Выравнивания и гомология

Практические вопросы

План занятия

0. Парное выравнивание: карта локального сходства

I. Биологический смысл выравнивания
   a. В каком выравнивании вероятнее ошибки: Множественное выравнивание vs парное
   b. Принцип ВальдадRP Survivorship bias (https://en.wikipedia.org/wiki/Survivorship_bias#In_the_military)

II. Выравнивание последовательностей и наложение структур
   a. RdRp

III. Смысл выравнивания «невыравниваемых» участков белков

Nucleoline – GAR домен.

Как доказать гомологичность последовательностей и их частей

Рекомбинация, слияние и разделение

домены

Блоки

Pfam и др.

Обзор.

Jalview

Выравнивание без гомологии. Выращивание сайтов

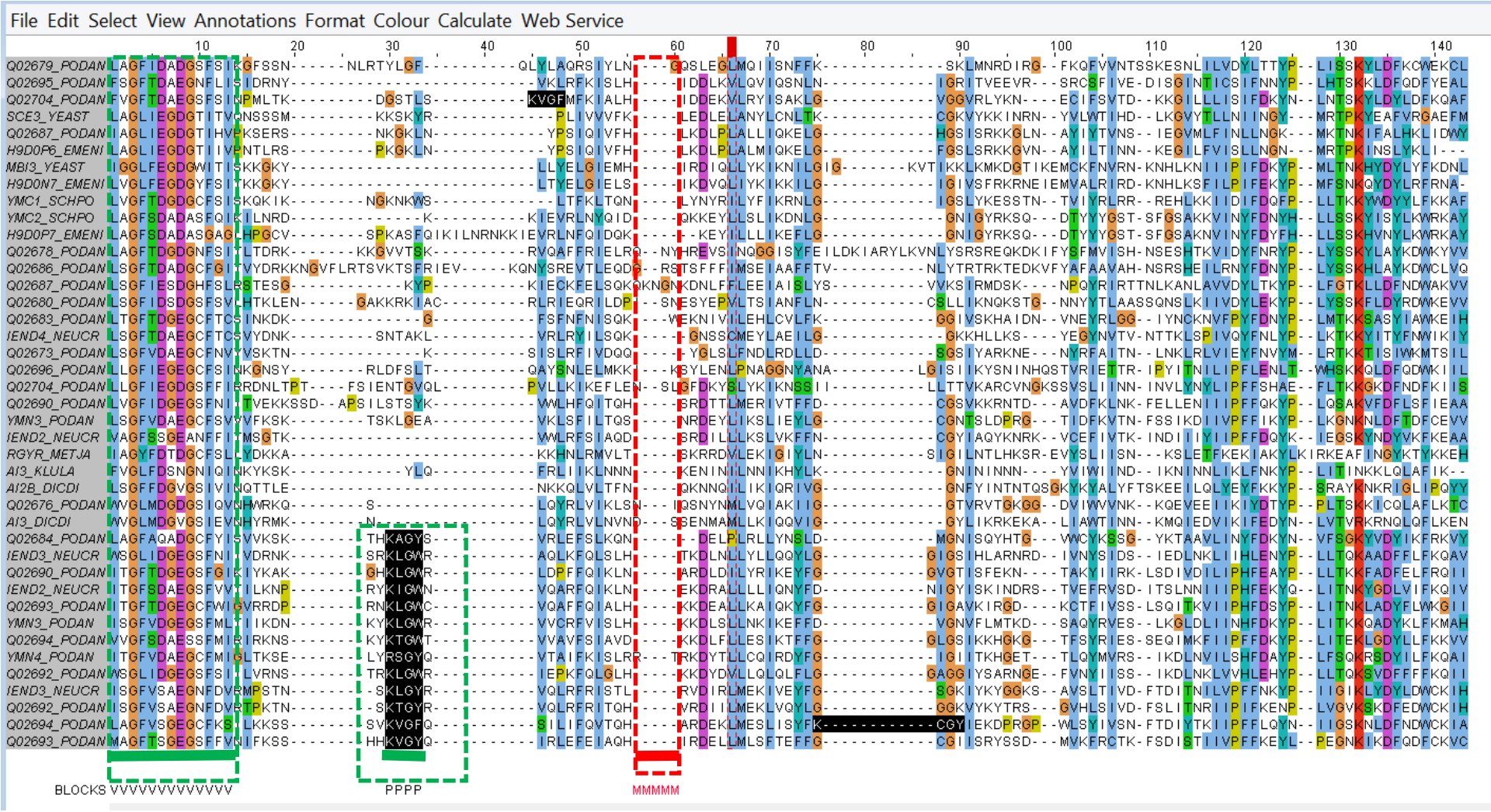
Алгоритмы и программы.

THIE_LACLS;THIE_MANSM;THIE_STRA3;THIE_LISIN;THIE_ANOFW;THIE_GEOTN;THIE_BACSU;THIE_BACA2;THIE_OCEIH;THIE_STAAB;THIE_STACT

Задание в классе

- Поиск во множественном выравнивании блоков «надежного» (плюс-блоки) и «ненадежного» (минус-блоки) выравнивания с помощью JalView
- Ссылки на два файла (Jalview project и protocol) должны появиться на странице pr14. Запись в очередь со ссылкой на pr14 должна появиться сегодня.

Блок, плюс-блоки и минус-блоки



[LVIVF].G[LVIVF].[DE].[DE .{0,5}

[KR].G[WFY]

Паттерны JalView

- [LVIMF].G[LVIVF].[DE].[DE] находит:
  - `MWGFAEAD`
  - `FAGLCDIE  и т.д.`
- [KR].G[WFY]
  - `RAGY и т.д.`
  - .{0,5} - от нуля до 5 любых букв подряд☺
- [ED]{5,10} - подскажите примеры!

1. Биологический смысл выравнивания

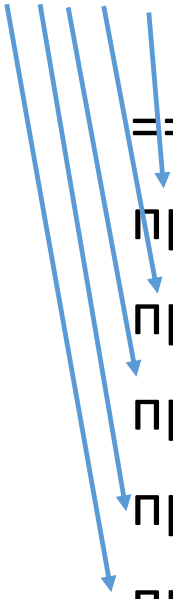
Выравнивание последовательностей потомков относительно предка

Гомологичные нуклеотиды ставим друг под другом

ПРЕДОК	1.	TATGCGAAT-GCCCTGAA	
сын	2.	TATGC**A**AAT-GCCCTGAA	замена
внук	3.	TATGC**A**AAT-GC**T**CTGAA	замена
правнук	4.	TATGC**A**AAT**C**GC**T**C**G**GAA	вставка 1 п.н.
праправнук	5.	TATGC**A**AA**A**C**C**GC**T**C**G**GAA	замена
прапраправнук	6.	TATGC**A**AA**-**C**C**GC**T**C**G**GAA	делеция 1п.н.

Выравнивание потомков относительно общего предка

Синий – вставка
Красный – замена
- Делеция
Относительно ПРЕДКА

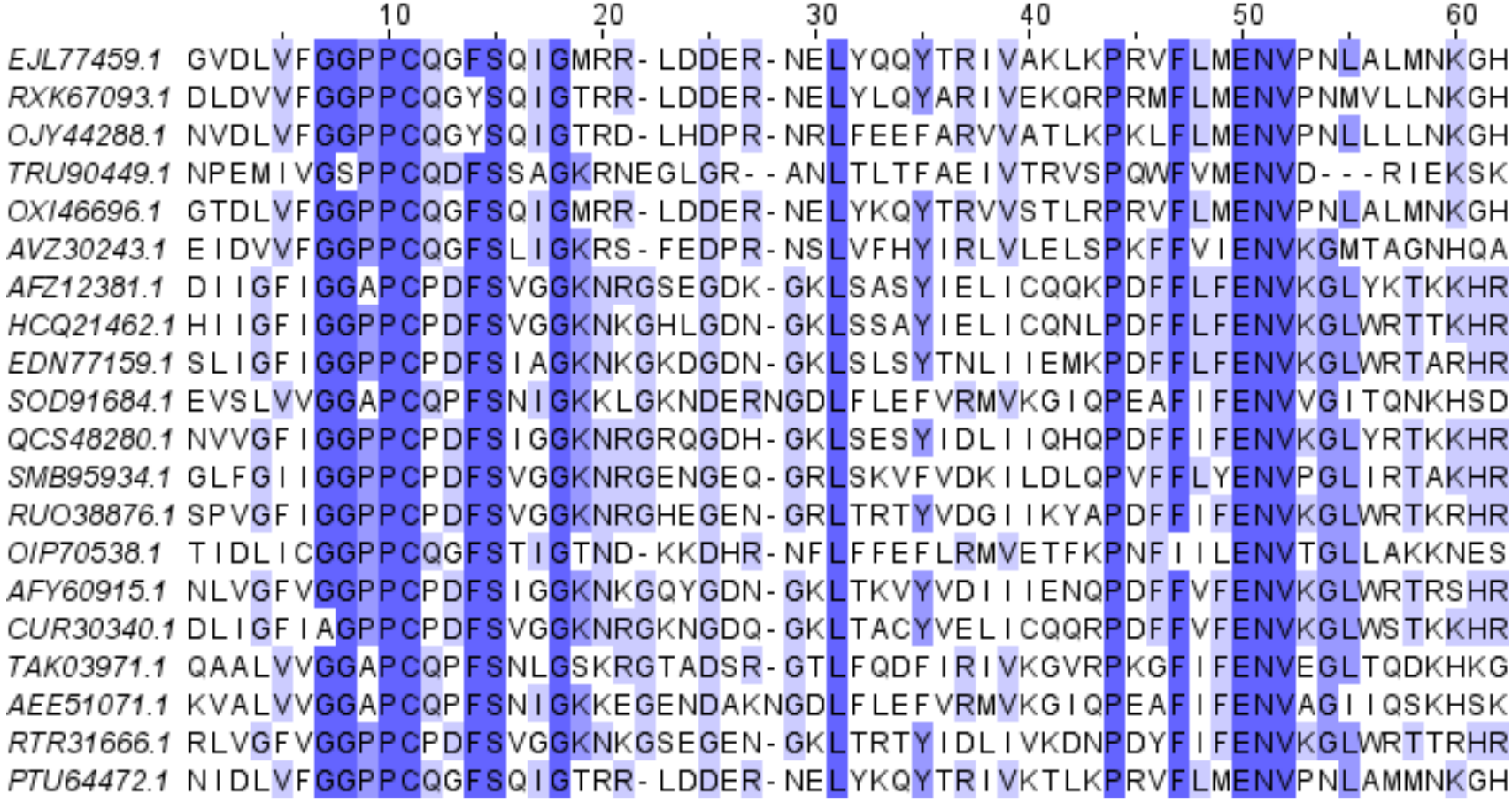


ПРЕДОК	**TAT--GCGAAT-GCCCTGAA**	
==========	================================	
праправнук1	**1.TAT--GCAAAACGCTCGGAA**	4 замены, 1 вставка
праправнук2	**2.TAT--GCGAA--GCTCTGAA**	1 делеция, 1 замена
праправнук3	**3.TAT--GCGTATCGCTCGGAA**	3 замены, 1 вставка
праправнук4	**4.TAT--GCGAAT-GC---GAA**	1 делеция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

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

Биологический смысл выравнивания

Обсуждаем смысл колонок и гэпов



Теория. Выравнивание – отражение эволюции последовательностей белков от общего предка

Кодоны стоящие в колонка произошли из одного кодона предка

Три понимания «правильного» выравнивания

Какое из пониманий выравнивания на предыдущем слайде?

1. ***Оптимальное выравнивание:*** наилучшее по весу

   Его ищут программы.

   Оптимальное выравнивание существует для любого набора последовательностей, даже негомологичных!

2. ***Эволюционное выравнивание***: запись, отражающая ход эволюции

   Не поддается достоверной реконструкции в большинстве реальных случаев; может отличаться от оптимального выравнивания.

   Алгоритм вычисления веса стараются выбрать так, чтобы можно было ожидать, что эволюционное выравнивание будет среди нескольких оптимальных.

   Для негомологичных последовательностей эволюционного выравнивания не существует!

3. ***Функциональное выравнивание:*** сопоставление функционально идентичных частей белков или нуклеиновых кислот

   Объясняет сохранение в эволюции одних частей белка и варьирование других. Поскольку функция и 3D-структура белка очень тесно связаны, функционально выровненные аминокислотные остатки должны иметь примерно одинаковое расположение в пространстве.

Никакое!

Не существует программ множественного выравнивания строящих оптимальное выравнивание!

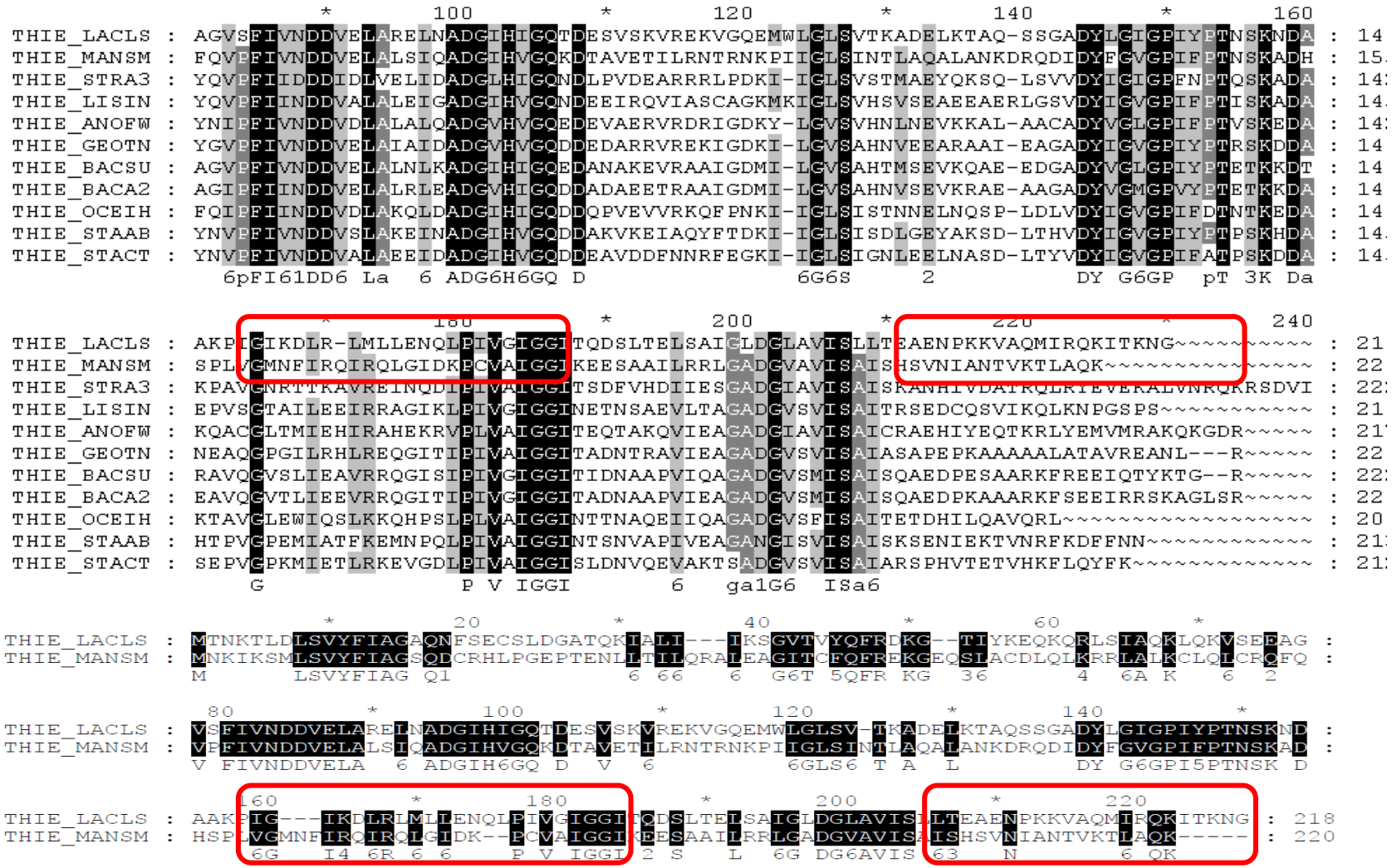
Почему?

Все программы множественного выравнивания - эвристические

2. Какое выравнивание правильнее?

Множественное даёт аргументы, опровергающие оптимальное парное выравнивание!

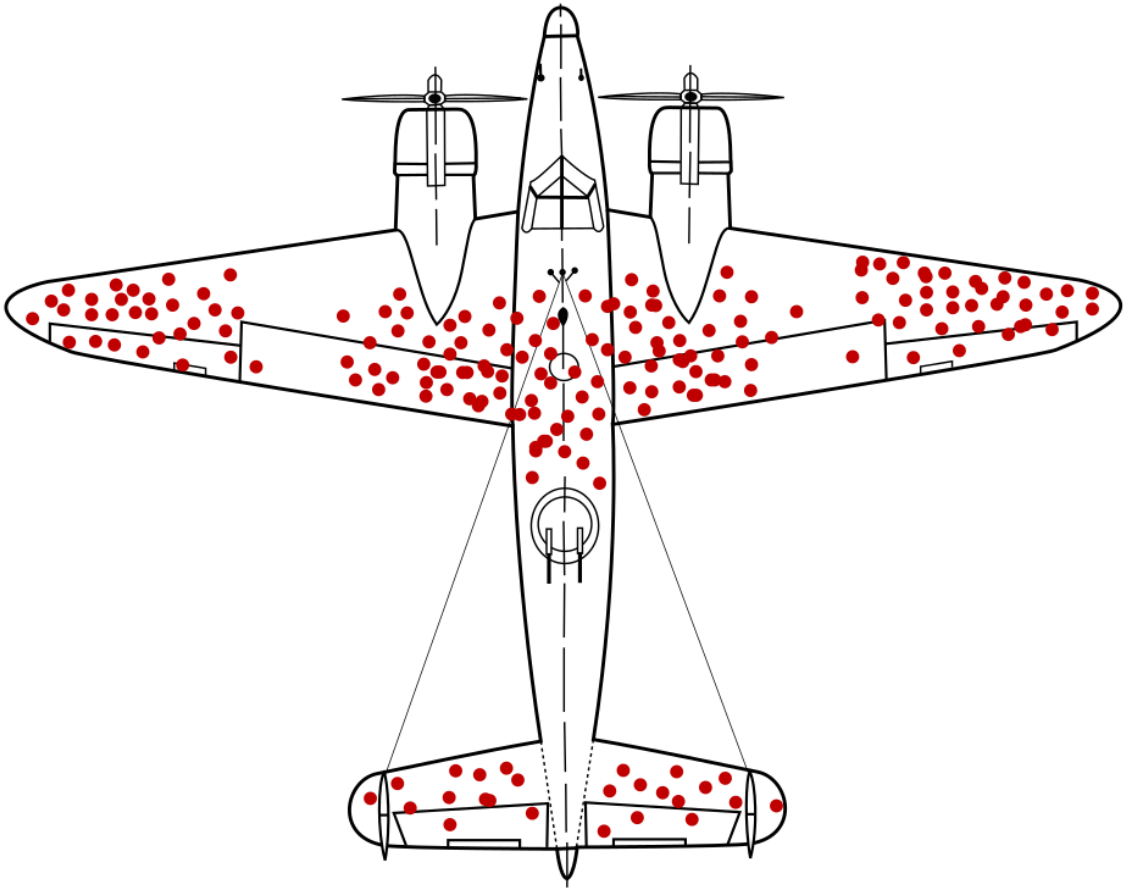
Пример. Из множественного выравнивания мы знаем, что между консервативными позициями бывает не более одной делеции



3. Принцип Вальда
(так назвал МГ)

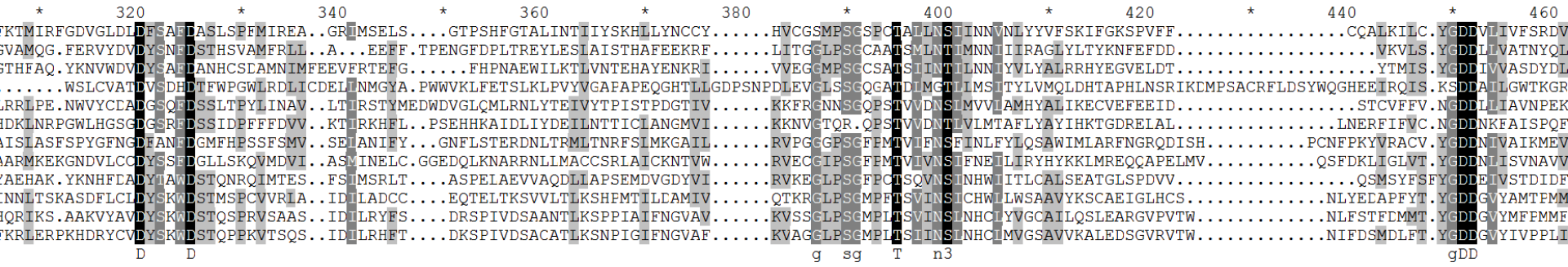
Пробоины на вернувшихся американских самолётах во время второй мировой войны.

Какие части укреплять броней?



Сохраняющееся (консервативное) – важно

РНК зависимая РНК полимераза (RdRP), консервативные участки



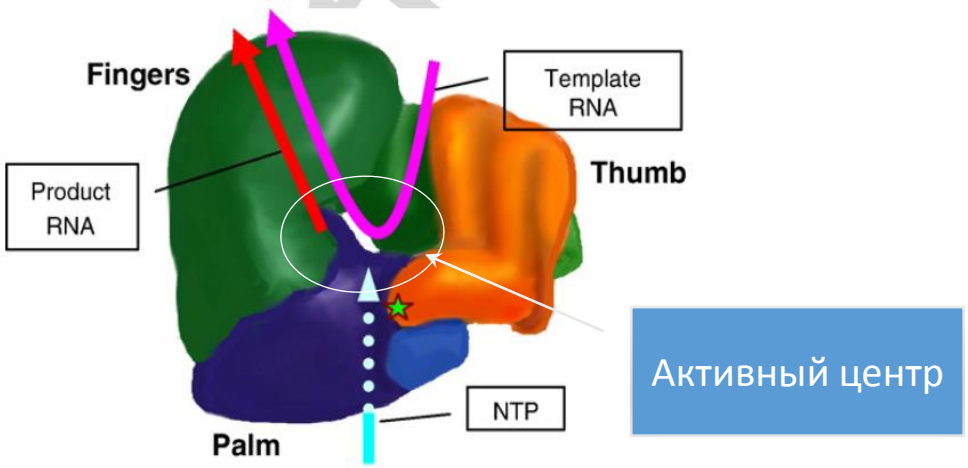
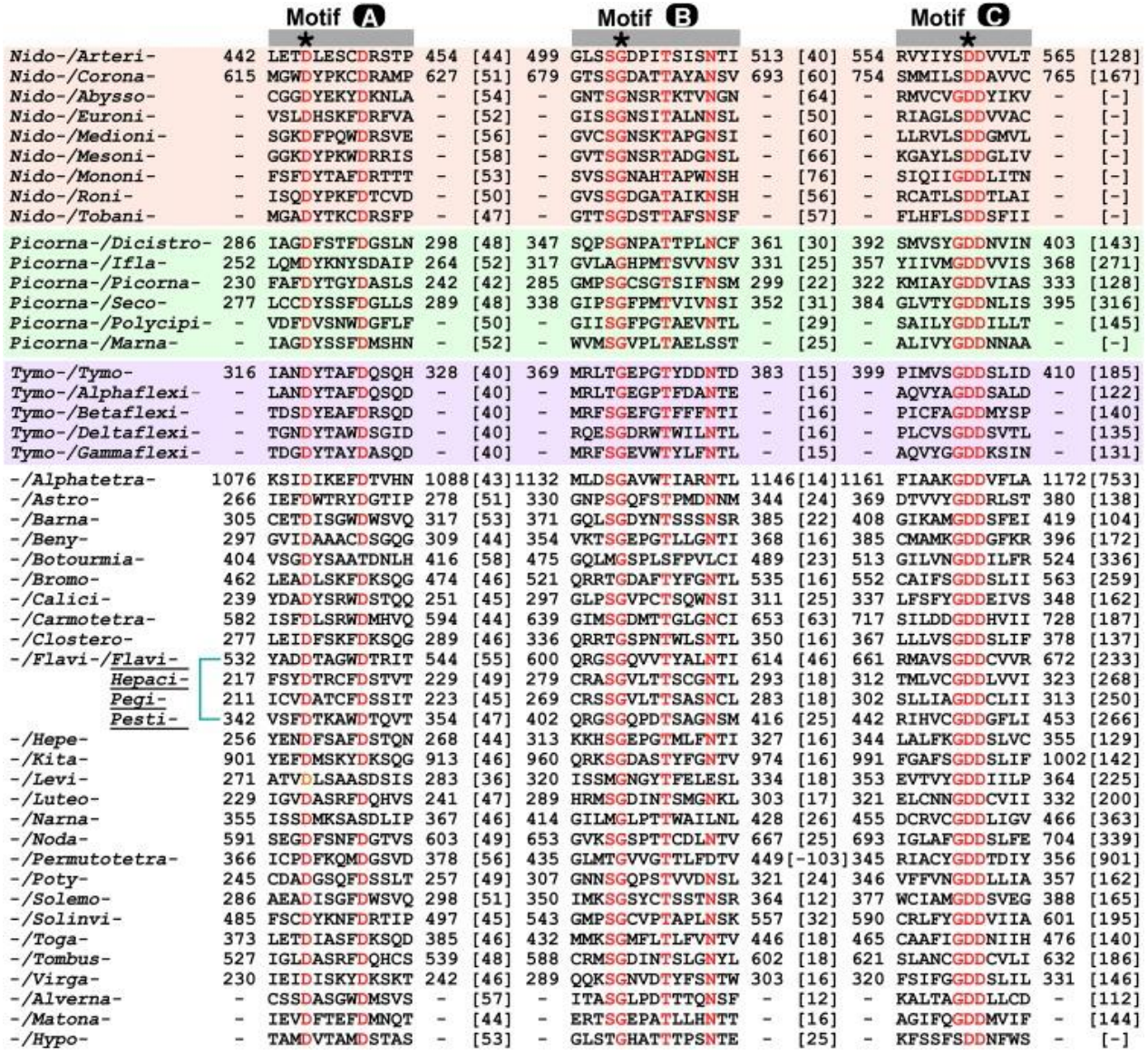


Fig. 1. Schematic architecture of "small" RdRP. The hairpin between the palm and thumb domains is in light blue. The predicted approximate location of MV RdRP Trp460 is marked by star.

На каких участках выравнивание правильное –
совпадает с эволюционным?

Рисунок мой :) хвастаюсь

Каталитические мотивы RdRP



		Motif A			Motif B			Motif C	
Nido-/Arteri-	442	LETDLESCDRSTP	454	[44]	499	GLSSGDPITSISNTI	513	[40]	554	RVYIYSDDVVLT	565	[128]
Nido-/Corona-	615	MGWDYPKCDRAMP	627	[51]	679	GTSSGDATTAYANSV	693	[60]	754	SMMILSDDAVVC	765	[167]
Nido-/Abysso-	-	CGGDYEKYDKNLA	-	[54]	-	GNTSGNSRTKTVNGN	-	[64]	-	RMVCVGDDYIKV	-	[-]
Nido-/Euroni-	-	VSLDHSKFDRFVA	-	[52]	-	GISSGNSITALNNSL	-	[50]	-	RIAGLSDDVVAC	-	[-]
Nido-/Medioni-	-	SGKDFPQWDRSVE	-	[56]	-	GVCSGNSKTAPGNSI	-	[60]	-	LLRVLSDDGMVL	-	[-]
Nido-/Mesoni-	-	GGKDYPKWDRRIS	-	[58]	-	GVTSGNSRTADGNSL	-	[66]	-	KGAYLSDDGLIV	-	[-]
Nido-/Mononi-	-	FSFDYTAFDRTTT	-	[53]	-	SVSSGNAHTAPWNSH	-	[76]	-	SIQIIGDDLITN	-	[-]
Nido-/Roni-	-	ISQDYPKFDTCVD	-	[50]	-	GVSSGDGATAIKNSH	-	[56]	-	RCATLSDDTLAI	-	[-]
Nido-/Tobani-	-	MGADYTKCDRSFP	-	[47]	-	GTTSGDSTTAFSNSF	-	[57]	-	FLHFLSDDSFII	-	[-]
Picorna-/Dicistro-	286	IAGDFSTFDGSLN	298	[48]	347	SQPSGNPATTPLNCF	361	[30]	392	SMVSYGDDNVIN	403	[143]
Picorna-/Ifla-	252	LQMDYKNYSDAIP	264	[52]	317	GVLAGHPMTSVVNSV	331	[25]	357	YIIVMGDDVVIS	368	[271]
Picorna-/Picorna-	230	FAFDYTGYDASLS	242	[42]	285	GMPSGCSGTSIFNSM	299	[22]	322	KMIAYGDDVIAS	333	[128]
Picorna-/Seco-	277	LCCDYSSFDGLLS	289	[48]	338	GIPSGFPMTVIVNSI	352	[31]	384	GLVTYGDDNLIS	395	[316]
Picorna-/Polycipi-	-	VDFDVSNWDGFLF	-	[50]	-	GIISGFPGTAEVNTL	-	[29]	-	SAILYGDDILLT	-	[145]
Picorna-/Marna-	-	IAGDYSSFDMSHN	-	[52]	-	WVMSGVPLTAELSST	-	[25]	-	ALIVYGDDNNAA	-	[-]
Tymo-/Tymo-	316	IANDYTAFDQSQH	328	[40]	369	MRLTGEPGTYDDNTD	383	[15]	399	PIMVSGDDSLID	410	[185]
Tymo-/Alphaflexi-	-	LANDYTAFDQSQD	-	[40]	-	MRLTGEGPTFDANTE	-	[16]	-	AQVYAGDDSALD	-	[122]
Tymo-/Betaflexi-	-	TDSDYEAFDRSQD	-	[40]	-	MRFSGEFGTFFFNTI	-	[16]	-	PICFAGDDMYSP	-	[140]
Tymo-/Deltaflexi-	-	TGNDYTAWDSGID	-	[40]	-	RQESGDRWTWILNTL	-	[16]	-	PLCVSGDDSVTL	-	[135]
Tymo-/Gammaflexi-	-	TDGDYTAYDASQD	-	[40]	-	MRFSGEVWTYLFNTL	-	[15]	-	AQVYGGDDKSIN	-	[131]
-/Alphatetra-	1076	KSIDIKEFDTVHN	1088	[43]	1132	MLDSGAVWTIARNTL	1146	[14]	1161	FIAAKGDDVFLA	1172	[753]
-/Astro-	266	IEFDWTRYDGTIP	278	[51]	330	GNPSGQFSTPMDNNM	344	[24]	369	DTVVYGDDRLST	380	[138]
-/Barna-	305	CETDISGWDWSVQ	317	[53]	371	GQLSGDYNTSSSNSR	385	[22]	408	GIKAMGDDSFEI	419	[104]
-/Beny-	297	GVIDAAACDSGQG	309	[44]	354	VKTSGEPGTLLGNTI	368	[16]	385	CMAMKGDDGFKR	396	[172]
-/Botourmia-	404	VSGDYSAATDNLH	416	[58]	475	GQLMGSPLSFPVLCI	489	[23]	513	GILVNGDDILFR	524	[336]
-/Bromo-	462	LEADLSKFDKSQG	474	[46]	521	QRRTGDAFTYFGNTL	535	[16]	552	CAIFSGDDSLII	563	[259]
-/Calici-	239	YDADYSRWDSTQQ	251	[45]	297	GLPSGVPCTSQWNSI	311	[25]	337	LFSFYGDDEIVS	348	[162]
-/Carmotetra-	582	ISFDLSRWDMHVQ	594	[44]	639	GIMSGDMTTGLGNCI	653	[63]	717	SILDDGDDHVII	728	[187]
-/Clostero-	277	LEIDFSKFDKSQG	289	[46]	336	QRRTGSPNTWLSNTL	350	[16]	367	LLLVSGDDSLIF	378	[137]
-/Flavi-/Flavi-	532	YADDTAGWDTRIT	544	[55]	600	QRGSGQVVTYALNTI	614	[46]	661	RMAVSGDDCVVR	672	[233]
Hepaci-	217	FSYDTRCFDSTVT	229	[49]	279	CRASGVLTTSCGNTL	293	[18]	312	TMLVCGDDLVVI	323	[268]
Pegi-	211	ICVDATCFDSSIT	223	[45]	269	CRSSGVLTTSASNCL	283	[18]	302	SLLIAGDDCLII	313	[250]
Pesti-	342	VSFDTKAWDTQVT	354	[47]	402	QRGSGQPDTSAGNSM	416	[25]	442	RIHVCGDDGFLI	453	[266]
-/Hepe-	256	YENDFSAFDSTQN	268	[44]	313	KKHSGEPGTMLFNTI	327	[16]	344	LALFKGDDSLVC	355	[129]
-/Kita-	901	YEFDMSKYDKSQG	913	[46]	960	QRKSGDASTYFGNTV	974	[16]	991	FGAFSGDDSLIF	1002	[142]
-/Levi-	271	ATVDLSAASDSIS	283	[36]	320	ISSMGNGYTFELESL	334	[18]	353	EVTVYGDDIILP	364	[225]
-/Luteo-	229	IGVDASRFDQHVS	241	[47]	289	HRMSGDINTSMGNKL	303	[17]	321	ELCNNGDDCVII	332	[200]
-/Narna-	355	ISSDMKSASDLIP	367	[46]	414	GILMGLPTTWAILNL	428	[26]	455	DCRVCGDDLIGV	466	[363]
-/Noda-	591	SEGDFSNFDGTVS	603	[49]	653	GVKSGSPTTCDLNTV	667	[25]	693	IGLAFGDDSLFE	704	[339]
-/Permutotetra-	366	ICPDFKQMDGSVD	378	[56]	435	GLMTGVVGTTLFDTV	449	[-103]	345	RIACYGDDTDIY	356	[901]
-/Poty-	245	CDADGSQFDSSLT	257	[49]	307	GNNSGQPSTVVDNSL	321	[24]	346	VFFVNGDDLLIA	357	[162]
-/Solemo-	286	AEADISGFDWSVQ	298	[51]	350	IMKSGSYCTSSTNSR	364	[12]	377	WCIAMGDDSVEG	388	[165]
-/Solinvi-	485	FSCDYKNFDRTIP	497	[45]	543	GMPSGCVPTAPLNSK	557	[32]	590	CRLFYGDDVIIA	601	[195]
-/Toga-	373	LETDIASFDKSQD	385	[46]	432	MMKSGMFLTLFVNTV	446	[18]	465	CAAFIGDDNIIH	476	[140]
-/Tombus-	527	IGLDASRFDQHCS	539	[48]	588	CRMSGDINTSLGNYL	602	[18]	621	SLANCGDDCVLI	632	[186]
-/Virga-	230	IEIDISKYDKSKT	242	[46]	289	QQKSGNVDTYFSNTW	303	[16]	320	FSIFGGDDSLIL	331	[146]
-/Alverna-	-	CSSDASGWDMSVS	-	[57]	-	ITASGLPDTTTQNSF	-	[12]	-	KALTAGDDLLCD	-	[112]
-/Matona-	-	IEVDFTEFDMNQT	-	[44]	-	ERTSGEPATLLHNTT	-	[16]	-	AGIFQGDDMVIF	-	[144]
-/Hypo-	-	TAMDVTAMDSTAS	-	[53]	-	GLSTGHATTTPSNTE	-	[25]	-	KFSSFSDDNFWS	-	[-]

The RdRP active site is surrounded by the palm, fingers, and thumb domains with seven catalytic motifs (motifs A–G) distributed within the palm (motifs A–E) and fingers (motifs F–G) (Poch et al., 1989; Gorbalenya et al., 2002; Bruenn, 2003; te Velthuis, 2014; Wu et al., 2015) (see an alignment of motif A–C of the 49 representative RdRP sequences in **Figure 2**).

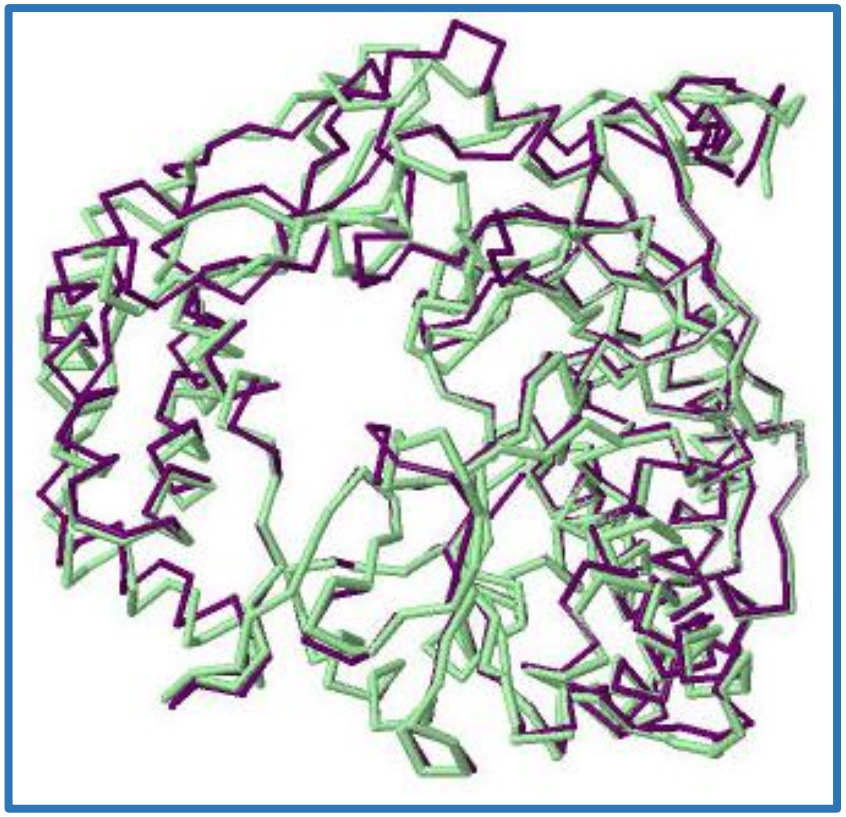
Hengxia Jia and Peng Gong, review, Front Microbiol, 2019

4. Выравнивание и совмещение структур

Для ГЛОБУЛЯРНЫХ белков правильность выравнивания может быть проверена совмещением структур – при наличии

Проверка выравнивания по совмещению полипептидных остовов 3Dструктур

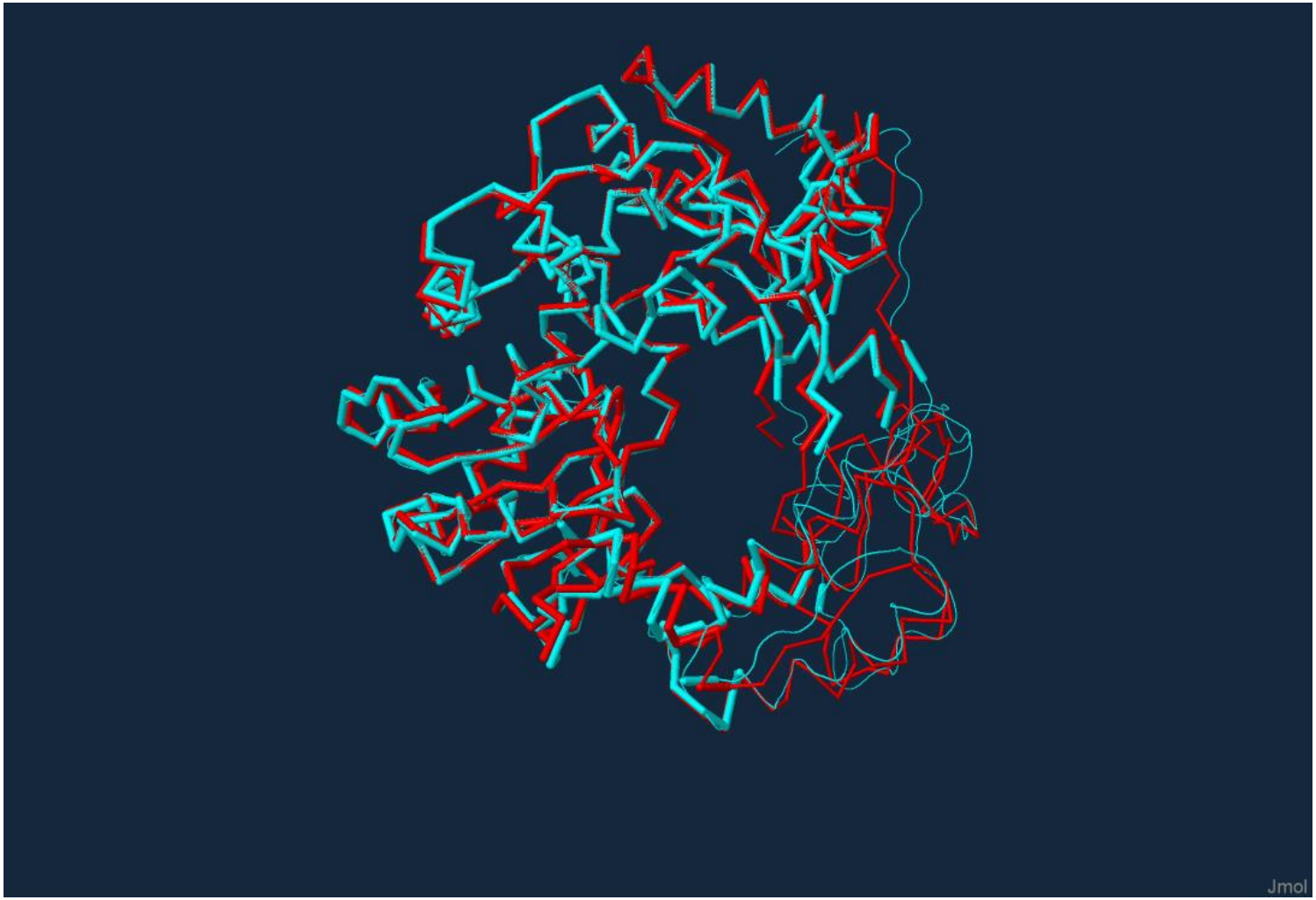
Структура консервативнее последовательностей



Совмещение полипептидных цепей двух RdRP +РНК вирусов.

Не успел посмотреть каких. Выбрал случайно из совмещения пяти структур

Совмещение 2x RdRp



Толстая backbone модель построена по сближенным (<2 ангстрем) C_alpha атомам

Как проверять выравнивание при отсутствии 3D структур?

Варианты

- Выделение блоков надежного выравнивнивания на основании сходства последовательностей (плюс-блоков)
- Поиск информации о функционально важных аминокислотных остатках (аннотации записей, литература)
- Предсказание вторичной структуры по последовательности (или по 3D структуре даже если она есть только у одной последовательности)
- Предсказание неструктурированных участков в белках
- Опыт и интуиция (!) Полезно воображать 3D структуру. Гэпы преимущественно в петлях между спиралями и тяжами бета-листов

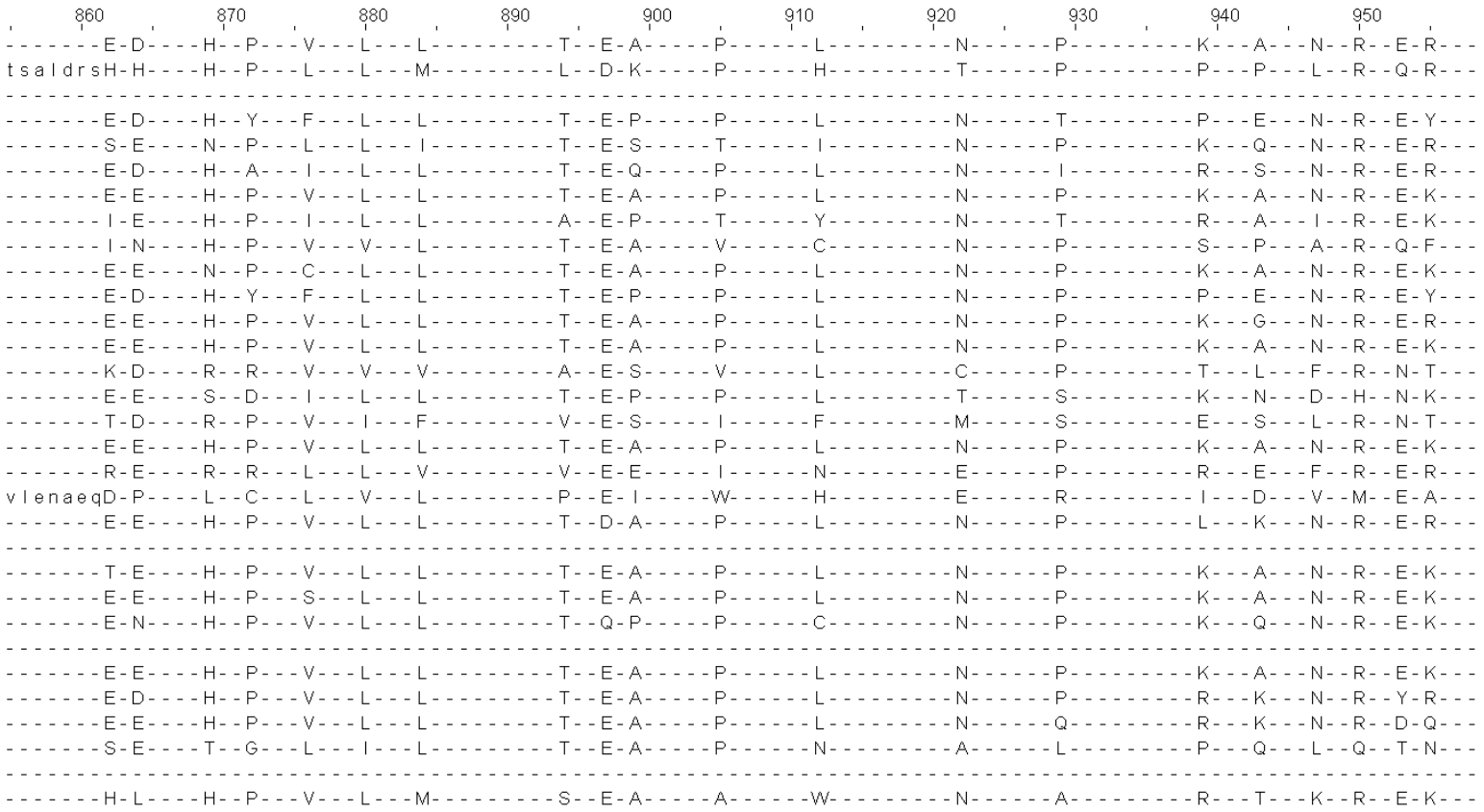
Программа-выравниватель лишь инструмент

- Программа-выравниватель выдает ГИПОТЕЗУ об эволюцинном выравнивании
  Гипотеза может быть верна в отдельных блоках и не верна в остальных блока или во всем выравнивании:)

- Решение за человеком, который обосновывает выравнивания опираясь на объективные данные о выравнивании и/или его блоках
- Анализируя выравнивание, полезно соображать как наблюдаемые распределения гэпов затрагивают структуру белка, не входит ли в противоречие с вероятным ходом эволюции

Иногда программа выдаёт такое выравнивание …

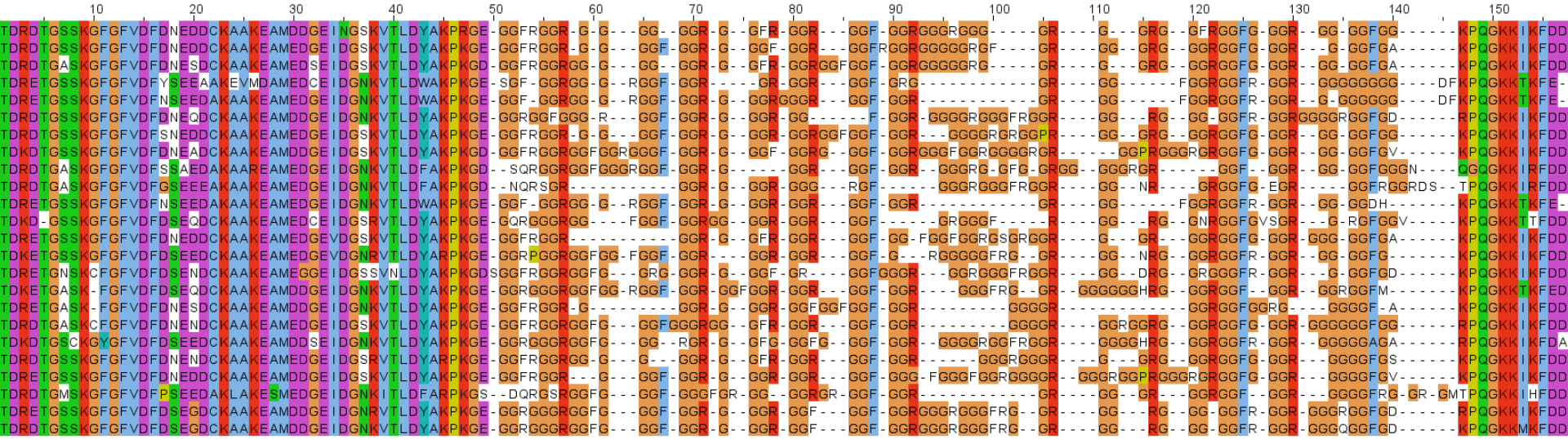
Это выравнивание не может быть правильным …..



~1400 последовательностей, почти в каждой позиции найдется какая-нибудь вставка хотя бы в одной последовательности

5. Выравнивание «невыравниваемых» участков белков

Был ли GAR у общего предка этих белков?



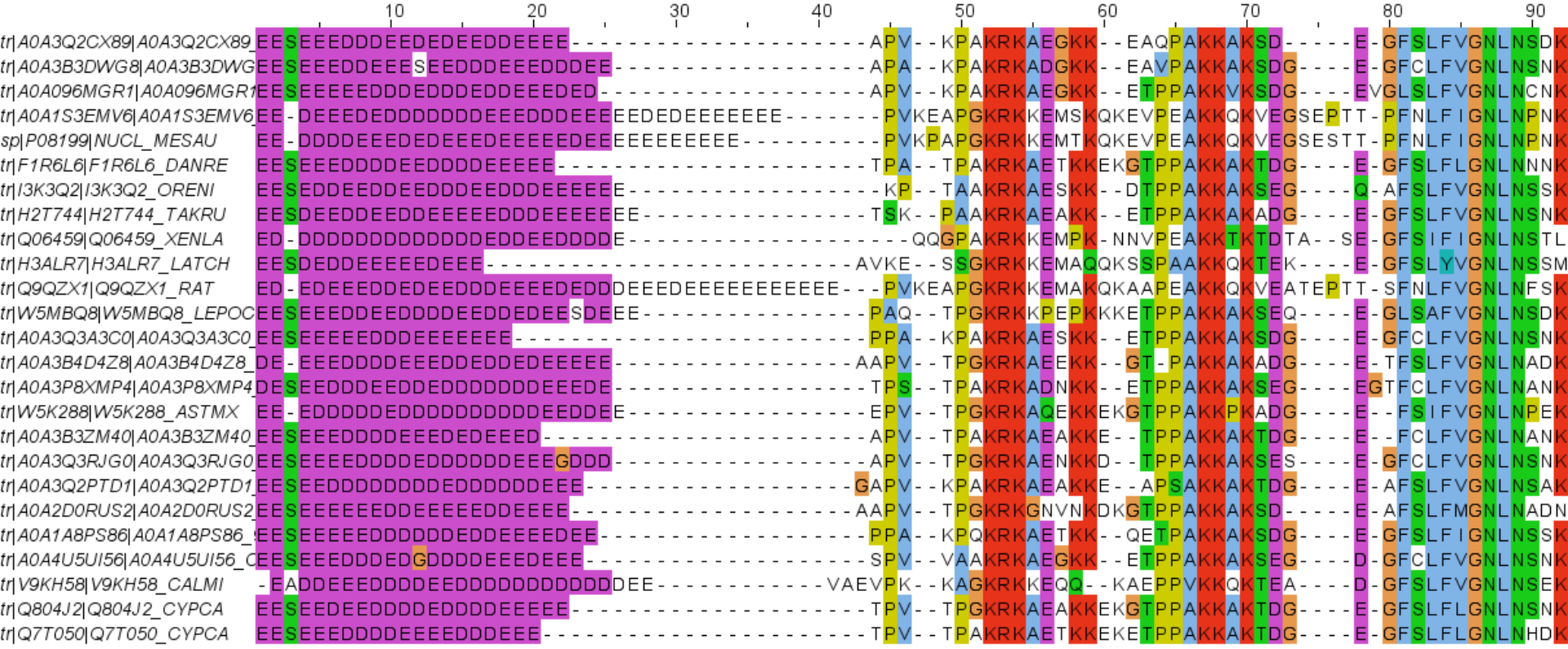
Предположительно был, так как он есть у нуклеолинов всех ныне живущих эукариот.

Утверждать, что остатки, стоящие в одной колонке произошли от одного и того же остатка у общего предка нельзя.

Из-за повторяющихся мотивов RGG и других, можно предположить, что GAR эволюционировали путем дупликации повторов.

Уникальная дупликация простых повторов в геноме индивидуума лежит в основе идентификации личности по ДНК

Nucleolin. Кислый участок и сигнал сигнал ядерной локализации



6. Домены БД Pfam

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

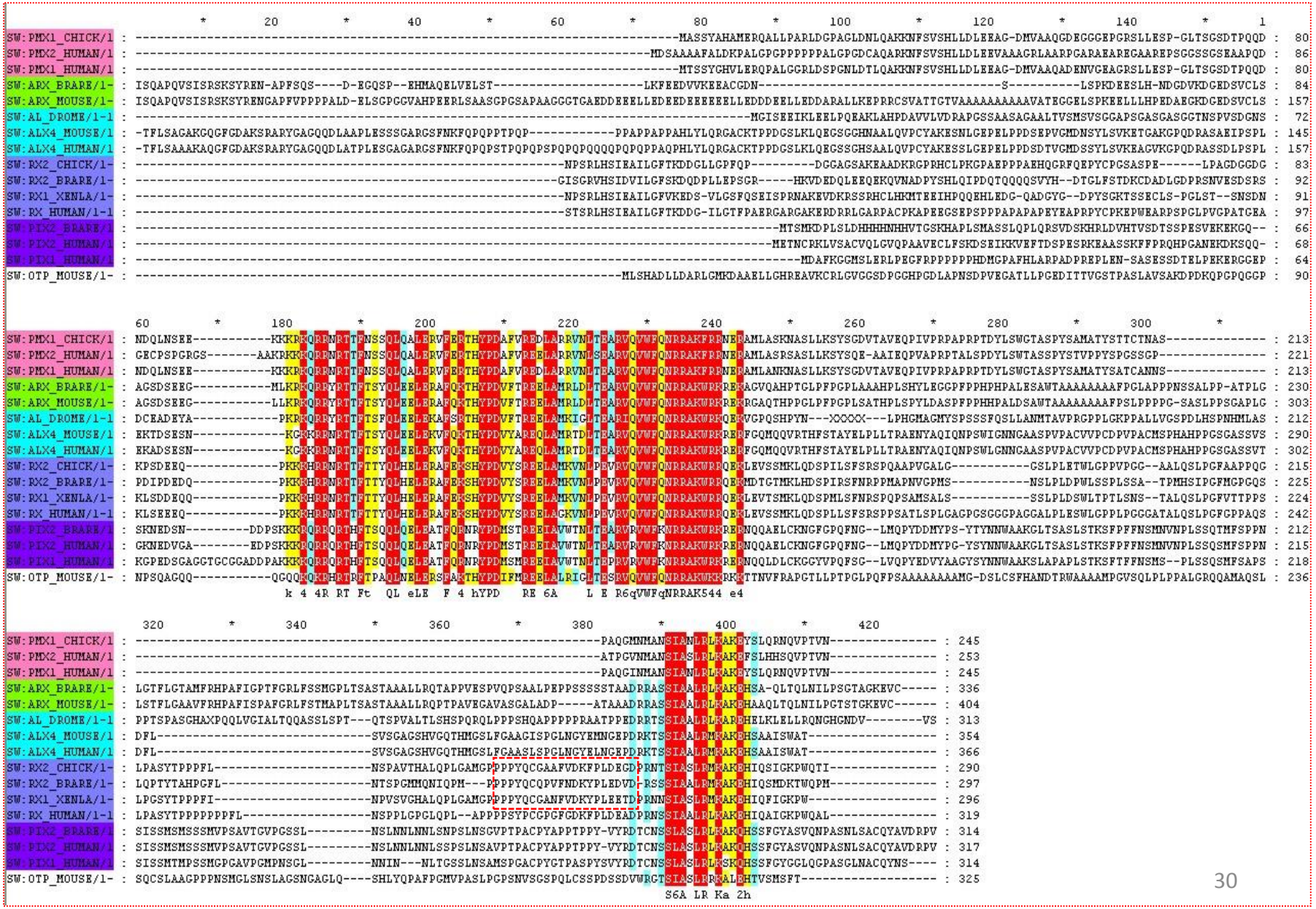
Терминологическая проблема.

ДОМЕН – набор фрагментов последовательностей и их выравнивание. Имеет 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)

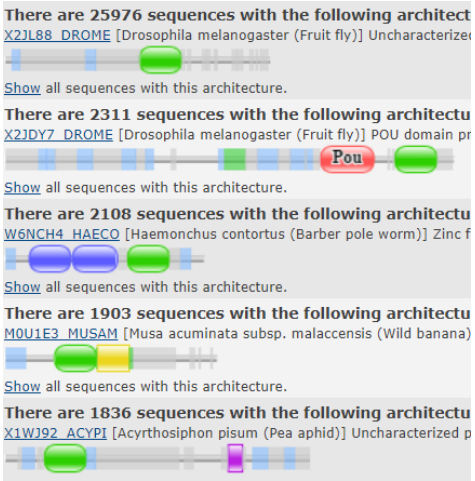
There are 1836 sequences with the following architecture:
Homeodomain, OAR

Гомеодомен является ДОМЕНОМ

Доказательство

- Выравнивание, свидетельствующее о гомологичности последовательностей

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



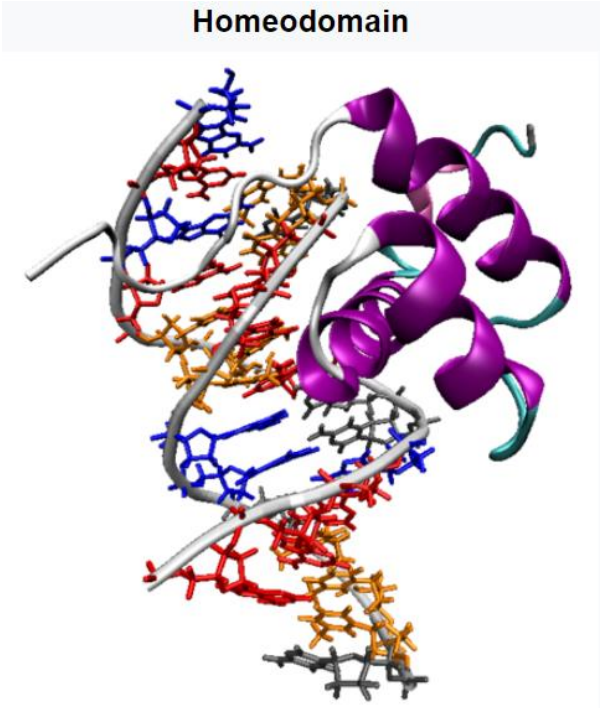
Эволюционные домены

- 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]



Что есть в Pfam

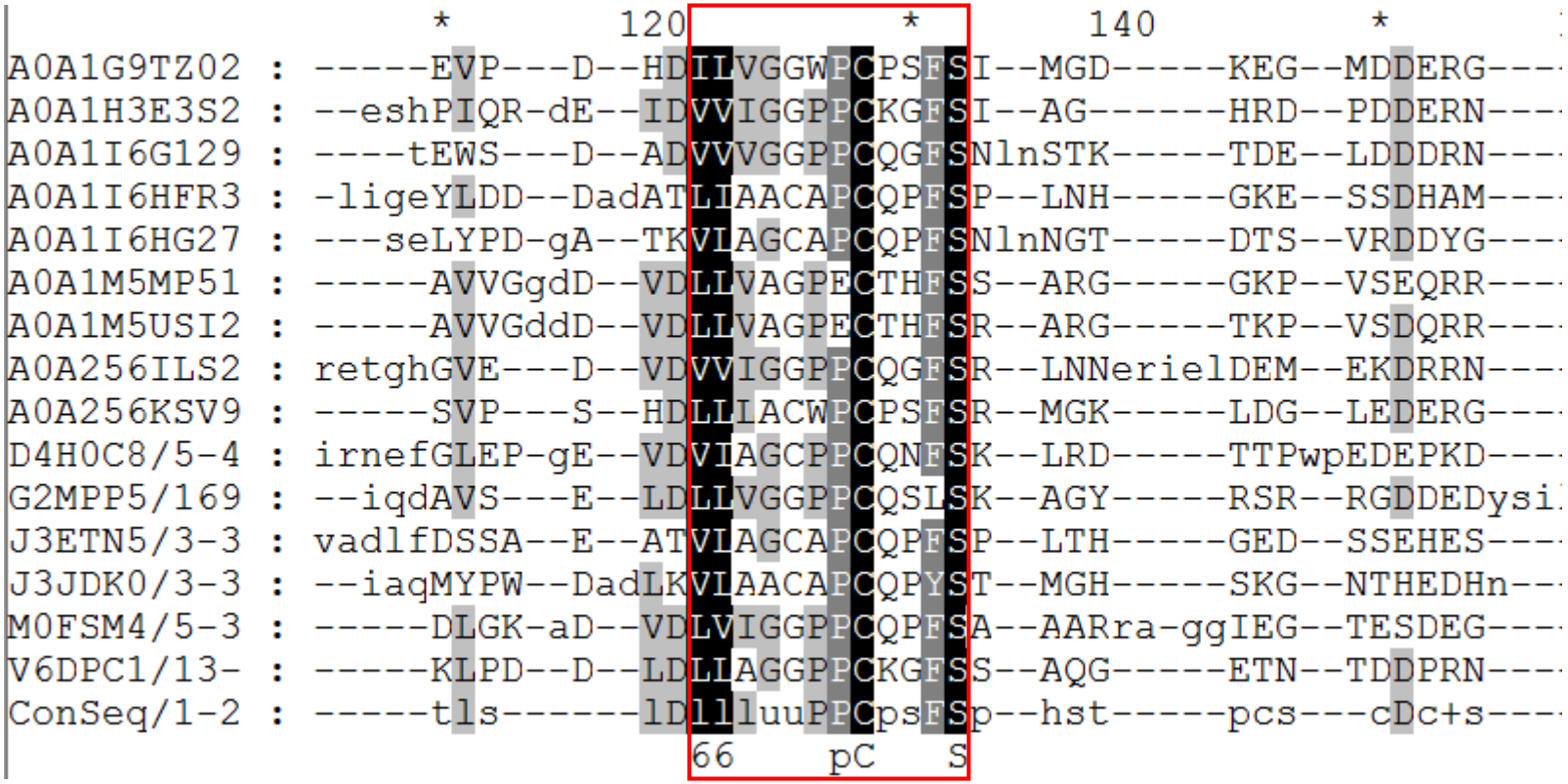
- Название домена, ID AC
- Выравнивания
  - Seed
  - Full
- Доменные архитектуры
- Описание функции
- Таксономическая распространенность
- 3D структуры
- HMM профиль выравнивания
- Clan

6. блоки

Консервативный блок выравнивания (плюс-блок)

- Подтверждается консервативными или функционально консервативными позиция во всех фрагментах
- 1я и последняя позиции такие
- В блоке нет гэпов
- Консервативных позиций достаточно, чтобы похожие последовательности почти не встречались в дпугих местах выравнивания.

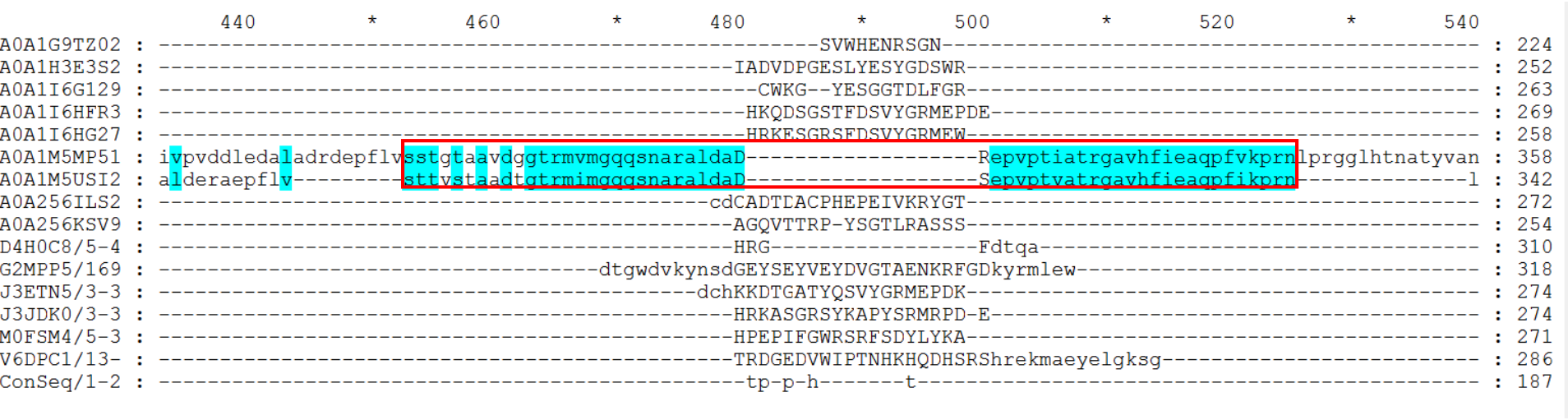
Вертикальный плюс-блок



В вертикальном блоке, по определению, содержатся фрагменты ИЗ ВСЕХ последовательностей.

Значит в вертикальном К-блоке мы предполагаем гомологичность всех фрагментов и всех остатков в каждой из колонок, неважно, консервативные они или нет

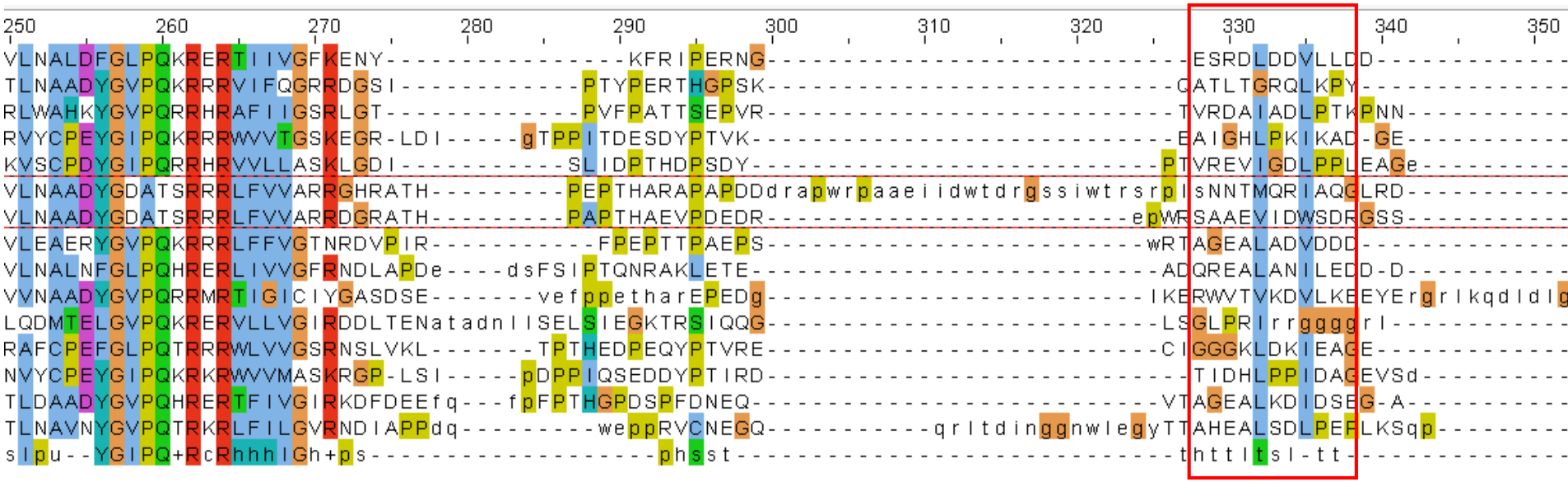
Частичный плюс-блок



Как бы назвать такие блоки? Пока «не вертикальный плюс-блок»

В таком блоке все фрагменты гомологичны и все остатки фрагментов из блока в каждой позиции блока предполагаем гомологичными.

Минус блок (не консервативный)



Выделить минус блок труднее. Для этого надо найти все плюс-блоки. То, что останется надо разбить на блоки, они и будут минус-блоки

В минус блоках мы не предполагаем наличие гомологичных остатков в любой позиции. Группировка остатков в одну позицию имеет целью сократить длину выравнивания, не более.

Но программы выравнивания пытаются выровнять невыравниваемое :(

JalView

Демонстрация

КОНЕЦ ПРЕЗЕНТАЦИИ