

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 Вальда RP 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 гомологию последовательностей и их частей

Рекомбинация, слияние и разделение

домены

Блоки

Pfam и др.

Обзор.

Jalview

Выравнивание без гомологии. Выращивание сайтов

Алгоритмы и программы.

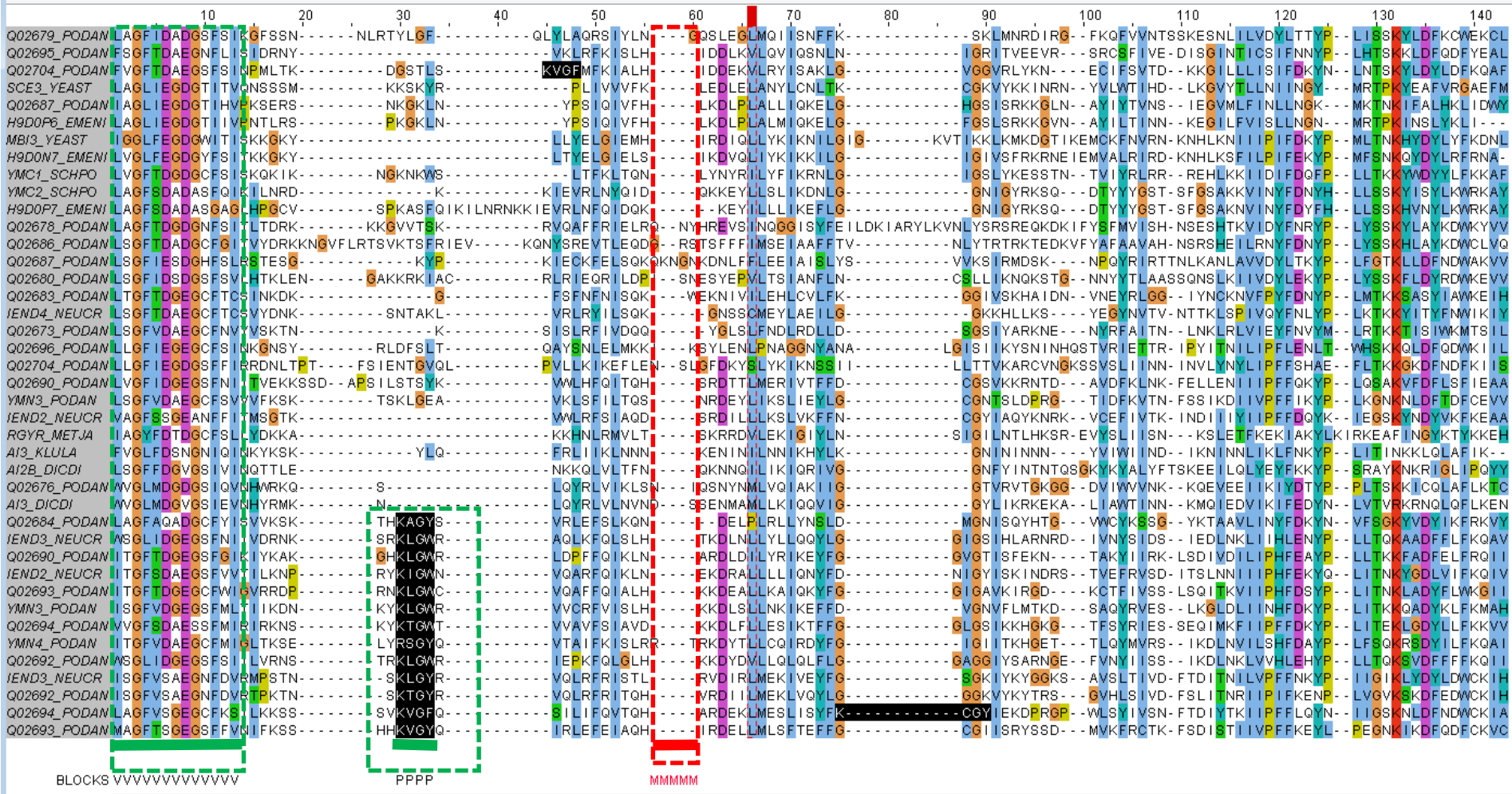
THE_LACLS;THE_MANSM;THE_STRA3;THE_L
ISIN;THE_ANOFW;THE_GEOTN;THE_BACSU;T
HIE_BACA2;THE_OCEIH;THE_STAAB;THE_STA
CT

Задание в классе

- Поиск во множественном выравнении 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

File Edit Select View Annotations Format Colour Calculate Web Service



[LVIVF].G[LVIVF].[DE].[DE

.{0,5}

[KR].G[WFY]

Паттерны JaView

- [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

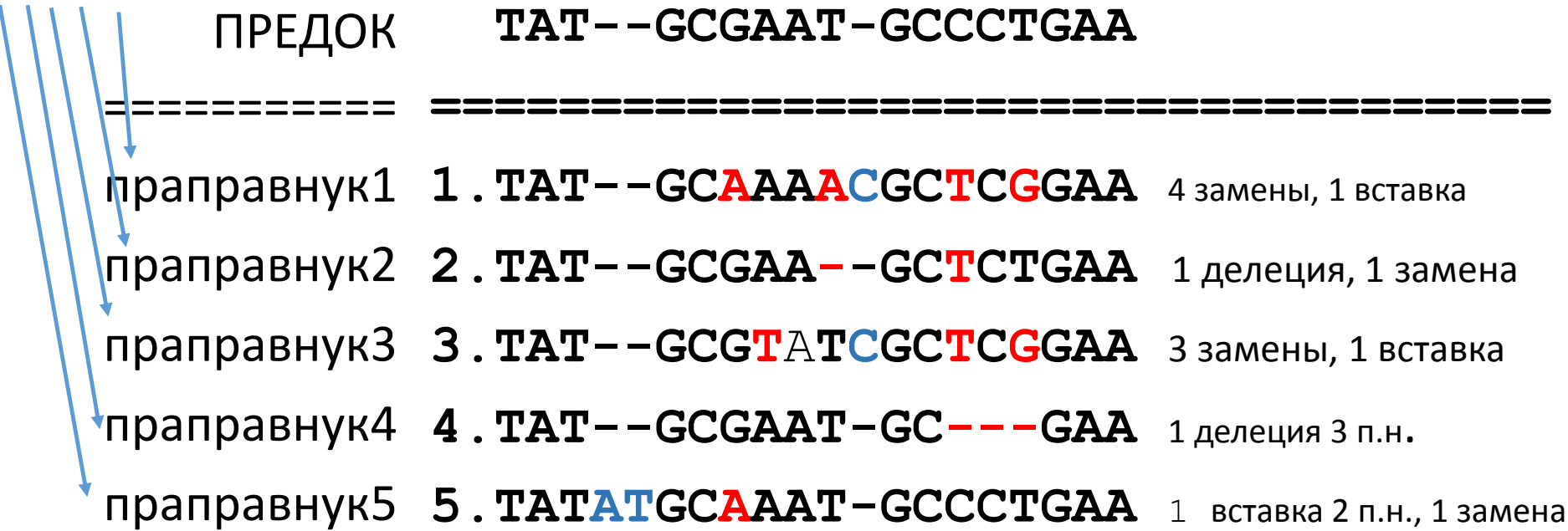
Выравнивание последовательностей потомков относительно предка

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 G CT C G G AA	замена
прапраправнук	6 .	TATGC A AA- C GC T C G GAA	делеция 1п.н.

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.

Биологический смысл выравнивания

Обсуждаем смысл колонок и гэпов

```
          10          20          30          40          50          60
EJL77459.1 GVDLVF GGPPCQGF SQIGMRR- LDDER- NEL YQQYTR I VAKLKPRVFLMENVPNLALMNKGH
RXK67093.1 DLDVVF GGPPCQGY SQIGTRR- LDDER- NEL YLQYAR I VEKQRPRMFLMENVPNMVL LNKGH
OJY44288.1 NVDLVF GGPPCQGY SQIGTRD- LHDPR- NRL FEEFARVVATLKP KLFMENVPNL LLLLNKGH
TRU90449.1 NPEMIV GSPPCQDF SSAGKRNEGLGR- - ANL TLTFAE I VTRVSPQWFVMENV D- - - RIEKSK
OXI46696.1 GTDLVF GGPPCQGF SQIGMRR- LDDER- NEL YKQYTRV VSTLRPRVFLMENVPNLALMNKGH
AVZ30243.1 EIDVVF GGPPCQGF SLIGKRS- FEDPR- NSL VFHY I RLVLELSPKFFV I ENVKGMTAGNHQA
AFZ12381.1 DIIGF I GGAPCPDF SVGGKNRGSEGDK- GKLSASY I EL I CQQKPDFFLFENVKGLYKTKKHR
HCQ21462.1 HIIGF I GGPPCPDF SVGGKNKGHLGDN- GKLSASY I EL I CQNLPDFFLFENVKGLWRTTKHR
EDN77159.1 SLIGF I GGPPCPDF S IAGKNKGKGDGN- GKLSLSYTNL I IEMKPDFFLFENVKGLWRTARHR
SOD91684.1 EVSLVV GGAPCQPF SNIGKKLGKNDERNGDL FLEFVRMVKGIQPEAF I FENVVGI TQNKHSD
QCS48280.1 NVVGF I GGPPCPDF S IGGKNRGRQGDH- GKLSesy I DL I IQHQPDFFLFENVKGLYRTKKHR
SMB95934.1 GLFG I I GGPPCPDF SVGGKNRGENGEQ- GRLSKVFVDK I LDLQP VFFLYENVPGL I RTAKHR
RUO38876.1 SPVGF I GGPPCPDF SVGGKNRGHEGEN- GRLTRTYVDG I IKYAPDFFLFENVKGLWRTKRHR
OIP70538.1 TIDL I CGPPCQGF STIGTND- KKDHR- NFL FFEFLRMVETFKPNF I ILENTGLLAKKNES
AFY60915.1 NLVGFV GGPPCPDF S IGGKNKGQYGDN- GKLTKVYVD I I IENQPDFFLFENVKGLWRTRSRHR
CUR30340.1 DLIGF I AGPPCPDF SVGGKNRGKNGDQ- GKLTACYVEL I CQQRPDFFLFENVKGLWSTKKHR
TAK03971.1 QAALVV GGAPCQPF SNL GSKRGTADSR- GTLFQDF I R I VKGVRPKGF I FENVEGLTQDKHKG
AEE51071.1 KVALVV GGAPCQPF SNIGKKEGENDAKNGDL FLEFVRMVKGIQPEAF I FENVAG I I QSKHSK
RTR31666.1 RLVGFV GGPPCPDF SVGGKNKGSEGEN- GKLTRTY I DL I VKDNPDYF I FENVKGLWRTTRHR
PTU64472.1 NIDLVF GGPPCQGF SQIGTRR- LDDER- NEL YKQYTR I VKTLKPRVFLMENVPNLAMMNKGH
```

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

```

      *           100           *           120           *           140           *           160
THIE_LACLS : AGVSFIVNDDVELARELNADGIHIGQTDESVSKVREKVGQEMWLGLSVTKADELKTAQ-SSGADYLGIGPIYPTNSKNDA : 14
THIE_MANSM : FQVPFIVNDDVELALSIQADGIHVGQKDTAVETILRNTRNKPIIGLSINTLAQALANKDRQDIDYFGVGPFIPTNSKADH : 15.
THIE_STRA3 : YQVPFIIDDDIDLVELIDADGLHIGQNDLPVDEARRRLPDKI-IGLSVSTMAEYQKSQ-LSVVDYIGIGPFENPTQSKADA : 14:
THIE_LISIN : YQVPFIINDDDVALALEIGADGIHVGQNDDEEIRQVIASCAGKMKIGLSVHVSVEAEEAERLGSVDYIIGVGPFIPTISKADA : 14.
THIE_ANOFW : YNIPFIVNDDVDLALALQADGVHVGQDEDEVAERVRDRIGDKY-LGVSVHNLNEVKKAL-AACADYVGLGPIFPTVSKEDA : 14:
THIE_GEOTN : YGVFPFIVNDDVELAIAIDADGVHVGQDDEDARRVREKIGDKI-LGVSAHNVEEARAAI-EAGADYIGVGPFIPTRSKDDA : 14:
THIE_BACSU : AGVPFIVNDDVELALNLKADGIHIGQEDANAKEVRAAIGDMI-LGVSAHTMSEVKQAE-EDGADYVGLGPIYPTETKKDT : 14:
THIE_BACA2 : AGIPFIIINDDDVELALRLBADGVHIGQDDADAEEETRAAIGDMI-LGVSAHNVESEVKRAE-AAGADYVGMGPVYPTETKKDA : 14:
THIE_OCEIH : FQIPFIIINDDDVDLAKQLDADGIHIGQDDQPVEVVRKQF PNKI-IGLSISTNNELNQSP-LDLVDYIIGVGFIFDNTKEDA : 14:
THIE_STAAB : YNVPFIVNDDVSLAKEINADGIHVGQDDAKVKEIAQYFTDKI-IGLSISDLGEYAKSD-LTHVDYIIGVGPFIPTPSKHDA : 14.
THIE_STACT : YNVPFIVNDDVALAEEDADGIHVGQDDEAVDDFNNRFEGKI-IGLSIGNLEELNASD-LTYVDYIIGVGFIFATPSKDDA : 14.
      6pFI6LDD6 La 6 ADG6H6GQ D 6G6S 2 DY G6GP pT 3K Da
  
```

```

      *           180           *           200           *           220           *           240
THIE_LACLS : AKPIGKDLR-LMLLENQLPIVGIGGITQDSLTELSAIGLDGLAVISLLTEAENPKKVAQMIRQKITKNG~~~~~ : 21
THIE_MANSM : SPLVGMNFIRQIRQLGIDKPCVAIGGIKEESAAILRRLGADGVAVISSAISHSVNIANTVKTLAQK~~~~~ : 22
THIE_STRA3 : KPAVENRTRKAVREINQDLEIVAIGGITSDFVHDIIESGADGIAVISAISRAMHVDATRQLREVERALVNRQRSDVI : 22:
THIE_LISIN : EPVSGTALLEEIRRAGIKLPIVGIGGINETNSAEVLTAGADGVSVISAITRSEDQSVIKQLKNPGSPS~~~~~ : 21
THIE_ANOFW : KQACGLTMIEHIRAHEKRVPLVAIGGITEQTAQVIEAGADGIAVISAICRAEHIYEQTKRRLYEMVMRAKQKGR~~~~~ : 21
THIE_GEOTN : NEAQPGLILRHLREQGITIPIVVAIGGITADNTRAVIEAGADGVSVISAIASAPEPKAAAAALATAVREANL---R~~~~~ : 22
THIE_BACSU : RAVQGVSLIEAVRRQGISIPVGIIGGITIDNAAPVIEAGADGVSMISAISQAEDPE SAARKFREEIQTYKTG--R~~~~~ : 22:
THIE_BACA2 : EAVQGVTLIEEVRRQGITIPVGIIGGITADNAAPVIEAGADGVSMISAISQAEDPKAAARKFSEEIRRSKAGLSR~~~~~ : 22
THIE_OCEIH : KTAGLEWIIQSLKKQHPSLPLVAIGGINTTNAQEIIQAGADGVSFISAITETHDHILQAVQRL~~~~~ : 20
THIE_STAAB : HTPVGPEMIATFKEMNPQLPIVVAIGGINTSNVAPIVEACANGISVISAISSKSENIEKTVNRFKDFNN~~~~~ : 21:
THIE_STACT : SEPVGPKMIETLRKEVGDLEIVAIGGISLDNVQEVAKTSADGVSVISAIARSPHVTETVHKFLQYFK~~~~~ : 21:
      G P V IGGI 6 galG6 I8a6
  
```

```

      *           20           *           40           *           60           *           80
THIE_LACLS : MTNKTLDLSVYFIAGAQNFSECSLDGATQKIALI---IKSGVTVYQFRDKG--TIIYKEQKQRLSTAKLQKVSEEAG :
THIE_MANSM : MNKIKSMLSVYFIAGSQDCRHLPGEPTENLLTILQRALEAGITCFQFREKGEQSLACDLQLKRLALKCLQLCRQFQ :
      M LSVYFIAG Q1 6 66 6 G6T 5QFR KG 36 4 6A K 6 2
  
```

```

      80           *           100           *           120           *           140           *
THIE_LACLS : VSFIVNDDVELARELNADGIHIGQTDESVSKVREKVGQEMWLGLSVTKADELKTAQSSGADYLGIGPIYPTNSKND :
THIE_MANSM : VEFIVNDDVELALSIQADGIHVGQKDTAVETILRNTRNKPIIGLSINTLAQALANKDRQDIDYFGVGPFIPTNSKAD :
      V FIVNDDVELA 6 ADGIH6GQ D V 6 6GLS6 T A L DY G6GPI5PTNSK D
  
```

```

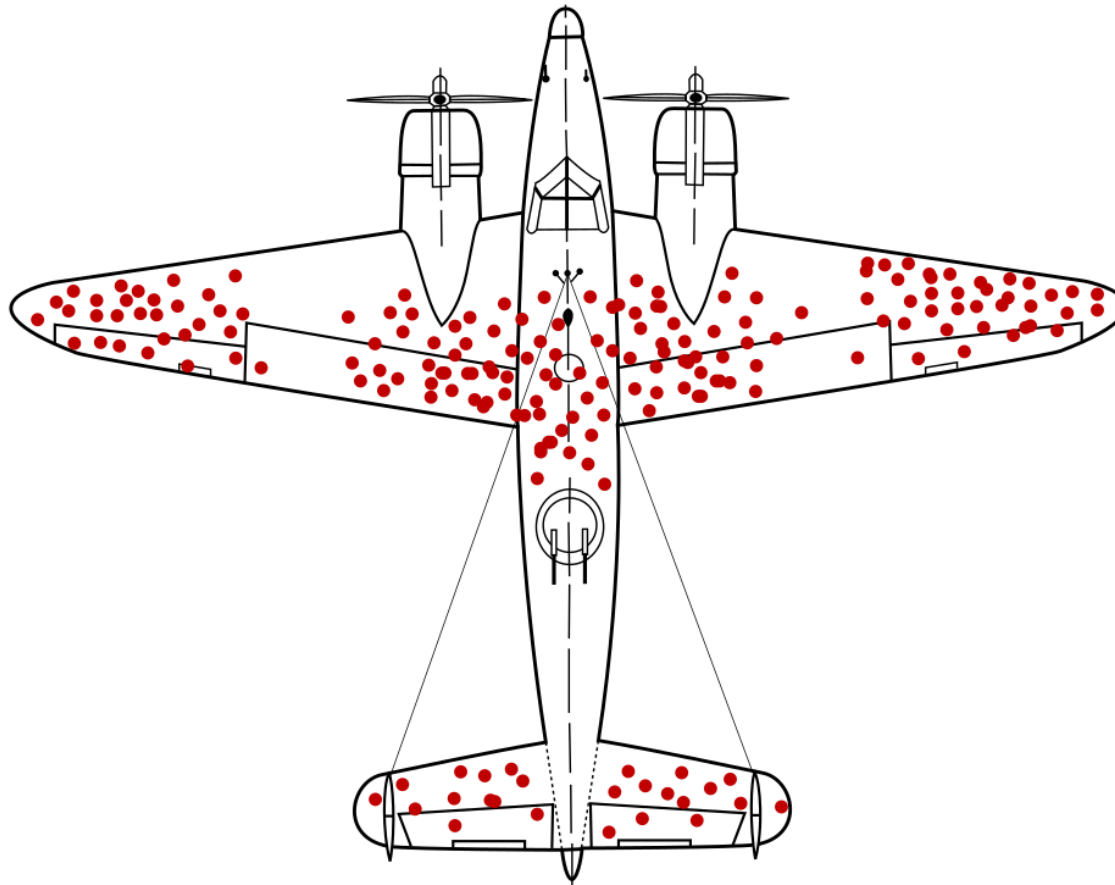
      160           *           180           *           200           *           220
THIE_LACLS : AAKPIG---IKDLRLMLLENQLPIVGIGGITQDSLTELSAIGLDGLAVISLLTEAENPKKVAQMIRQKITKNG : 218
THIE_MANSM : HSPVGMNFIRQIRQLGIDK--PCVAIGGIKEESAAILRRLGADGVAVISSAISHSVNIANTVKTLAQK----- : 220
      6G 14 6R 6 6 P V IGGI 2 S L 6G DG6AVIS 63 N 6 QK
  
```

3. Принцип Вальда

(так называл МГ)

Пробоины на вернувшихся американских самолётах во время второй мировой войны.

Какие части укреплять броней?

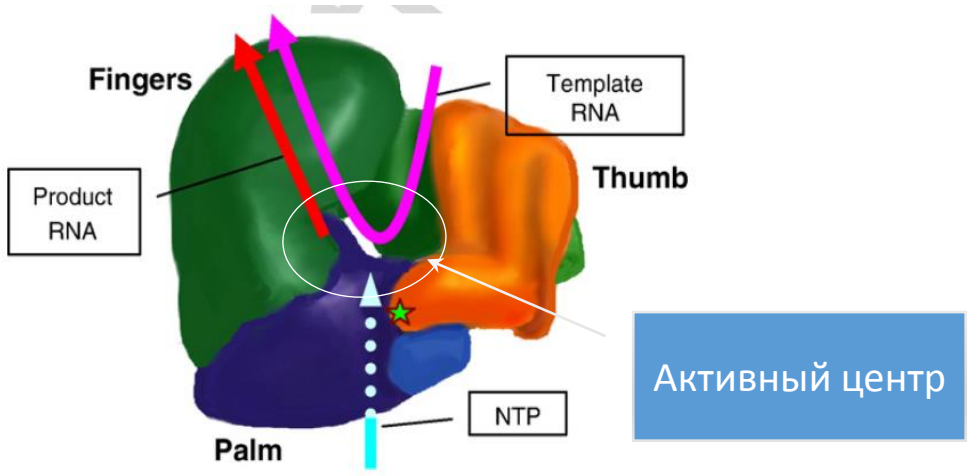


Сохраняющееся (консервативное)
– важно

РНК зависимая РНК полимераза (RdRP), консервативные участки

```

*      320      *      340      *      360      *      380      *      400      *      420      *      440      *      460
FKTMIRFGDVGLLDIFSAFDASLSPPFMIREA..GRIMSELS...GTPSHFGTALINTIIYSKHLIYNCCY.....HVCGSMPSGSECTALINSTINNVLYYVFSKIFGKSPVFF.....CQALKILC.YGDDVIVFVSRD
EVAMQG.FERVYDVVYSNEDSTHSVAMFRLL..A...EEFF.TPENGFDPLTREYLESIAISTHAFEKRF.....LTGGLPSCCAATSMINTIMNNIIIRAGLYLTYNFEFDD.....VKVLS.YGDDLVATNYQL
ETHFAQ.YKNVWVDVLYSADANHCSDAMNMFEEVFERTEFG.....FHPNAEWILKTLVNTTEHAYENKRI.....VVEGCMPSGCSATSIINTILNNIYVLYALRRHYEGVELDT.....YTMIS.YGDDIVVASYDYL
.....WSLCVATIVSDHDTFWPGWLRDLICDELINMGYA.PWVVKLFETSLKLPVYVGAFAPEQGHLLGDPSPNDLEVGLSSGQGATDLMGTLIMSTIYLVMLQDHTAPHLNSRIKDMPSACRFLDSYWQGHEETROIS.KSDDAILGWTKGR
LRLRLE.NWVYCADAGSQEDSSSLTPYLINAV..LTLRSTYMEDWDVGLQMLRNLYTEIVYTPISTPDGTIV.....KKFRGNNSGQPSIVDNLSLMVVIAMHYALIKECFEVEEID.....STCVFFV.NGDDLIIVAVNPEK
HDKLNRPGLWLGSGDGRDSSIDPFFFDVV..KTKRKHEL..PSEHHRaidLIYDEILNTTICLANGMVI.....KKNVGTQR.QPSTVDNTLVMITAFLYAYIHKTGDRELAL.....LNERFIFVC.NGDDNKFAISPQF
AISLASFSFYGFNCFANEDGMFHPSSFSMV..SELANIFY...GNFLSTERDNLTRMLTNRFSLMKGAIL.....RVPGGSPSGFFMTVFNSEINLFYLQSAWIMLARFNGRQDISH.....PCNFPKYVRACV.YGDDNIVAIMEV
AARMKEKGNVDVLCODYSSEFDGLLSKQVMDVI..ASVINELC.GGEDQLKNARRNLLMACCSRIAICKNTVW.....RVECGIPSGFFMTVFNSEINLFYLRHYHKIMREQQAPELMV.....QSFDKLIGLVT.YGDDNLSVNAV
YAEHAK.YKNHFDADYIADSTQNRQIMTES..FSIMSRLT...ASPELAEVVAQDLAPSEMVDG DYVI.....RVKEGLPSGFFPCTSQVNSINHWITLCALEATGLSPDVV.....QMSYFYSFYGDDIVSTDIDF
NNLTSKASDFLCLDYSKFDSTMSPCVVRIA..IDLADCC...EQTELTKSVVLTILKSHFMTILAMIV.....QTKRGLPSGMPFTSVINSHCHWLLWSAAVYKSCAEIGLHCS.....NLYEDAPFYT.YGDDGVYAMTEMM
IQRIKS.AAKVYAVDYSKFDSTQSPRVSAAAS..IDLRYFS...DRSPIVDSAANTLKSPPIAIFNGVAV.....KVSSEGLPSGMPFTSVINSHCHWLLWSAAVYKSCAEIGLHCS.....NLYEDAPFYT.YGDDGVYAMTEMM
TKRLERPKHdryCVLYSKFDSTQPPKVTSSQS..IDILRHFT...DKSPIVDSACATLKSNPIGIFNGVAF.....KVAGGLPSGMPFTSVINSHCHWLLWSAAVYKSCAEIGLHCS.....NLYEDAPFYT.YGDDGVYAMTEMM
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

Каталитические мотивы RdRP

	Motif A			Motif B			Motif C		
	*		*	*		*		*	
Nido-/Arteri-	442	LETDLLESCDRSTP	454 [44]	499	GLSSGDPITTSISNTI	513 [40]	554	RVYIYSDVVVLT	565 [128]
Nido-/Corona-	615	MGWDYPKCDRAMP	627 [51]	679	GTSSGDATTAYANSV	693 [60]	754	SMMILSDDAVVC	765 [167]
Nido-/Alyss-	-	CGGDYKEYDKNLA	- [54]	-	GNTSGNSRRTKTVNGN	- [64]	-	RMVVCVGGDYIKV	- [-]
Nido-/Euroni-	-	VSLDHSKFDRFVA	- [52]	-	GISGNSITALNNSL	- [50]	-	RIAGLSDDVVAC	- [-]
Nido-/Medioni-	-	SGKDFPQWDRSVE	- [56]	-	GVCSSGNSKTAPGNSI	- [60]	-	LLRVLSDDGMLV	- [-]
Nido-/Mesoni-	-	GGKDYPKWDRRIS	- [58]	-	GVTSGNSRTADGNSL	- [66]	-	KGAYLSDDGLIV	- [-]
Nido-/Mononi-	-	FSFDYTAFDRTTT	- [53]	-	SVSSGNAHTAPWNHS	- [76]	-	SIQIIGDDLITN	- [-]
Nido-/Roni-	-	ISQDYPKFDTCVD	- [50]	-	GVSSGDGATAIKNSH	- [56]	-	RCATLSDDTLAI	- [-]
Nido-/Tobani-	-	MGADYTKCDRSFP	- [47]	-	GTTSGDSTTAFNSNF	- [57]	-	FLHFLSDDSFII	- [-]
Picornia-/Dicistro-	286	IAGDFSTFDGSLN	298 [48]	347	SQPSGNPATTPLNCF	361 [30]	392	SMVSYGDDNVIN	403 [143]
Picornia-/Ifla-	252	LQMDYKNYSDAIP	264 [52]	317	GVLAGHPMTSVVNSV	331 [25]	357	YIIVMGDDVVIS	368 [271]
Picornia-/Picornia-	230	FAPDYTYGDASLS	242 [42]	285	GMPSGCSGTISFNMS	299 [22]	322	KMIAYGDDVIAS	333 [128]
Picornia-/Seco-	277	LCCDYSSFDGLLS	289 [48]	338	GIPSGFPMTVIVNSI	352 [31]	384	GLVTYGGDNLIS	395 [316]
Picornia-/Polycipi-	-	VDFDVSNWDGFLF	- [50]	-	GIIISGFPGTAEVNTL	- [29]	-	SAILYGDDILLT	- [145]
Picornia-/Marna-	-	IAGDYSSFDMSHN	- [52]	-	WVMSGVPLTAELSST	- [25]	-	ALIVYGDDNNA	- [-]
Tymo-/Tymo-	316	IANDYTAFDQSQH	328 [40]	369	MRLTGEFPGTYDDNTD	383 [15]	399	PIMVSGDDSLID	410 [185]
Tymo-/Alphaflexi-	-	LANDYTAFDQSQD	- [40]	-	MRLTGEFPTFDANTE	- [16]	-	AQVYAGDDSALD	- [122]
Tymo-/Betaflexi-	-	TDSDYEAFFDRSQD	- [40]	-	MRFSGEFPTFFFNTI	- [16]	-	PICFAGDDMYSP	- [140]
Tymo-/Deltaflexi-	-	TGNDYTAWDSGID	- [40]	-	RQESGDRWTWLLNTL	- [16]	-	PLCVSGDDSVTL	- [135]
Tymo-/Gammaflexi-	-	TDGDYTAIDASQD	- [40]	-	MRFSGEVWTYLFTL	- [15]	-	AQVYGGDDKSIN	- [131]
-/Alphatetra-	1076	KSIDIKEFDTVHN	1088 [43]	1132	MLDSCAVWTIARNTL	1146 [14]	1161	FIAAKGDDVFLA	1172 [753]
-/Astro-	266	IEFDWTRYDGTIP	278 [51]	330	GNPSSGFSTPMDNNM	344 [24]	369	DTVVYGGDRLST	380 [138]
-/Barna-	305	CETDLSGWDWSVQ	317 [53]	371	GQLSGDYNTSSSNRS	385 [22]	408	GIKAMGDDSFEI	419 [104]
-/Beny-	297	GVIDAAACDSGQG	309 [44]	354	VKTSCEPGTLLGNTI	368 [16]	385	CMAMKGDDGFKR	396 [172]
-/Botourmia-	404	VSGDYSAAATDNLH	416 [58]	475	GQMLSGSPLSFPVLCI	489 [23]	513	GILVNGDDILFR	524 [336]
-/Bromo-	462	LEADLSKFDKSSQG	474 [46]	521	QRRTGDAFTYFGNTL	535 [16]	552	CAIFSGDDSLII	563 [259]
-/Calici-	239	YDADYSRWDSTQQ	251 [45]	297	GLPSGVPECTSQWNSI	311 [25]	337	LFSFYGGDEIVS	348 [162]
-/Carmotetra-	582	ISFDLSRWDHMQ	594 [44]	639	GIMSGDMTTLGLNCI	653 [63]	717	SILDDGDDHVII	728 [187]
-/Clostero-	277	LEIDFSKFDKSSQG	289 [46]	336	QRRTGSPNTWLSNTL	350 [16]	367	LLVSDGDDSLIF	378 [137]
-/Flavi-/Flavi-	532	YADDTAGWDTGRIT	544 [55]	600	QRSGSQVPTYALNTI	614 [46]	661	RMVAVSGDDCVVR	672 [233]
-/Hepaci-	217	FSYDTRCFDSTVT	229 [49]	279	CRASGVLTTSCGNTL	293 [18]	312	TMLVCGDDLVVI	323 [268]
-/Pegi-	211	ICVDATCFDSSIT	223 [45]	269	CRSSGVLTTASANCL	283 [18]	302	SLLIAGDDCLII	313 [250]
-/Pesti-	342	VSFDTKAWDTQVT	354 [47]	402	QRSGSQPDTTSAGNSM	416 [25]	442	RIHVCGDDGFLI	453 [266]
-/Hepe-	256	YENDFSAFDSTQN	268 [44]	313	KKHSGEPGTMFLNTI	327 [16]	344	LALFKGDDSLVC	355 [129]
-/Kita-	901	YEFDMSKYDKSSQG	913 [46]	960	QRKSGDASTYFQNTV	974 [16]	991	FGAFSGDDSLIF	1002 [142]
-/Levi-	271	ATVDLSAASDSIS	283 [36]	320	ISSMNGYTYFLES	334 [18]	353	EVTVYGGDIILP	364 [225]
-/Luteo-	229	IGVDASRFDQHVS	241 [47]	289	HRMSGDINTSMGNKL	303 [17]	321	ELCNNGDDCVII	332 [200]
-/Narna-	355	ISSDMKSASDLIP	367 [44]	414	GILMGLPTTWAILNL	428 [26]	455	DCRVCGDDLIGV	466 [363]
-/Noda-	591	SEGDFSNFDGTVS	603 [49]	653	GKSGSPTTCDLNTV	667 [25]	693	IGLAFGDDSLFV	704 [339]
-/Permutotetra-	366	ICPDFKQMDGSVD	378 [56]	435	GLMTGVVGTTLFDTV	449 [-103]	345	RIACYGDDTDIY	356 [901]
-/Poty-	245	CDADGSGQFDDSSLT	257 [49]	307	GNNSGQPSVVVDNSL	321 [24]	346	WFFVNGDDLLIA	357 [162]
-/Solemo-	286	AEADISGFDWSVQ	298 [51]	350	IMKSGSYCTTSTNSR	364 [12]	377	WCIAMGDDSVEG	388 [165]
-/Solinv-	485	FSCDYKNFDRITIP	497 [45]	543	GMPSGCVPTAPLNSK	557 [32]	590	CRLFYGDDVIIA	601 [195]
-/Toga-	373	LETDIASFDDKSDQ	385 [46]	432	MMKSGMFLTTFVNTV	446 [18]	465	CAAFIIGDDNIH	476 [140]
-/Tombus-	527	IGLDASRFDQHCS	539 [48]	588	CRMSGDINTSLGNL	602 [18]	621	SLANCGDDCVLI	632 [186]
-/Virga-	230	IEIDISKYDKSKT	242 [46]	289	QKSGSNVDTYFSNTW	303 [16]	320	FSIFGGDDSLIL	331 [146]
-/Alverna-	-	CSSDASGWDMSVS	- [57]	-	ITASGLPDTTQNSF	- [12]	-	KALTAGDDLLCD	- [112]
-/Matona-	-	IEVDFTFEFDMNQT	- [44]	-	ERTSGEPATLLHNT	- [16]	-	AGIFQGGDMVIF	- [144]
-/Hypo-	-	TAMDVTAMDSTAS	- [53]	-	GLSTGHATTPSNTT	- [25]	-	KFSSFSDDNFWS	- [-]

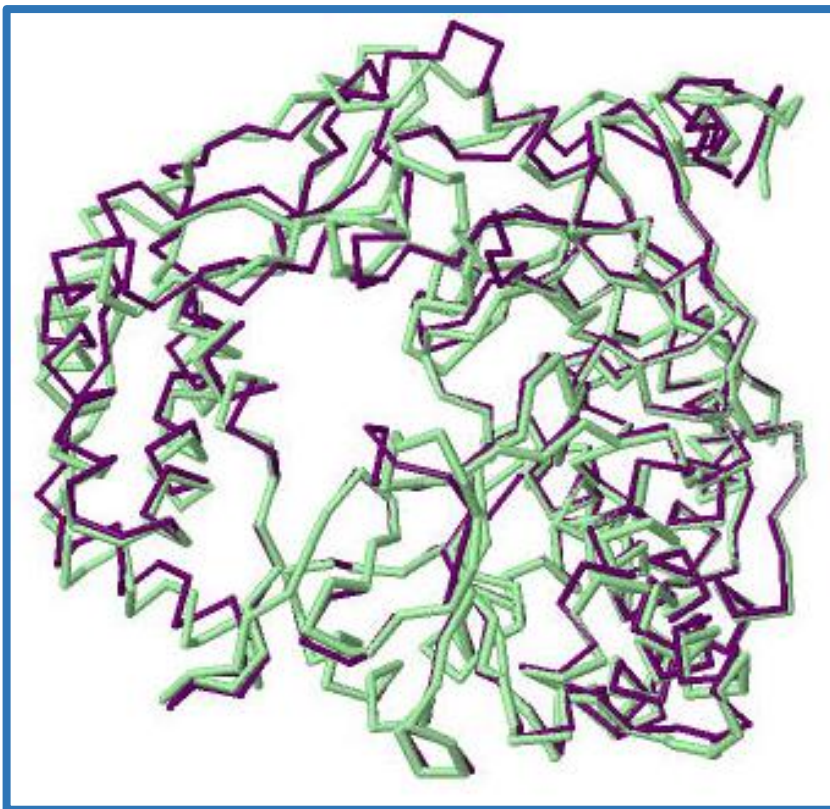
The RdRP active site is surrounded by the palm, fingers, and thumb domains with seven catalytic motifs (motifs A-G) distributed within the palm (motifs A-E) and fingers (motifs F-G) (Poch et al., 1989; Gorbalenya et al., 2002; Bruenn, 2003; te Velthuis, 2014; Wu et al., 2015) (see an alignment of motif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

4. Выравнивание и совмещение структур

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

Проверка выравнивания по совмещению полипептидных остовов 3Dструктур

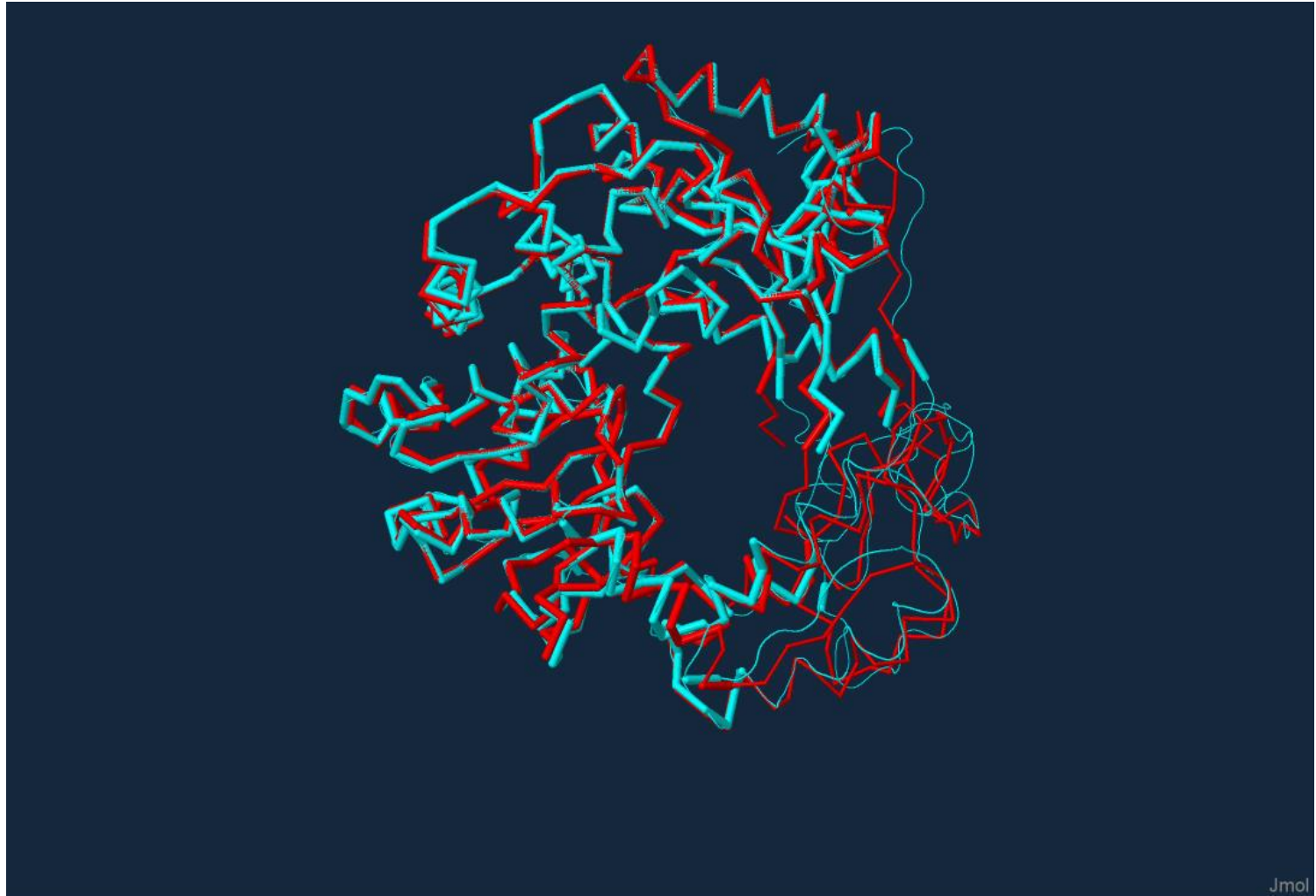
Структура консервативнее последовательностей



Совмещение полипептидных цепей
двух RdRP +РНК вирусов.

Не успел посмотреть каких. Выбрал случайно
из совмещения пяти структур

Совмещение 2х RdRp



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

Как проверять выравнивание при отсутствии 3D структур?

Варианты

- Выделение блоков надежного выравнивания 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 вообразить 3D структуру. Гэпы преимущественно в петлях между спиралями и тяжами бета-листов

Программа-выравниватель лишь инструмент

- Программа-выравниватель выдает ГИПОТЕЗУ об эволюционном выравнивании
Гипотеза может быть верна в отдельных блоках и не верна в остальных блоках 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 ...

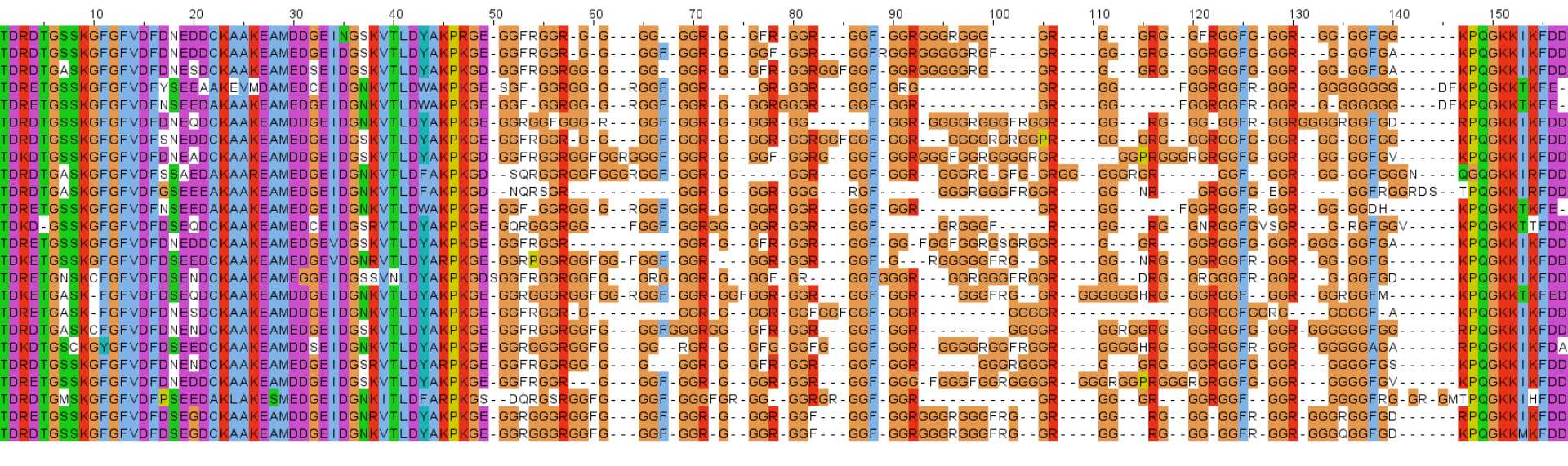
Это выравнивание не может быть правильным

```
      860      870      880      890      900      910      920      930      940      950
-----E-D---H--P---V---L---L-----T--E-A---P-----L-----N-----P-----K---A---N--R--E-R---
tsaldrsH-H---H--P---L---L---M-----L--D-K---P-----H-----T-----P-----P---P---L--R--Q-R---
-----E-D---H--Y---F---L---L-----T--E-P---P-----L-----N-----T-----P---E---N--R--E-Y---
-----S-E---N--P---L---L---I-----T--E-S---T-----I-----N-----P-----K---Q---N--R--E-R---
-----E-D---H--A---I---L---L-----T--E-Q---P-----L-----N-----I-----R---S---N--R--E-R---
-----E-E---H--P---V---L---L-----T--E-A---P-----L-----N-----P-----K---A---N--R--E-K---
-----I-E---H--P---I---L---L-----A--E-P---T-----Y-----N-----T-----R---A---I--R--E-K---
-----I-N---H--P---V---V---L-----T--E-A---V-----C-----N-----P-----S---P---A--R--Q-F---
-----E-E---N--P---C---L---L-----T--E-A---P-----L-----N-----P-----K---A---N--R--E-K---
-----E-D---H--Y---F---L---L-----T--E-P---P-----L-----N-----P-----P---E---N--R--E-Y---
-----E-E---H--P---V---L---L-----T--E-A---P-----L-----N-----P-----K---G---N--R--E-R---
-----E-E---H--P---V---L---L-----T--E-A---P-----L-----N-----P-----K---A---N--R--E-K---
-----K-D---R--R---V---V---V-----A--E-S---V-----L-----C-----P-----T---L---F--R--N-T---
-----E-E---S--D---I---L---L-----T--E-P---P-----L-----T-----S-----K---N---D--H--N-K---
-----T-D---R--P---V---I---F-----V--E-S---I-----F-----M-----S-----S---L--R--N-T---
-----E-E---H--P---V---L---L-----T--E-A---P-----L-----N-----P-----K---A---N--R--E-K---
-----R-E---R--R---L---L---V-----V--E-E---I-----N-----E-----P-----R---E---F--R--E-R---
vlenaeqD-P---L--C---L---V---L-----P--E-I---W-----H-----E-----R-----I---D---V--M--E-A---
-----E-E---H--P---V---L---L-----T--D-A---P-----L-----N-----P-----L---K---N--R--E-R---
-----T-E---H--P---V---L---L-----T--E-A---P-----L-----N-----P-----K---A---N--R--E-K---
-----E-E---H--P---S---L---L-----T--E-A---P-----L-----N-----P-----K---A---N--R--E-K---
-----E-N---H--P---V---L---L-----T--Q-P---P-----C-----N-----P-----K---Q---N--R--E-K---
-----E-E---H--P---V---L---L-----T--E-A---P-----L-----N-----P-----K---A---N--R--E-K---
-----E-D---H--P---V---L---L-----T--E-A---P-----L-----N-----P-----R---K---N--R--Y-R---
-----E-E---H--P---V---L---L-----T--E-A---P-----L-----N-----Q-----R---K---N--R--D-Q---
-----S-E---T--G---L---I---L-----T--E-A---P-----N-----A-----L-----P---Q---L--Q--T-N---
-----H-L---H--P---V---L---M-----S--E-A---A-----W-----N-----A-----R---T---K--R--E-K---
```

~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?



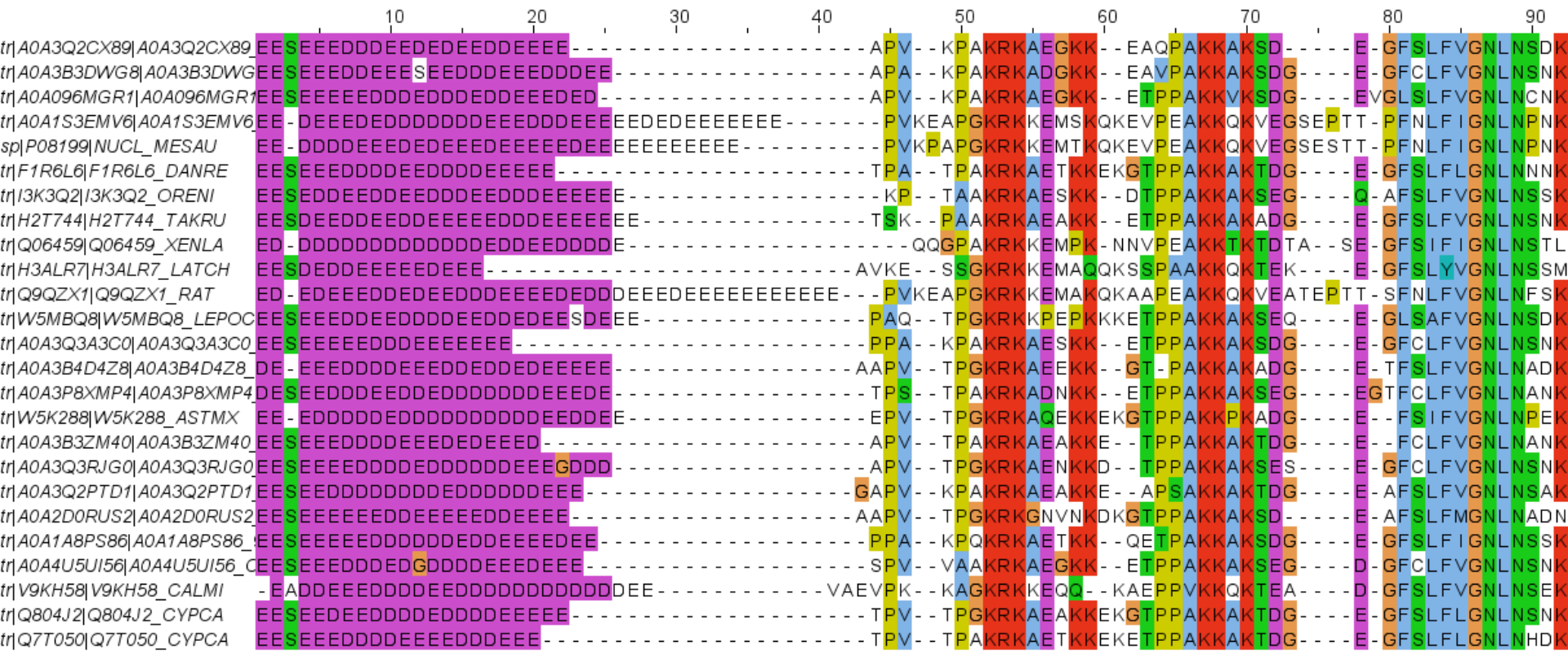
Предположительно был, так как он есть у нуклеолинов всех ныне живущих эукариот.

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 дубликации повторов.

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

Nucleolin. Кислый участок и 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

```
SW: PMX1_CHICK/1 : -----*-----20-----*-----40-----*-----60-----*-----80-----*-----100-----*-----120-----*-----140-----*-----1 : 80
SW: PMX2_HUMAN/1 : -----*-----20-----*-----40-----*-----60-----*-----80-----*-----100-----*-----120-----*-----140-----*-----1 : 86
SW: PMX1_HUMAN/1 : -----*-----20-----*-----40-----*-----60-----*-----80-----*-----100-----*-----120-----*-----140-----*-----1 : 80
SW: ARX_BRARE/1- : ISQAPQVVISRSKSYREN-APFSQS---D-EGQSP--EHMAQLVELST-----LKFEEDEVVKEEACQDN-----S-----LSPKDEESLH-NDGVDKCDSDSVCLS : 84
SW: ARX_MOUSE/1- : ISQAPQVVISRSKSYRENGCAPVFPVPPALD-ELSGPCGVAHPERERLSAASGPGSAPAAAGCGTCAEDDEEELLEDEDEEEEEELEDDDEELLEDDARALLKEPERRCVATTCTVA-----AVATEGGELSPEKRELLLHPEDARCKDGEDSVCLS : 157
SW: AL_DROME/1-1 : -----*-----20-----*-----40-----*-----60-----*-----80-----*-----100-----*-----120-----*-----140-----*-----1 : 72
SW: ALX4_MOUSE/1 : -----*-----20-----*-----40-----*-----60-----*-----80-----*-----100-----*-----120-----*-----140-----*-----1 : 145
SW: ALX4_HUMAN/1 : -----*-----20-----*-----40-----*-----60-----*-----80-----*-----100-----*-----120-----*-----140-----*-----1 : 157
SW: RX2_CHICK/1- : -----*-----20-----*-----40-----*-----60-----*-----80-----*-----100-----*-----120-----*-----140-----*-----1 : 83
SW: RX2_BRARE/1- : -----*-----20-----*-----40-----*-----60-----*-----80-----*-----100-----*-----120-----*-----140-----*-----1 : 92
SW: RX1_XENLA/1- : -----*-----20-----*-----40-----*-----60-----*-----80-----*-----100-----*-----120-----*-----140-----*-----1 : 91
SW: RX_HUMAN/1-1 : -----*-----20-----*-----40-----*-----60-----*-----80-----*-----100-----*-----120-----*-----140-----*-----1 : 97
SW: PIX2_BRARE/1 : -----*-----20-----*-----40-----*-----60-----*-----80-----*-----100-----*-----120-----*-----140-----*-----1 : 66
SW: PIX2_HUMAN/1 : -----*-----20-----*-----40-----*-----60-----*-----80-----*-----100-----*-----120-----*-----140-----*-----1 : 68
SW: PIX1_HUMAN/1 : -----*-----20-----*-----40-----*-----60-----*-----80-----*-----100-----*-----120-----*-----140-----*-----1 : 64
SW: OTP_MOUSE/1- : -----*-----20-----*-----40-----*-----60-----*-----80-----*-----100-----*-----120-----*-----140-----*-----1 : 90

SW: PMX1_CHICK/1 : NDQLNSEE-----KTKRQRNRRTFTFNSSQLQALERWERETHYDPAFVRBDLARRVNIETARVQVVFQNRRAKFRNNEAMLASKNASLLKSYSGDVTAVEQPIVPRPAPRPTDYLSWGTASPYSAMATYSTTCTMAS----- : 213
SW: PMX2_HUMAN/1 : GCPCSPGRCG-----AAKRKKQRNRRTFTFNSSQLQALERWERETHYDPAFVREELARRVNIETARVQVVFQNRRAKFRNNEAMLASRSASLLKSYSGEAAETQVPAVPRPTALSPDYLSWTASSPYSTVPVFPYSGSGCP----- : 221
SW: PMX1_HUMAN/1 : NDQLNSEE-----KTKRQRNRRTFTFNSSQLQALERWERETHYDPAFVRBDLARRVNIETARVQVVFQNRRAKFRNNEAMLANKNASLLKSYSGDVTAVEQPIVPRPAPRPTDYLSWGTASPYSAMATYSTATCANMS----- : 213
SW: ARX_BRARE/1- : ACGDSEEG-----HLKRRQRNRRTFTFTSYQLEELERAFQKTHYDPAFVREELARRLDLDEARVNIETARVQVVFQNRRAKWRNREARCAQTHPPGLFPFPCPLSATHPSPYLDASPPPHHPALDSAWTAAAAAAFPPLPPP- : 230
SW: ARX_MOUSE/1- : ACGDSEEG-----LLKRRQRNRRTFTFTSYQLEELERAFQKTHYDPAFVREELARRLDLDEARVNIETARVQVVFQNRRAKWRNREARCAQTHPPGLFPFPCPLSATHPSPYLDASPPPHHPALDSAWTAAAAAAFPPLPPP- : 303
SW: AL_DROME/1-1 : ECRADETY-----PKRQRNRRTFTFTSYQLEELERAFQKTHYDPAFVREELARRLDLDEARVNIETARVQVVFQNRRAKWRNREARCAQTHPPGLFPFPCPLSATHPSPYLDASPPPHHPALDSAWTAAAAAAFPPLPPP- : 212
SW: ALX4_MOUSE/1 : EKTDSSEN-----KCKRQRNRRTFTFTSYQLEELERAFQKTHYDPAFVREELARRLDLDEARVNIETARVQVVFQNRRAKWRNREARCAQTHPPGLFPFPCPLSATHPSPYLDASPPPHHPALDSAWTAAAAAAFPPLPPP- : 290
SW: ALX4_HUMAN/1 : EKADSESN-----KCKRQRNRRTFTFTSYQLEELERAFQKTHYDPAFVREELARRLDLDEARVNIETARVQVVFQNRRAKWRNREARCAQTHPPGLFPFPCPLSATHPSPYLDASPPPHHPALDSAWTAAAAAAFPPLPPP- : 302
SW: RX2_CHICK/1- : KPDSREQ-----PKRQRNRRTFTFTSYQLEELERAFQKTHYDPAFVREELARRVNIETARVQVVFQNRRAKWRNREARCAQTHPPGLFPFPCPLSATHPSPYLDASPPPHHPALDSAWTAAAAAAFPPLPPP- : 215
SW: RX2_BRARE/1- : PDIPDEDQ-----PKRQRNRRTFTFTSYQLEELERAFQKTHYDPAFVREELARRVNIETARVQVVFQNRRAKWRNREARCAQTHPPGLFPFPCPLSATHPSPYLDASPPPHHPALDSAWTAAAAAAFPPLPPP- : 225
SW: RX1_XENLA/1- : KLSDDDEQ-----PKRQRNRRTFTFTSYQLEELERAFQKTHYDPAFVREELARRVNIETARVQVVFQNRRAKWRNREARCAQTHPPGLFPFPCPLSATHPSPYLDASPPPHHPALDSAWTAAAAAAFPPLPPP- : 224
SW: RX_HUMAN/1-1 : KLSDEEQ-----PKRQRNRRTFTFTSYQLEELERAFQKTHYDPAFVREELARRVNIETARVQVVFQNRRAKWRNREARCAQTHPPGLFPFPCPLSATHPSPYLDASPPPHHPALDSAWTAAAAAAFPPLPPP- : 242
SW: PIX2_BRARE/1 : SKNEDSW-----DDPSKRRQRNRRTFTFTSYQLEELERAFQKTHYDPAFVREELARRVNIETARVQVVFQNRRAKWRNREARCAQTHPPGLFPFPCPLSATHPSPYLDASPPPHHPALDSAWTAAAAAAFPPLPPP- : 212
SW: PIX2_HUMAN/1 : GKNEDVGA-----EDPSKRRQRNRRTFTFTSYQLEELERAFQKTHYDPAFVREELARRVNIETARVQVVFQNRRAKWRNREARCAQTHPPGLFPFPCPLSATHPSPYLDASPPPHHPALDSAWTAAAAAAFPPLPPP- : 215
SW: PIX1_HUMAN/1 : KCPEDSCAGCTGCCGADDPAKRRQRNRRTFTFTSYQLEELERAFQKTHYDPAFVREELARRVNIETARVQVVFQNRRAKWRNREARCAQTHPPGLFPFPCPLSATHPSPYLDASPPPHHPALDSAWTAAAAAAFPPLPPP- : 218
SW: OTP_MOUSE/1- : NPSQACQQ-----CQCKRQRNRRTFTFTSYQLEELERAFQKTHYDPAFVREELARRVNIETARVQVVFQNRRAKWRNREARCAQTHPPGLFPFPCPLSATHPSPYLDASPPPHHPALDSAWTAAAAAAFPPLPPP- : 236
      k 4 4R RT Ft QL EL E R 4 YTPD RE 6A L E R6qVVFQNRRAK54 e4

SW: PMX1_CHICK/1 : -----*-----320-----*-----340-----*-----360-----*-----380-----*-----400-----*-----420-----* : 245
SW: PMX2_HUMAN/1 : -----*-----320-----*-----340-----*-----360-----*-----380-----*-----400-----*-----420-----* : 253
SW: PMX1_HUMAN/1 : -----*-----320-----*-----340-----*-----360-----*-----380-----*-----400-----*-----420-----* : 245
SW: ARX_BRARE/1- : LGTFLGTAAMFRHPAFIGTFCRGLFSSMCLPTSASTAAALLRQTAPPVPSVPQSAALPEPPSSSSSTAADRRASSTAAALPLRKAHDSA-QLTQLNLIPSGTACKKEVC----- : 336
SW: ARX_MOUSE/1- : LGTFLGAAVFRHPAFIFSPAGRLFTMAPLTSAASTAAALLRQTAPPVAVGASGALADP-----ATAAADRRASSTAAALPLRKAHQAQITLQNLNLPGTSTCKKEVC----- : 404
SW: AL_DROME/1-1 : PPTSPASGAXPQLVGLIALTQQASSLSPT---QTSFVALTLSSHSPRQLPPSHQAPPPPPAAATPPEDRRRTSSIAALPLRKAHDELKLELLRQNGHCNDV-----VS : 313
SW: ALX4_MOUSE/1 : DFL-----SVSGACSHVQTHMCSLFGAACISPLNGCYELNCEPDRKTSIAALPLRKAHDSAAISWAT----- : 354
SW: ALX4_HUMAN/1 : DFL-----SVSGACSHVQTHMCSLFGAACISPLNGCYELNCEPDRKTSIAALPLRKAHDSAAISWAT----- : 366
SW: RX2_CHICK/1- : LPASYTTPPPFL-----NSPVAHTHALQPLGAMGPPPPYQCGAAVDFKPLDEGDPNRTSSIAALPLRKAHDSHIQSGKPWQTI----- : 290
SW: RX2_BRARE/1- : LQPTYTAHPGFL-----NTSPGMHNTQIQLM---PPPPYQCPVVFNDKYPLEEDVD---RSSSTAAALPLRKAHDSHIQSGDKTQWPM----- : 297
SW: RX1_XENLA/1- : LPSGYTTPPPFI-----NPVSVGHALQPLGAMGPPPPYQCGAANFDKYPLEEDVDPMNSIASLPLRKAHDSHIQFICKPW--- : 296
SW: RX_HUMAN/1-1 : LPASYTTPPPPPFL-----NSPPLCPGLQPL---APPPSYFCGPFGDKRPLLEADPMNSIASLPLRKAHDSHIQAIQKFPWQAL----- : 319
SW: PIX2_BRARE/1 : SISSMSSMSSMVPASVTCVPGSSL-----NSLNNLNLNSPLNSAVTTPACPYAPPTPPY-VYRDTCNSSLASLPLRKAHDSHSGFYASVQNPASNLNLSACQYAVDRPV : 314
SW: PIX2_HUMAN/1 : SISSMSSMSSMVPASVTCVPGSSL-----NSLNNLNLNSPLNSAVTTPACPYAPPTPPY-VYRDTCNSSLASLPLRKAHDSHSGFYASVQNPASNLNLSACQYAVDRPV : 317
SW: PIX1_HUMAN/1 : SISMTMPSSMCPGAVPGMNSCL-----MNIN---MLTGSSLNSAMSFCPCPYCTPASPYVYRDTCNSSLASLPLRKAHDSHSGFYCGGLCPASGLNACQYNS----- : 314
SW: OTP_MOUSE/1- : SQCSLAAGPPPNMCLNSLNSLACNGACLQ---SHLYQAPFGMVPASLPCGSMNSGSPQLCCSSPDSVWVRCSTASLPLRKAHDSHSGFYASVQNPASNLNLSACQYAVDRPV : 325
      S6A LR Ka 2h
```

Домены принято изображать так

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



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

Гомеодомен является ДОМЕНОМ

Доказательство

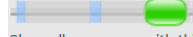
- Выравнивание, свидетельствующее о гомологичности последовательностей
- Представленность домена в неомологичных белках (обычно, но не обязательно)

```

G4VRE6_SCHMA/203-259
HME2B_DAHRE/173-229
HM1N_BOVWQ/373-429
TIEHES_HELRO/41-97
HM05_CAEL/36-92
G3IBX4_CRIGR/25-81
DLX3B_DAHRE/126-182
F1QFR3_DAHRE/136-192
B8A5N9_DAHRE/135-191
Q91967_CHICK/77-133
HM19_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
BSH1_DROME/275-331
H2XU06_CIOIM/470-524
HM19_CAEL/95-151
SLOU_DROME/546-602
F6VWQ6_XENTR/112-168
TIN_DROME/302-358
NKX25_RAT/138-194
H0XK12_OTOGA/100-156
HM09_CAEL/71-127
H2VEX2_TAKRU/13-59
UX517_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
K48BZ1_SOLL/24-79
PHO2_YEAST/78-134
W0X9_ARATH/52-114
W0X9_ORYS/111-72
W0X2_ORYS/24-85
W0X4_ARATH/87-148
W0X1_ARATH/73-134
W0X2_ARATH/11-72
W0X5_ORYS/41-102
WUS_SOLL/25-85
W0X6_ARATH/58-119
YHP1_YEAST/174-230
Y0X1_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

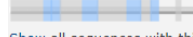
```

There are 25976 sequences with the following architecture **X2JL88_DROME** [Drosophila melanogaster (Fruit fly)] Uncharacterized



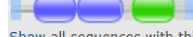
Show all sequences with this architecture.

There are 2311 sequences with the following architecture **X2JDY7_DROME** [Drosophila melanogaster (Fruit fly)] POU domain pr



Show all sequences with this architecture.

There are 2108 sequences with the following architecture **W6NCH4_HAECO** [Haemochus contortus (Barber pole worm)] Zinc f



Show all sequences with this architecture.

There are 1903 sequences with the following architecture **MOU1E3_MUSAM** [Musa acuminata subsp. malaccensis (Wild banana)]



Show all sequences with this architecture.

There are 1836 sequences with the following architecture **X1WJ92_ACYPI** [Acyrtosiphon pisum (Pea aphid)] Uncharacterized p



Эволюционные домены

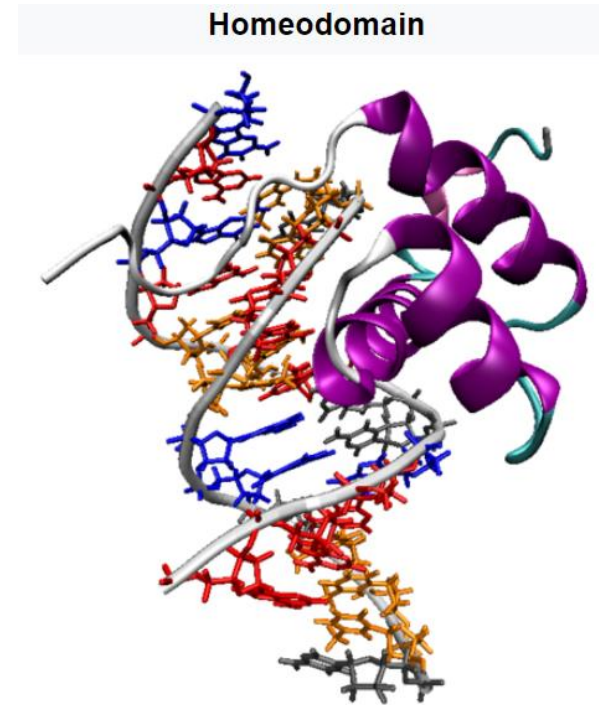
- Имеют определенную функцию (не всегда известна)

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 распространённость
- 3D структуры
- HMM профиль выравнивания
- Clan

6. блоки

Консервативный блок выравнивания (плюс-блок)

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

Вертикальный плюс-блок

```

          *          120          *          140          *
A0A1G9TZ02 : -----EVP---D--HDILVGGWPCPSFSI--MGD-----KEG--MDDERG----
A0A1H3E3S2 : --eshPIQR-dE--IDVVIGGPPCKGFESI--AG-----HRD--PDDERN----
A0A1I6G129 : ----tEWS---D--ADVVGPPCQGFSNlnSTK-----TDE--LDDDRN----
A0A1I6HFR3 : -ligeYLDD--DadATLLIACAPCQPFSP--LNH-----GKE--SSDHAM----
A0A1I6HG27 : ---seLYPD-gA--TKVLAGCAPCQPFSNlnNGT-----DTS--VRDDYG----
A0A1M5MP51 : -----AVVGgdD--VDLLVAGPECTHFS--ARG-----GKP--VSEQRR----
A0A1M5USI2 : -----AVVGddD--VDLLVAGPECTHFSR--ARG-----TKP--VSDQRR----
A0A256ILS2 : retghGVE---D--VDVVIGGPPCQGFSR--LNNeielDEM--EKDRRN----
A0A256KSV9 : -----SVP---S--HDLLIACWPCPSFSR--MGK-----LDG--LEDERG----
D4H0C8/5-4 : irnefGLEP-gE--VDVIAGCPQCQNEFSK--LRD-----TTPwpEDEPKD----
G2MPP5/169 : --iqdAVS---E--LDLLVGGPPCQSLSK--AGY-----RSR--RGDDEDysi.
J3ETN5/3-3 : vadlfDSSA--E--ATVLAGCAPCQPFSP--LTH-----GED--SSEHES----
J3JDK0/3-3 : --iaqMYPW--DadLKVLAACAPCQPYST--MGH-----SKG--NTHEDHn---
M0FSM4/5-3 : -----DLGK-aD--VDLVIGGPPCQPFSA--AARra-ggIEG--TESDEG----
V6DPC1/13- : -----KLDP--D--LDLLAGGPPCKGFSS--AQG-----ETN--TDDPRN----
ConSeq/1-2 : ----tls-----LDllluuPPCpsFSp--hst-----pcs---cDc+s----
          66          pC          S

```

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

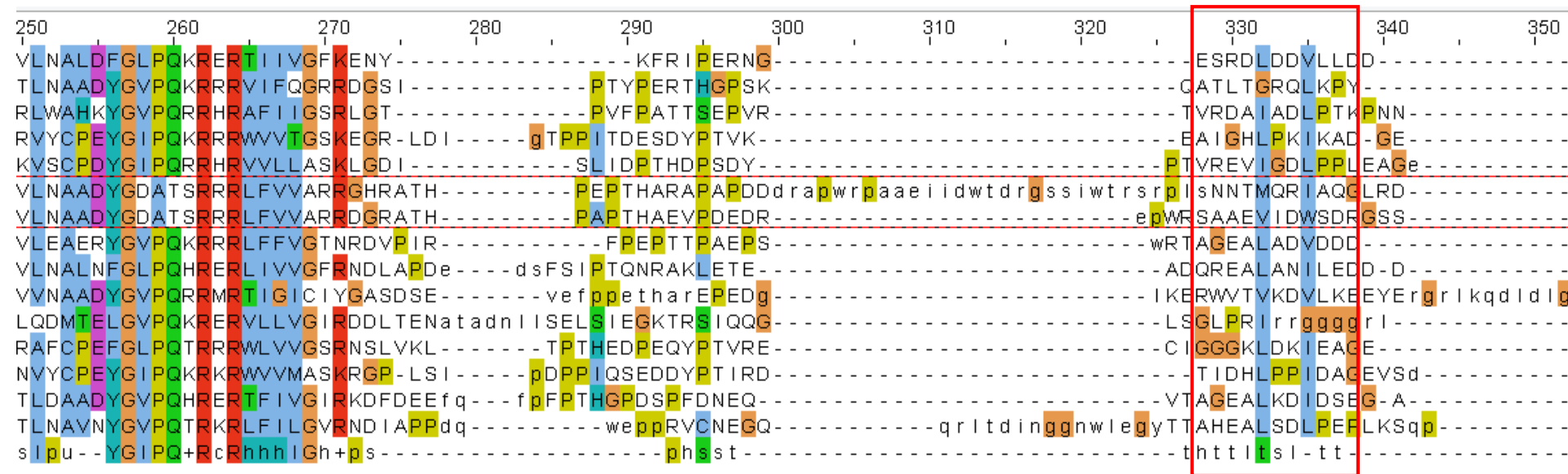
Частичный плюс-блок

	440	*	460	*	480	*	500	*	520	*	540					
A0A1G9TZ02	:	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	: 224				
A0A1H3E3S2	:	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	: 252				
A0A1I6G129	:	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	: 263				
A0A1I6HFR3	:	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	: 269				
A0A1I6HG27	:	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	: 258				
A0A1M5MP51	:	ivpvddledal	adrdepflvs	sttqtaav	dggt	rmvmgqqs	naralda	D	-----	-----	Repvptiatrgavh	fieaqp	fvkprnl	prgglh	tnatyvan	: 358
A0A1M5USI2	:	alderaepflv	-----	sttvstaadt	gtrmim	gqqs	naralda	D	-----	-----	Sevpvtvatrgavh	fieaqp	fikprn	-----	-----	: 342
A0A256ILS2	:	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	: 272
A0A256KSV9	:	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	: 254
D4H0C8/5-4	:	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	: 310
G2MPP5/169	:	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	: 318
J3ETN5/3-3	:	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	: 274
J3JDK0/3-3	:	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	: 274
M0FSM4/5-3	:	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	: 271
V6DPC1/13-	:	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	: 286
ConSeq/1-2	:	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	: 187

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