPPPRAATPPEDRRTSSIAALPLAKRHSYSLQRNQVPTVN-----VS	: 313
SW: ALX4_MOUSE/1	DFL-----SVSGACSHVQTHMCSLFGAACISPLNGCYELNCGPDRKTSIAALPLAKRHSYAASISWAT-----	: 354
SW: ALX4_HUMAN/1	DFL-----SVSGACSHVQTHMCSLFGAACISPLNGCYELNCGPDRKTSIAALPLAKRHSYAASISWAT-----	: 366
SW: RX2_CHICK/1	LPASYTTPPFL-----NSPVTGHALQPLGAMGPPPPYQCGAFAVDFKFLDEGDPNRTSSIAALPLAKRHSYSLQRNQVPTVN-----	: 290
SW: RX2_BRARE/1	LQPTYTAHPCFL-----NTSPGMHNQIQPM---PPPPYQVFPVNDKYLEEDV-SSSIAALPLAKRHSYSLQRNQVPTVN-----	: 297
SW: RX1_XENLA/1	LPGSYTPPPFI-----NPVSVGHALQPLGAMGPPPPYQCGAFAVDFKFLDEEDPDMSSIAALPLAKRHSYSLQRNQVPTVN-----	: 296
SW: RX_HUMAN/1-1	LPASYTTPPPPPFL-----NSPFLGCPQLPL---APPPSYPCGCFGDKRFLDEADPNSSIAALPLAKRHSYSLQRNQVPTVN-----	: 319
SW: PIX2_BRARE/1	SISMSMSMSMVPASVTCVPGSSL-----NSLNNLNLNSPNSLNSAVTTPACPYAPPTPPY-VYRDTCNSSLASLPLAKRHSYSLQRNQVPTVN-----	: 314
SW: PIX2_HUMAN/1	SISMSMSMSMVPASVTCVPGSSL-----NSLNNLNLNSPNSLNSAVTTPACPYAPPTPPY-VYRDTCNSSLASLPLAKRHSYSLQRNQVPTVN-----	: 317
SW: PIX1_HUMAN/1	SISMTMPSSMCPGAVPGMNSCL-----MNIN---MLTGSSLNSAMSFGACPYCTPASPYVYRDTCNSSLASLPLAKRHSYSLQRNQVPTVN-----	: 314
SW: OTP_MOUSE/1	SQCSLAAGPPPNMCLNSLAGSNGACLG---SHLYQAPFGMVPASLPGSPNSVSGSLQCLSSPDSVDRGTSIAALPLAKRHSYSLQRNQVPTVN-----	: 325

Гомеодомен является ДОМЕНОМ

Доказательство

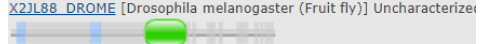
- Выравнивание, со сходимостью, свидетельствующим о гомологичности последовательностей
- Представленность домена в негомологичных белках (обычно, но не обязательно)

```

G4VRE6_SCHMA/203-259
HME2B_DANRE/173-229
HM1N_BOVWQ/373-429
TIEHES_HELRO/41-97
HM05_CAEL/36-92
G3IBX4_CRIGR/25-81
DLX3B_DANRE/136-192
B8A5N9_DANRE/135-191
Q91967_CHICK/77-133
DLX3B_DANRE/126-182
HM04_CAEL/193-159
HM23_CAEL/212-268
MSX3_MOUSE/88-144
HM30_CAEL/96-152
BARH2_RAT/230-286
BARH1_DROME/300-356
BARX2_MOUSE/138-194
DSH_DROME/275-331
H2XU06_CIOIV/470-524
HM19_CAEL/95-151
SLOU_DROME/546-602
F6VWQ6_XENTR/112-168
TIN_DROME/302-358
NXK25_RAT/138-194
H0XK12_OTOGA/100-156
HM09_CAEL/71-127
H2VEX2_TAKRU/135-69
U3K517_FICAL/59-115
TLX3_CHICK/173-229
U3ZQ6_FICAL/136-188
LBX1_MOUSE/126-182
G4VGG4_SCHMA/38-94
BCD_DROME/98-153
BCD_DROME/98-153 (55)
VENTX_HUMAN/92-148
VENT1_XENTR/128-184
Q804C9_XENTR/190-246
K48B21_SOLLCL/24-79
PHO2_YEAST/78-134
WOX9_ARATH/52-113
WOX9_ORYSJ/111-72
WOX2_ORYSJ/24-85
WOX4_ARATH/87-148
WOX1_ARATH/73-134
WOX2_ARATH/11-72
WOX5_ORYSJ/41-102
WUS_SOLLCL/25-85
WOX6_ARATH/58-119
YHP1_YEAST/174-230
YOX1_YEAST/177-233
HARA_DICDI/163-219
PHX1_SCHPO/169-223
CUT_DROME/1746-1802
CUX2_MOUSE/1114-1170
CUX1_MOUSE/1240-1296
Q22810_CAEL/212-268
HBX2_DICDI/486-542

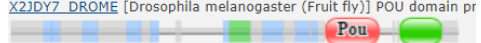
```

There are 25976 sequences with the following architect



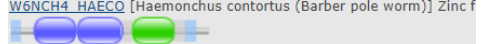
[Show all sequences with this architecture.](#)

There are 2311 sequences with the following architectu



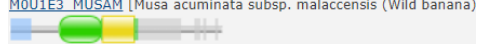
[Show all sequences with this architecture.](#)

There are 2108 sequences with the following architectu



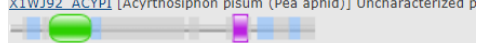
[Show all sequences with this architecture.](#)

There are 1903 sequences with the following architectu



[Show all sequences with this architecture.](#)

There are 1836 sequences with the following architectu



Эволюционные домены

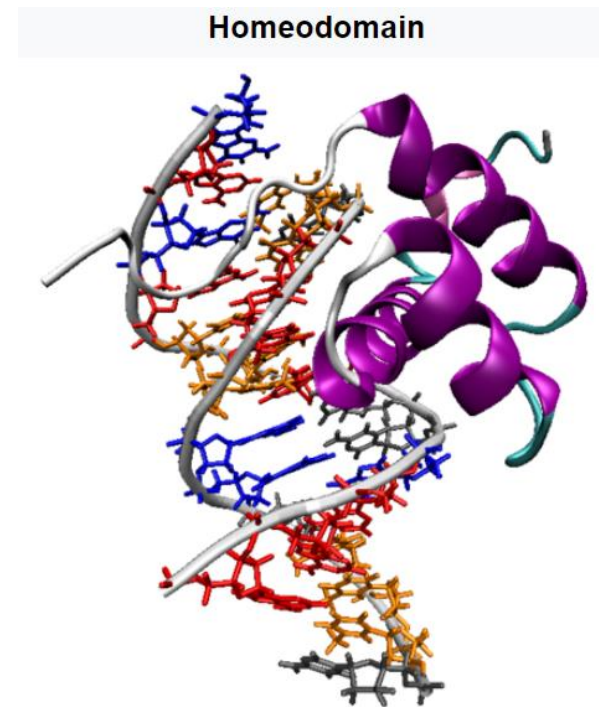
- Имеют определенную функцию (не всегда известна)

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]

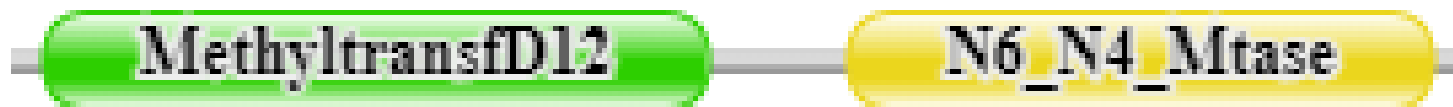


Как выровнять эти две последовательности?

There are 9 sequences with the following architecture:

MethyltransfD12, N6_N4_Mtase

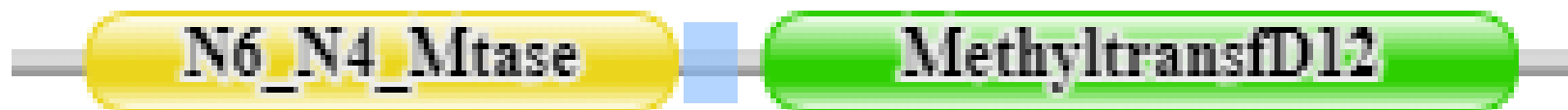
[A0A2Z5QVW5](#) [9MICC](#) [**D12-N6_N4**]



There are 5 sequences with the following architecture:

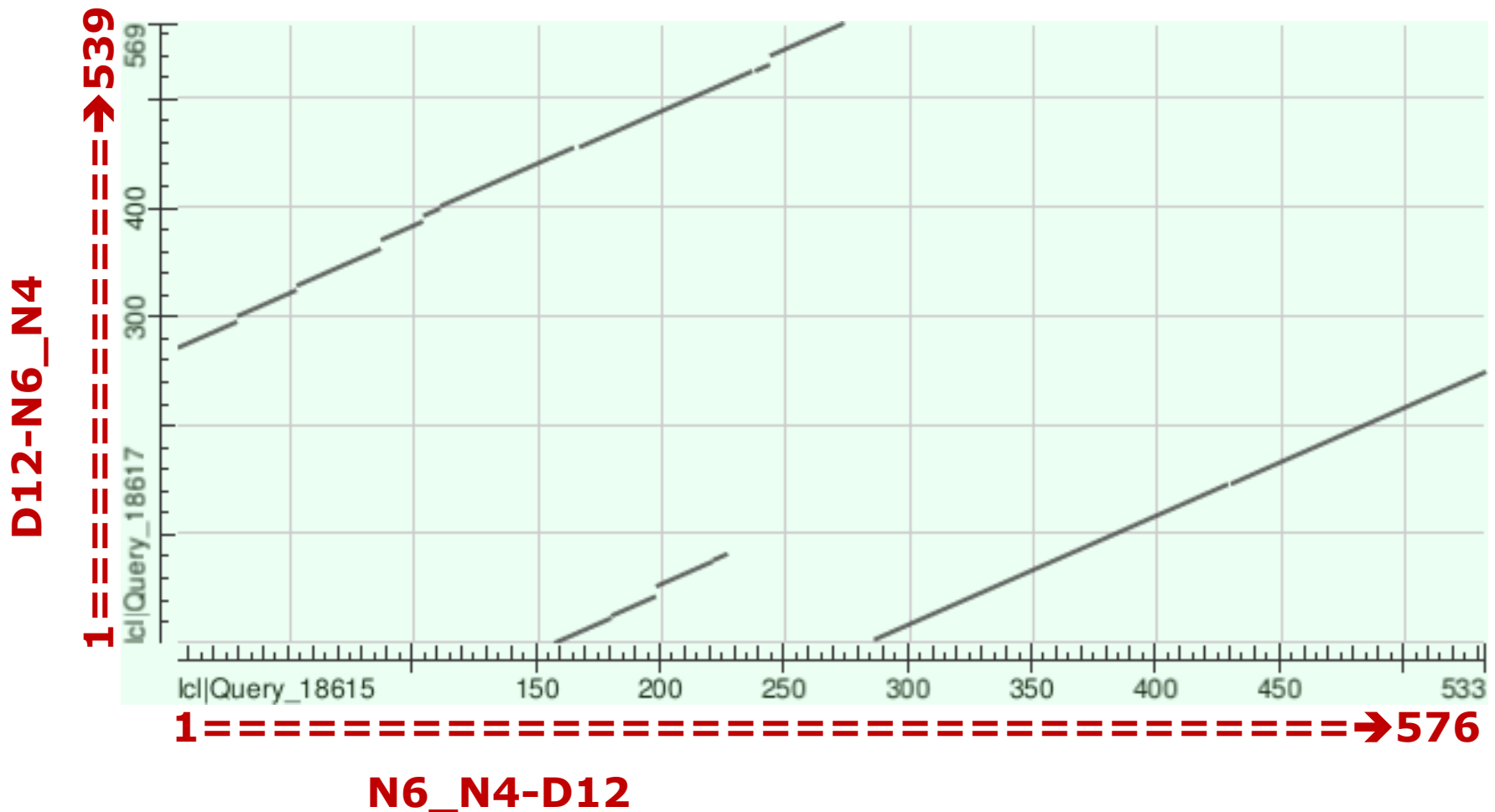
N6_N4_Mtase, MethyltransfD12

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



Как такое может возникнуть?

Лучшее парное выравнивание:
алгоритм множественных локальных
выравниваний.

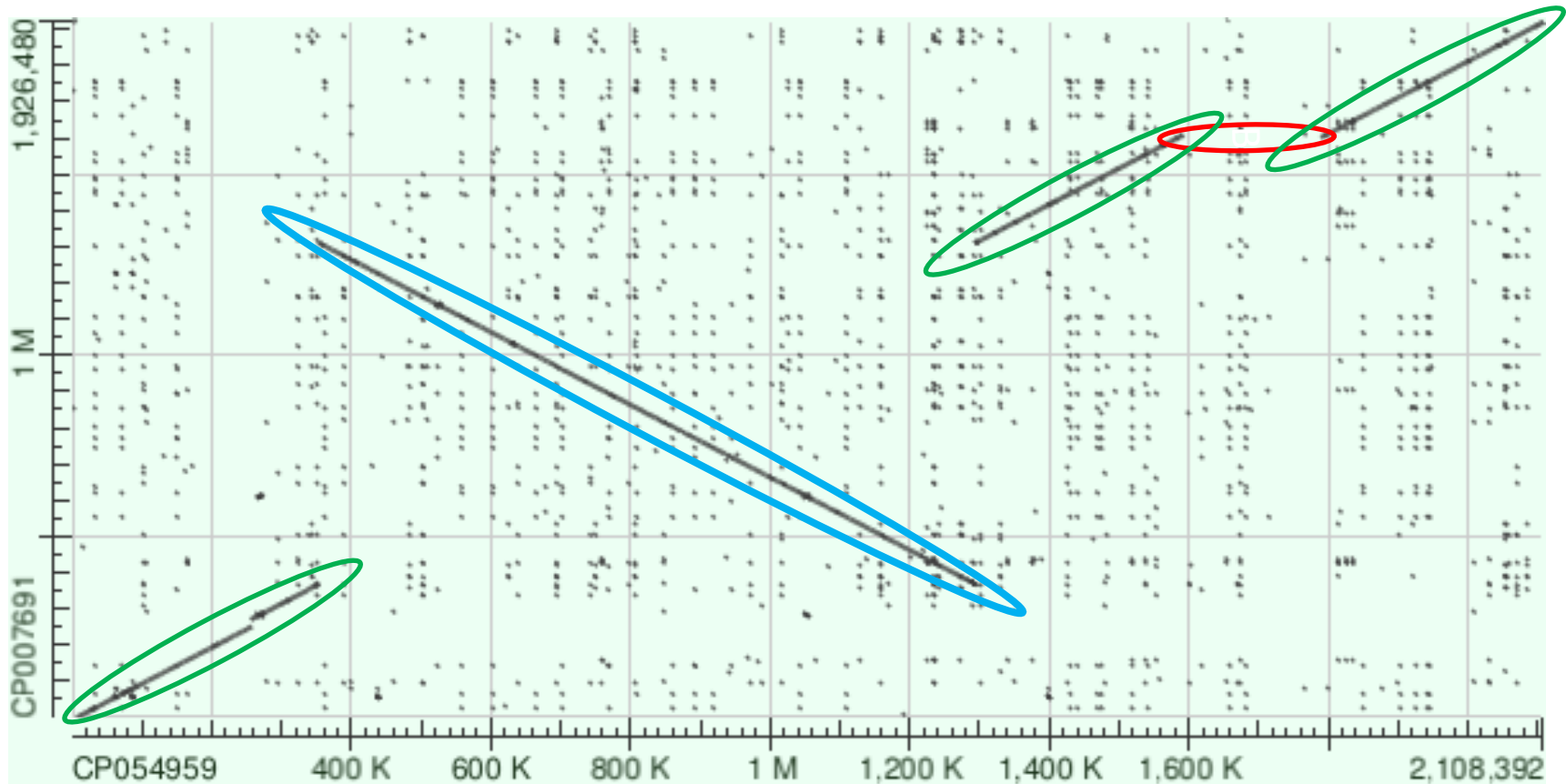


Эволюция геномов бактерий

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

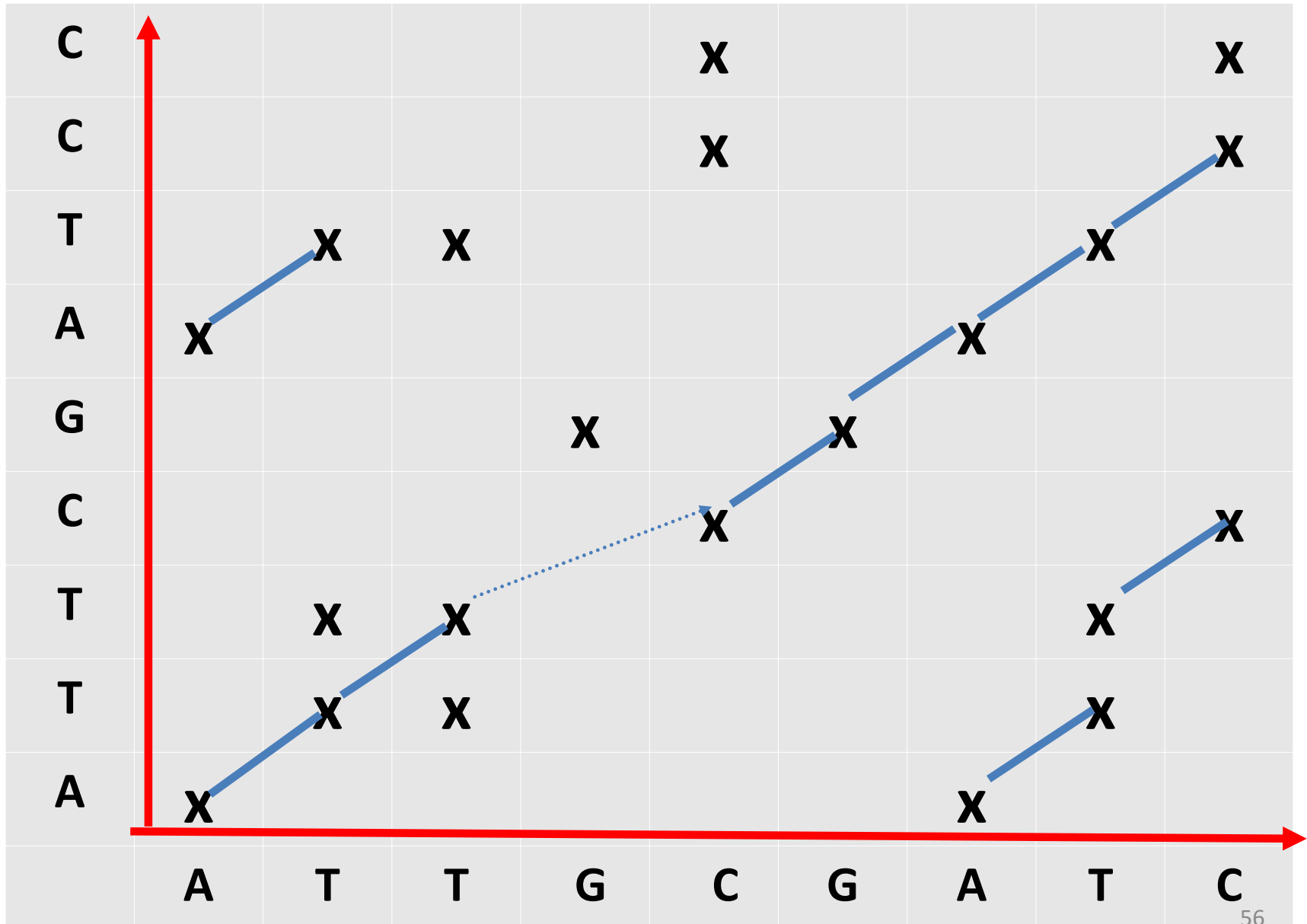


Сходство прямых цепочек

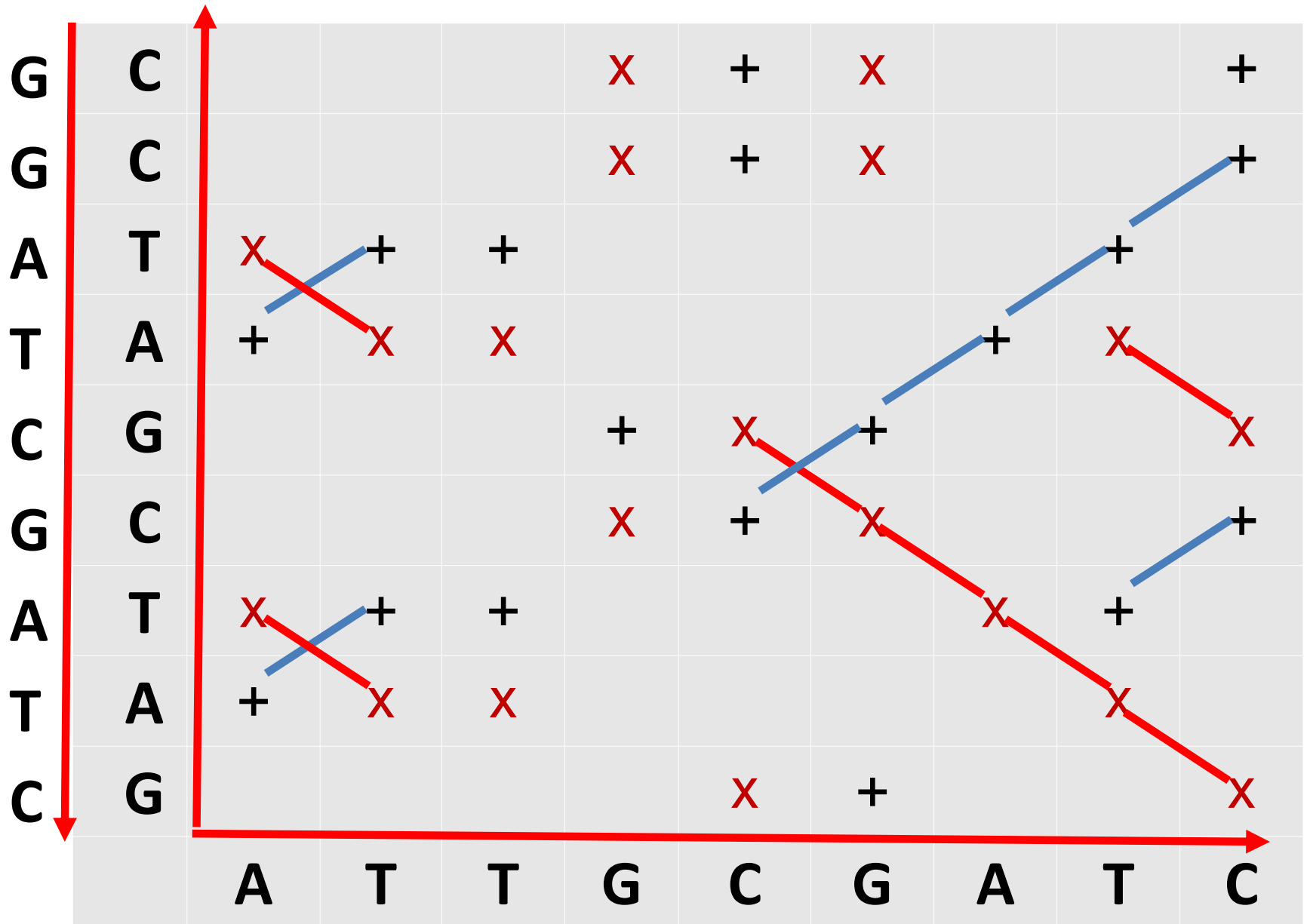
Инверсия - сходство прямой цепочки
с комплементарной второго генома

Делеция / вставка

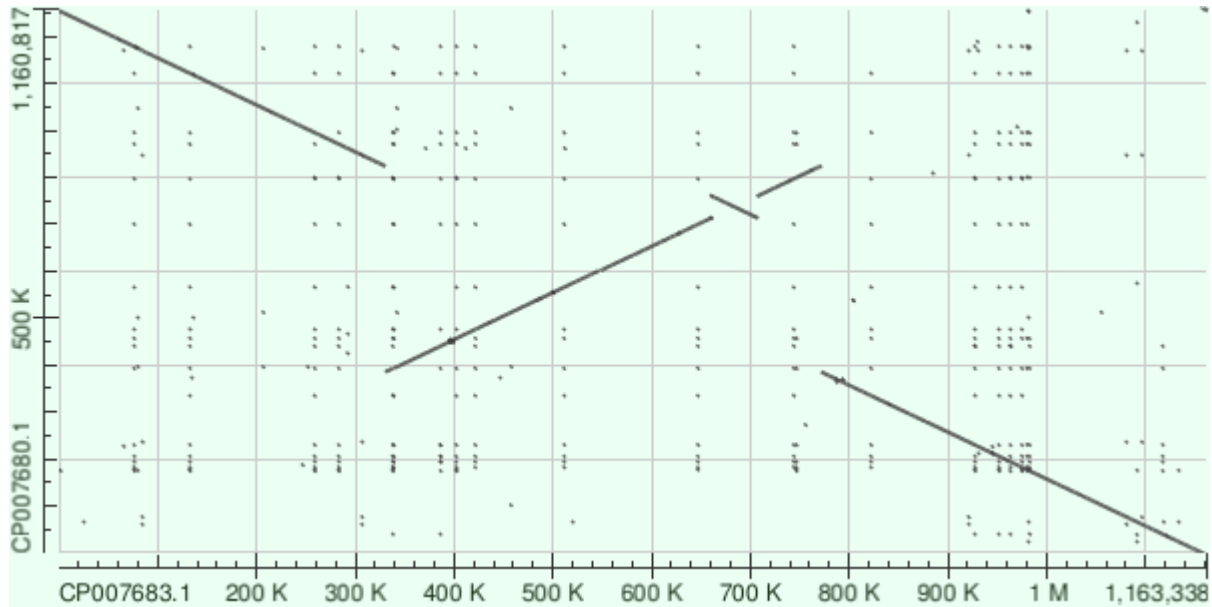
Объяснение Карты локального сходства



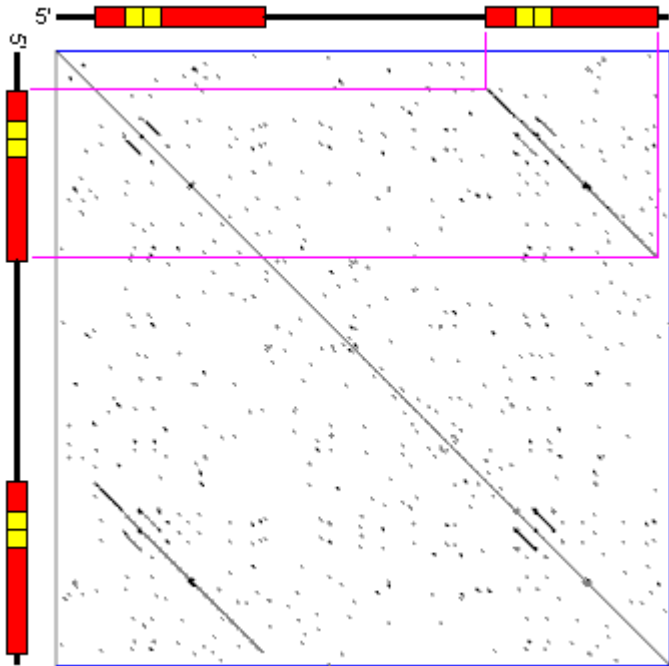
Карта сходства с учетом комплементарной цепочки



Что видим на этой карте?



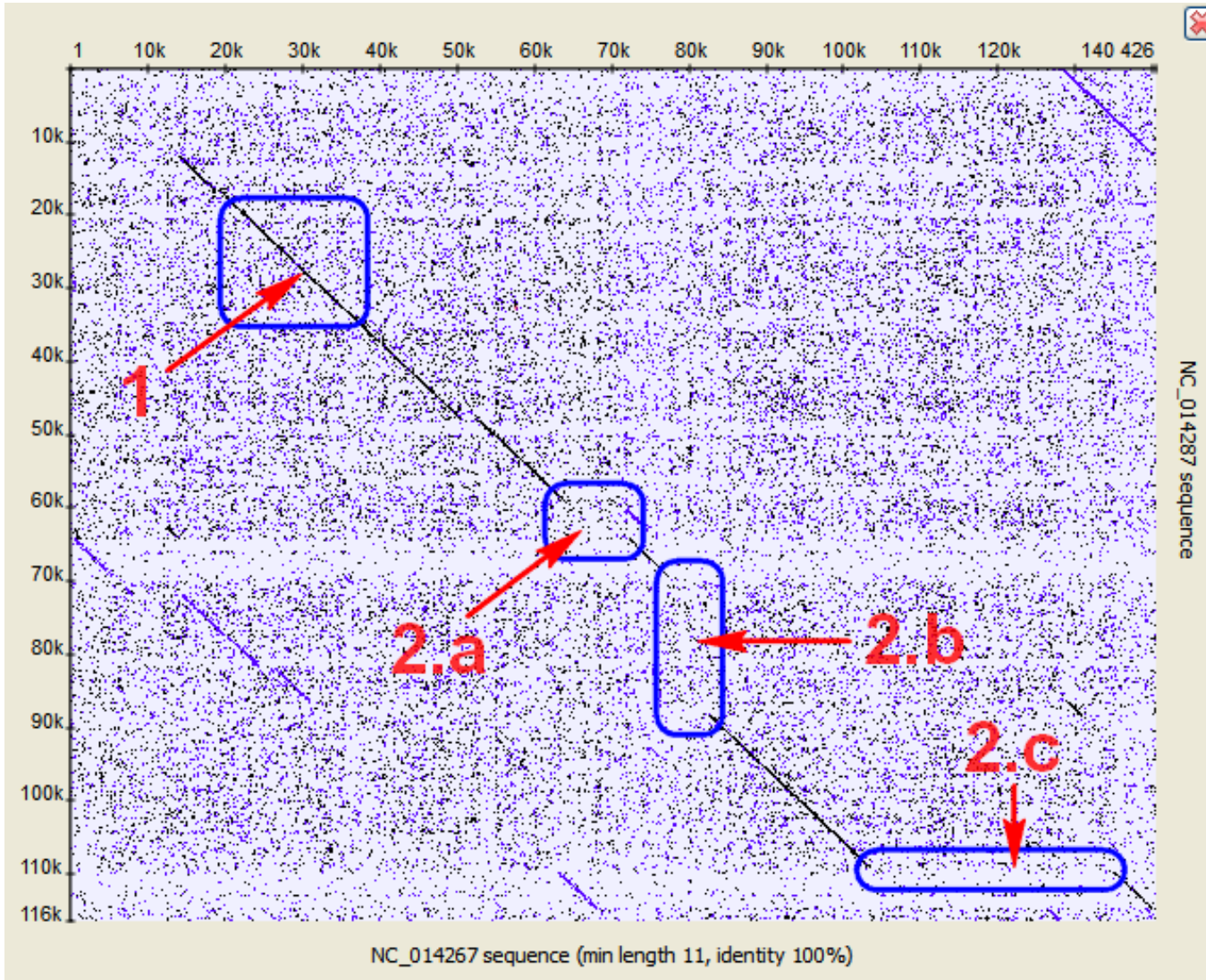
Дупликация



КОНЕЦ ПРЕЗЕНТАЦИИ

Interpreting Dotplot: Identifying Matches, Mutations, Inversions, etc

Создатель Yuliya Algaer, 2014



Эволюция белков

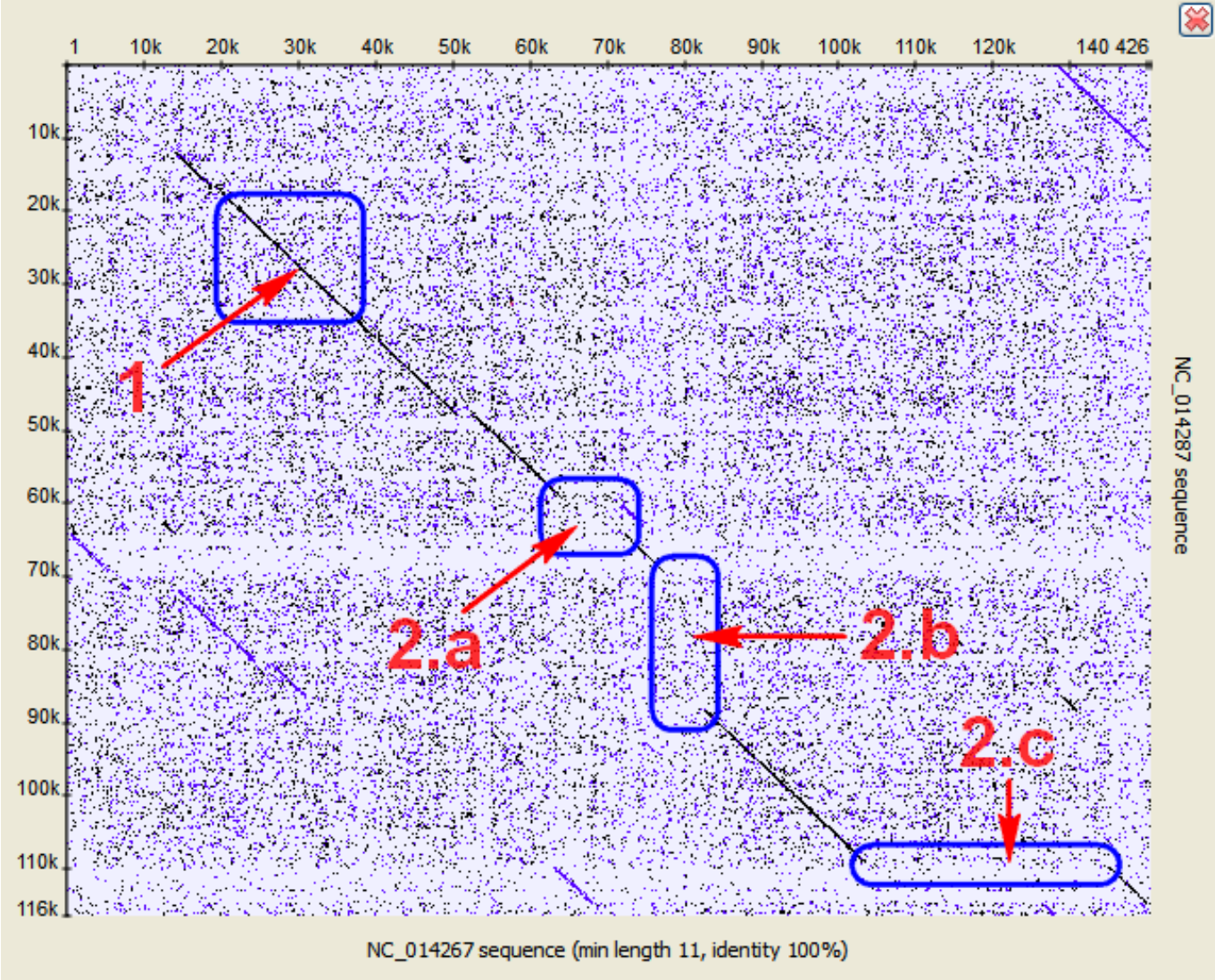
Локальная - небольшие изменения в гене
(Замены а.к. Делеции Вставки)

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!

Interpreting Dotplot: Identifying Matches, Mutations, Inversions, etc

Создатель Yuliya Algaer, 2014



ВЫРАВНИВАНИЕ

DNAK_THEAC 82 KFKVFDKEFTPQQISAFILQKIKKDA-EAFLGEPVNEAVITVPAYFNDNQR 131
 DNAK_PICTO 82 KYKIFGKEYTPQQISAFILQKIKRDA-EAFLGEPVTDAVITVPAYFNDNQR 131
 HSCA_ACIF2 116 RLRTVAGEKSPVEVSAEILRVLKERAVETLGGEPEGAVITVPAYFDEAQR 165
 HSCA_ACIF5 116 RLRTVAGEKSPVEVSAEILRVLKERAVETLGGEPEGAVITVPAYFDEAQR 165

DNAK_THEAC 132 QATKDAGT IAGFDVKRIINEPTAAALAYGVDKSGKSEKILVFDLGGGTLDV 182
 DNAK_PICTO 132 QATKDAGAIAGLNVRRINEPTAACLAYGIDKLNQTLKIVIYDLGGGTLDV 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