Выравнивания и домены

и разное ААл

План занятия

- Корона вирус.
- Множественное выравнивание vs парное
- Принцип Вальда
  - Пример из жизни
  - RdRP
  - Nucleoline – GAR домен.

Survivorship bias (https://en.wikipedia.org/wiki/Survivorship_bias#In_the_military)

- Блоки
- Принцип гомологии консервативных последовательностей
  - Гомология
    - блоки
  - Рекомбинация, домены
  - выращивание
- Гомология белков, домены.
- Pfam и др.
  - Обзор.
- Jalview
- 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

Коронавирус SARS-CoV-2

Research Article Journal of Optical Communications and Networking 1

Сравнение белка 6VSB коронавируса SARS-CoV-2 с протеомами родственных белков из рода Betacoronavirus

Марков Иван[1,*], Владимиров Даниил[1], Кряквин Максим[1]

[1] Факультет биоинженерии и биоинформатики, Московский государственный университет, 119234, ГСП-1, Ленинские горы МГУ 1, стр. 73 Москва
[*] Автор статьи: ivan@markov.im
[**] Соавтор отчета, научный руководитель: aba@belozersky.msu.ru

Compiled 11 апреля 2020 г.

В текущее время наблюдается пандемия коронавируса SARS-CoV-2. Масштаб этого события необычайно велик, и поэтому интересны причины столь быстрого распространия. Чтобы выявить эти причины, было произведено сравнение ключевого для заражения белка 6VSB. Результаты сравнения описаны в этой статье.

5. ЛИЧНЫЙ ВКЛАД

Вклад авторов: статья создана совместным трудом ВД, КМ и МИ. Поиск похожих белков был произведен КМ, кроме того, им была проведена редакция текста. Филогенетическое дерево и идея использовать 3D-структуры принадлежат ВД. Выравнивания и оформление статьи выполнил МИ.

1. ВСТУПЛЕНИЕ

Для сравнения был выбран гликопротеин шипа 6VSB, связывающийся с ангиотензинпревращающим ферментом 2(ACE2) и обеспечивающий проникновение вируса в клетку.

Белок еще не получил идентификатора UniProt, но на момент написания статьи имеет описанную 3D структуру.

На изображнеии 1 представлена шариковая модель белка.

2. МЕТОДЫ И МАТЕРИАЛЫ

Сведения о гликопротеине предоставлены согласно статье[1] авторов криомикроскопии.

Для сравнения последовательности этого белка был использован сервис BLAST.

Иллюстрации создавались с помощью 3D визулизаторов молекул JMol и Chimera.

Фиогенетическое дерево было построенно при помощи программы MEGAX.

Для оформления данной работы использовалась система компьютерной вёрстки LaTeX.

3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

A. Параметры и первичные результаты поиска BLAST

Для поискового запроса были выбраны параметры ожидаемого порога(expect threshold) равного 0.1, а размера слов(word size) - 3. Поиск BLAST выделил 71 схожую последовательность в роде Betacoroanvirus, наиболее интересными из которых являются четыре с покрытием последовательности около 93%. Графическая визуализация степени схожести представлена на изображении 2.

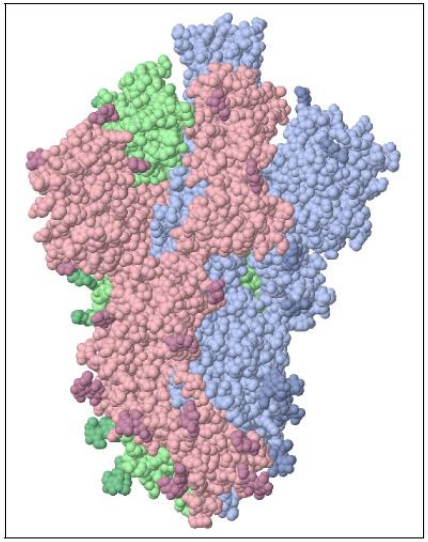
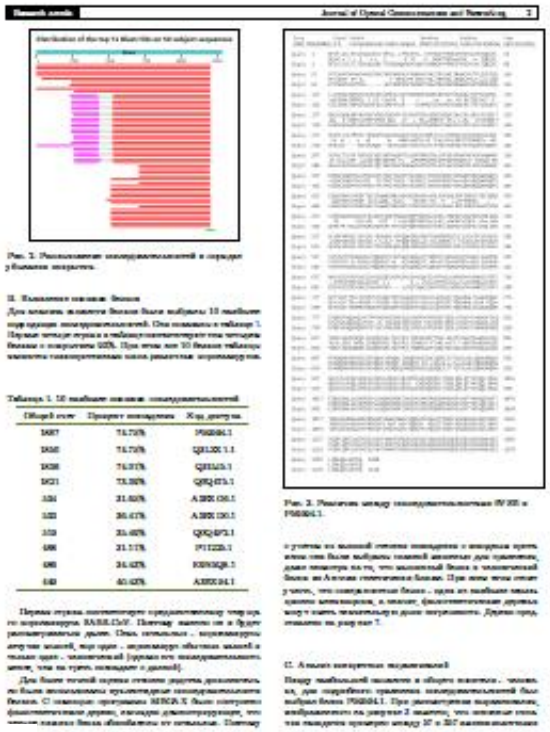
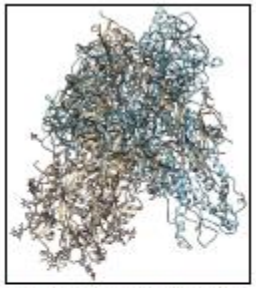
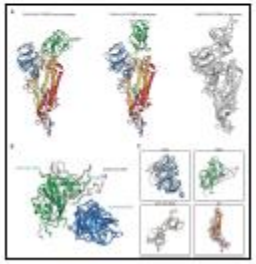
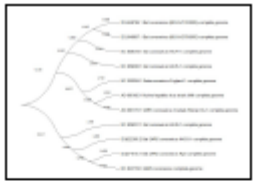


Рис. 1. Шариковая модель шипа 6VSB с раскраской по цепям.









Сравнение белка 6VSB коронавируса SARS-CoV-2 с протеомами родственных белков из рода Betacoronavirus

Марков Иван[1,*], Владимиров Данил[1], Кривин Максим[1], Алексинский Андрей[2, **]

[1] Факультет биоинженерии и биоинформатики, Московский государственный университет, 119234, РЕН 1, Ленинские горы МГУ 1, стр 73 Москва.
[*] Автор статьи: markov@ibi.ru
[**] Соавтор статьи, научный руководитель: alex@ibi.msu.ru

Опубликовано 11 марта 2020 г.

В текущее время наблюдается пандемия коронавируса SARS-CoV-2. Масштаб этого события необходимо знать, в силу интересы причины стал, быстрого распространения. Чтобы выявить эти причины, было проведено сравнение ключевого для заражения белка 6VSB. Результаты сравнения описаны в этой статье.

1. ВСТУПЛЕНИЕ

Для сравнения был выбран гомодимерная форма 6VSB, представляющий в антигенно-специфицирующие фрагмент S2(ACE2) и обеспечивающий прикрепление вируса к клетке.

Белок еще не получил идентификатора UniProt, но на момент написания статьи имеет номер по ID структуру.

На изображении 1 представлена пространственная модель белка.

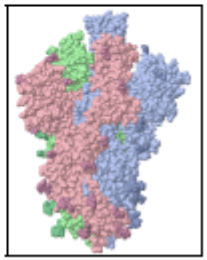


Рис. 1. Пространственная модель белка 6VSB в ленточной интерпретации.

2. МЕТОДЫ И МАТЕРИАЛЫ

Сведения о геномных последовательностях описаны статьи[1] которая кратко изложена.

Для сравнения последовательности этого белка был использован сервис BLAST.

Колонтитулы подвергались с помощью 3D визуализатора molsoft ZMM и UCSF Chimera.

Филогенетические деревья были построены при помощи программы MEGAX.

Для обработки данной работы использовались системы компьютерной обучения WEKA.

3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

A. Параметры и значения результаты поиска BLAST

Для сравнения вирусов были выбраны параметры сравнения: порог(expect) threshold(t) равного 0.1, и размер word(w) равный 3. Поиск BLAST выдал 75 вирусов по последовательность в ряде Betacoronavirus, наиболее которых были из которых являются четыре с покрытием геномных геномов свыше 95%. Графическая визуализация отмечена на изображении 2.

Рис. 2. Распределение геномных последовательностей в порядке убывания покрытия.

B. Выполнение множественного выравнивания

Для анализа являются белков были выбраны 10 наиболее подходящих последовательностей. Они показаны в таблице 1. Первые четыре строки в таблице соответствуют тем четырем белкам с покрытием 95%. При этом все 10 белков таблицы являются коронавирусными, кроме рассмотрения коронавирусов.

Таблица 1. 10 наиболее похожих последовательностей

Общий счет	Процент покрытия	Код доступа
1887	74.74%	P0DTC2.1
1856	74.74%	QHU36.1.1
1838	74.91%	QHU36.1
1821	73.38%	QHQ07.1
1324	31.60%	AJF08130.1
1322	30.41%	AJF08130.1
1312	30.46%	QHQ07.1
698	21.11%	P17223.1
698	24.42%	KHN03.1
640	40.42%	AJF08141.1

Первая строка соответствует предсказанному вирусу по коронавирусу SARS-CoV. Поэтому именно он и будет рассматриваться дальше. Очень интересно, что коронавирусы дают три общих значений, при одно - коронавирус обычных значений и только одно - человеческий (другие все коронавирусы имеют, что на уровне, исходящем с данной).

Для более точной оценки сходства результаты дополнительно были использованы предсказанные последовательности белков. С помощью программы MEGA X были построены филогенетические деревья, показывающие, что четыре похожих белка обособлены от остальных. Поэтому

Рис. 3. Различия между последовательностями 6VSB и P0DTC2.1.

и учитывая их высокий процент совпадения с вирусом пять или шесть были выбраны полной аналогии для сравнения, даже несмотря на то, что человеческий белок и человеческий белок на Алтаях отличаются близкие. При этом этот может у часть, что совпадающие белки - одна из наиболее сильно адаптированных, а значит, филогенетическая дерева могут иметь человеческую даже погрешность. Дерево представлено на рисунке 7.

C. Анализ конкретных выравниваний

Между наибольшей похожести в общем плотности - человека, для подробного сравнения последовательностей был выбран белок P0DTC2.1. При рассмотрении выравнивания, изображенного на рисунке 2 видно, что основные отличия выпадают примерно между 37 и 207 аминокислотами

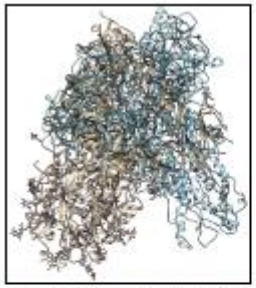


Рис. 4. 3D структуры 6VSB и P0DTC2.1, выделенные между собой, желтым обозначен белок типа SARS-CoV-2, голубым - SARS-CoV.

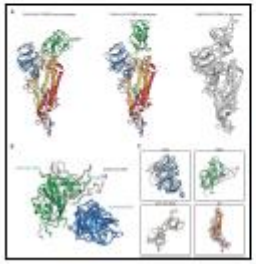


Рис. 5. Домены коронавирусов SARS-CoV2 и SARS-CoV.

Рис. 6. Различия между последовательностями QHZ83.1.1 и QHU36.1.

отличаются. Они отличаются на аминокислоте с каждой стороны. Кроме того, отличия во вторичной структуре белка сильно белка SARS-CoV-2 на месте альфа-спиралей SARS-CoV, можно рассмотреть при выравнивании 3D структур на рисунке 4.

Более подробное сравнение, представленное в статье показывает, что основные отличия в строении белка типа SARS-CoV и SARS-CoV-2 является расположение доменов связывания рецепторов[2][3]. Именно это факт обеспечивает более прикрепление клетки, чем в 10 раз[4] опять с клеткой человека и, значит, проходит в человеческий организм сильнее. Расположение доменов показано на изображении 5.

Выравнивание белков, имеющих номера выше четвертой не дает геномного выхода, потому как различия геномов TOR2 и рассмотрены рассмотрено по этой последовательности. Если сказать, что это один наиболее белок 6VSB.

Вывод, выявление вируса, причины и патогенез коронавирусов

Рис. 7. Дерево отличия родства между 10 наиболее похожими последовательностями.

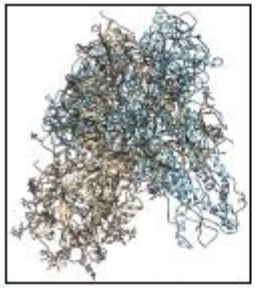


Рис. 4. 3D структуры 6VSB и P0DTC2.1, выделенные между собой, желтым обозначен белок типа SARS-CoV-2, голубым - SARS-CoV.

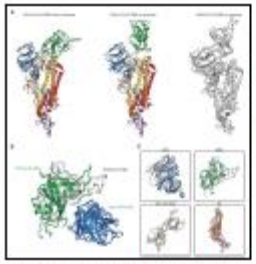


Рис. 5. Домены коронавирусов SARS-CoV2 и SARS-CoV.

Рис. 6. Различия между последовательностями QHZ83.1.1 и QHU36.1.

отличаются. Они отличаются на аминокислоте с каждой стороны. Кроме того, отличия во вторичной структуре белка сильно белка SARS-CoV-2 на месте альфа-спиралей SARS-CoV, можно рассмотреть при выравнивании 3D структур на рисунке 4.

Более подробное сравнение, представленное в статье показывает, что основные отличия в строении белка типа SARS-CoV и SARS-CoV-2 является расположение доменов связывания рецепторов[2][3]. Именно это факт обеспечивает более прикрепление клетки, чем в 10 раз[4] опять с клеткой человека и, значит, проходит в человеческий организм сильнее. Расположение доменов показано на изображении 5.

Выравнивание белков, имеющих номера выше четвертой не дает геномного выхода, потому как различия геномов TOR2 и рассмотрены рассмотрено по этой последовательности. Если сказать, что это один наиболее белок 6VSB.

Вывод, выявление вируса, причины и патогенез коронавирусов

4. ЗАКЛЮЧЕНИЕ

Основной причиной быстрого распространения SARS-CoV-2 является наличие геномных доменов связывания рецепторов относительно человека SARS-CoV. Сравнение геномов, филогенетического генерирования геномов с геномными вирусами показывает, что, опять, отличается вирус, в силу которого он этого не проявляется в вирусе.

5. СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

Все результаты представлены на основании статьи[1] и приложений репозитория Github геномов SARS-CoV-2.

БЛАГОДАРНОСТИ

Авторы благодарят Алексинского Андрея за рекомендации и методики проведения исследования, Лыжинского Артура за помощь в использовании визуализатора 3D структуры, Русанова Ивана и Сокурова Сергея за помощь на вопросы по работе BLAST.

СПИСОК ЛИТЕРАТУРЫ

1. D. Wrapp, N. Wang, K. S. Corbett, J. A. Goldsmith, C.-L. Hsieh, O. Abiona, B. S. Graham, and J. S. McLellan, "Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation," Science 367, 1260-1263 (2020), https://science.sciencemag.org/content/367/6483/1260.

1. Множественное выравнивание содержит больше информации, чем парное

```
                    *        20         *        40         *        60         *        80
THIE_LACLS : ~~~~~MTNKTLDLSVYFIAGAQNF---SECSLDGATQKIALIIKSGVTVYQFRDKGTIYK-EQKQ-RLSIAQKLQKVSEE :  70
THIE_MANSM : ~~~~~MNKIKSMLSVYFIAGSQDCRHLPGEPTENLLTILQRALEAGITCFQFREKGEQSLACDLQLKRRLALKCLQLCRQ :  75
THIE_STRA3 : ~~~~~~~~MKDTLKLYFVCGTVDC------SRKNILTVVEEALQAGITLFQFREKGFTAL-QGKE-KIAMAKQLQILCKQ :  64
THIE_LISIN : ~~~~~~~~MRAELAVYFIAGTQDI------VRGTLPSVLEEALKAGITCFQYREKGAGALQTASE-RKEMALECQQLCAK :  65
THIE_ANOFW : ~~~~~~~MMKQKLSLYFVMGSIDC-------TKDPLAVLDEAIKGGITMFQFREKGKGAL-TGIE-KYRLAEKLLERCRM :  64
THIE_GEOTN : MARITSEEMKERLAVYFIMGSQNS-------ERPAEDVLKEALDGGVTLFQFREKGSAAL-EGEE-KEALARQLQRLCRT :  71
THIE_BACSU : MTRISREMMKELLSVYFIMGSNNT-------KADPVTVVQKALKGGATLYQFREKGGDAL-TGEA-RIKFAEKAQAACRE :  71
THIE_BACA2 : MTRISREMMKDMLSVYFIMGSNNT-------SADPVSVVEKAIEGGATLFQFREKGSGSL-TGEE-RLLFAKRVQDVCRQ :  71
THIE_OCEIH : ~~~~~MKFDKHMLRKYFIMGSQNC-------HRDPREILKEAASAGITAFQYREKGKNSL-TGTA-KVELAKDLKAICHD :  66
THIE_STAAB : ~~~~~~MFNQSYLNVYFICGTSNV-----PSHRTIHEVLEAALKAGITLFQFREKGESAL-KGND-KLVLAKELQHLCHQ :  67
THIE_STACT : ~~~~~~MFQSKDLNVYFICGTQDI-----PEGRTIQEVLKEALEGGITLYQFREKGNGAK-TGQD-KVALAKELQALCKS :  67
                         L  YF6 G  1              6  a   G T 5Q5R7KG   l       4   A      c

                    *       100         *       120         *       140         *       160
THIE_LACLS : AGVSFIVNDDVELARELNADGIHIGQTDESVSKVREKVGQEMWLGLSVTKADELKTAQ-SSGADYLGIGPIYPTNSKNDA : 149
THIE_MANSM : FQVPFIVNDDVELALSIQADGIHVGQKDTAVETILRNTRNKPIIGLSINTLAQALANKDRQDIDYFGVGPIFPTNSKADH : 155
THIE_STRA3 : YQVPFIIDDDIDLVELIDADGLHIGQNDLPVDEARRRLPDKI-IGLSVSTMAEYQKSQ-LSVVDYIGIGPFNPTQSKADA : 142
THIE_LISIN : YQVPFIINDDVALALEIGADGIHVGQNDEEIRQVIASCAGKMKIGLSVHSVSEAEEAERLGSVDYIGVGPIFPTISKADA : 145
THIE_ANOFW : YNIPFIVNDDVDLALALQADGVHVGQEDEVAERVRDRIGDKY-LGVSVHNLNEVKKAL-AACADYVGLGPIFPTVSKEDA : 142
THIE_GEOTN : YGVPFIVNDDVELAIAIDADGVHVGQDDEDARRVREKIGDKI-LGVSAHNVEEARAAI-EAGADYIGVGPIYPTRSKDDA : 149
THIE_BACSU : AGVPFIVNDDVELALNLKADGIHIGQEDANAKEVRAAIGDMI-LGVSAHTMSEVKQAE-EDGADYVGLGPIYPTETKKDT : 149
THIE_BACA2 : AGIPFIINDDVELALRLEADGVHIGQDDADAEETRAAIGDMI-LGVSAHNVSEVKRAE-AAGADYVGMGPVYPTETKKDA : 149
THIE_OCEIH : FQIPFIINDDVDLAKQLDADGIHIGQDDQPVEVVRKQFPNKI-IGLSISTNNELNQSP-LDLVDYIGVGPIFDTNTKEDA : 144
THIE_STAAB : YNVPFIVNDDVSLAKEINADGIHVGQDDAKVKEIAQYFTDKI-IGLSISDLGEYAKSD-LTHVDYIGVGPIYPTPSKHDA : 145
THIE_STACT : YNVPFIVNDDVALAEEIDADGIHVGQDDEAVDDFNNRFEGKI-IGLSIGNLEELNASD-LTYVDYIGVGPIFATPSKDDA : 145
               6pFI61DD6 La   6 ADG6H6GQ D                 6G6S      2          DY G6GP   pT 3K Da

                    *       180         *       200         *       220         *       240
THIE_LACLS : AKPIGIKDLR-LMLLENQLPIVGIGGITQDSLTELSAIGLDGLAVISLLTEAENPKKVAQMIRQKITKNG~~~~~~~~~~ : 218
THIE_MANSM : SPLVGMNFIRQIRQLGIDKPCVAIGGIKEESAAILRRLGADGVAVISAISHSVNIANTVKTLAQK~~~~~~~~~~~~~~~ : 220
THIE_STRA3 : KPAVGNRTTKAVREINQDIPIVAIGGITSDFVHDIIESGADGIAVISAISKANHIVDATRQLRYEVEKALVNRQKRSDVI : 222
THIE_LISIN : EPVSGTAILEEIRRAGIKLPIVGIGGINETNSAEVLTAGADGVSVISAITRSEDCQSVIKQLKNPGSPS~~~~~~~~~~~ : 214
THIE_ANOFW : KQACGLTMIEHIRAHEKRVPLVAIGGITEQTAKQVIEAGADGIAVISAICRAEHIYEQTKRLYEMVMRAKQKGDR~~~~~ : 217
THIE_GEOTN : NEAQGPGILRHLREQGITIPIVAIGGITADNTRAVIEAGADGVSVISAIASAPEPKAAAAALATAVREANL---R~~~~~ : 221
THIE_BACSU : RAVQGVSLIEAVRRQGISIPIVGIGGITIDNAAPVIQAGADGVSMISAISQAEDPESAARKFREEIQTYKTG--R~~~~~ : 222
THIE_BACA2 : EAVQGVTLIEEVRRQGITIPIVGIGGITADNAAPVIEAGADGVSMISAISQAEDPKAAARKFSEEIRRSKAGLSR~~~~~ : 224
THIE_OCEIH : KTAVGLEWIQSLKKQHPSLPLVAIGGINTTNAQEIIQAGADGVSFISAITETDHILQAVQRL~~~~~~~~~~~~~~~~~~ : 206
THIE_STAAB : HTPVGPEMIATFKEMNPQLPIVAIGGINTSNVAPIVEAGANGISVISAISKSENIEKTVNRFKDFFNN~~~~~~~~~~~~ : 213
THIE_STACT : SEPVGPKMIETLRKEVGDLPIVAIGGISLDNVQEVAKTSADGVSVISAIARSPHVTETVHKFLQYFK~~~~~~~~~~~~~ : 212

THIE_LACLS   1 MTNKTLDLSVYFIAGAQNFSECSLDGATQKIALI---IKSGVTVYQFRDKG--TIYKEQKQRLSIAQKLQKVSEEAG  72
THIE_MANSM   1 MNKIKSMLSVYFIAGSQDCRHLPGEPTENLLTILQRALEAGITCFQFREKGEQSLACDLQLKRRLALKCLQLCRQFQ  77

THIE_LACLS  73 VSFIVNDDVELARELNADGIHIGQTDESVSKVREKVGQEMWLGLSV-TKADELKTAQSSGADYLGIGPIYPTNSKND 148
THIE_MANSM  78 VPFIVNDDVELALSIQADGIHVGQKDTAVETILRNTRNKPIIGLSINTLAQALANKDRQDIDYFGVGPIFPTNSKAD 154

THIE_LACLS 149 AAKPIG---IKDLRLMLLENQLPIVGIGGITQDSLTELSAIGLDGLAVISLLTEAENPKKVAQMIRQKITKNG     218
THIE_MANSM 155 HSPLVGMNFIRQIRQLGIDK--PCVAIGGIKEESAAILRRLGADGVAVISAISHSVNIANTVKTLAQK----     220
```

```
                        *        20         *        40         *        60         *        80
THIE_LACLS : ~~~~~MTNKTLDLSVYFIAGAQNF---SECSLDGATQKIALIIKSGVTVYQFRDKGTIYK-EQKQ-RLSIAQKLQKVSEE : 70
THIE_MANSM : ~~~~~MNKIKSMLSVYFIAGSQDCRHLPGEPTENLLTILQRALEAGITCFQFREKGEQSLACDLQLKRRLALKCLQLCRQ : 75
THIE_STRA3 : ~~~~~~~~~MKDTLKLYFVCGTVDC------SRKNILTVVEEALQAGITLFQFREKGFTAL-QGKE-KIAMAKQLQILCKQ : 64
THIE_LISIN : ~~~~~~~~~MRAELAVYFIAGTQDI------VRGTLPSVLEEALKAGITCFQYREKGAGALQTASE-RKEMALECQQLCAK : 65
THIE_ANOFW : ~~~~~~~MMKQKLSLYFVMGSIDC-------TKDPLAVLDEAIKGGITMFQFREKGKGAL-TGIE-KYRLAEKLLERCRM : 64
THIE_GEOTN : MARITSEEMKERLAVYFIMGSQNS-------ERPAEDVLKEALDGGVTLFQFREKGSAAL-EGEE-KEALARQLQRLCRT : 71
THIE_BACSU : MTRISREMMKELLSVYFIMGSNNT-------KADPVTVVQKALKGGATLYQFREKGGDAL-TGEA-RIKFAEKAQAACRE : 71
THIE_BACA2 : MTRISREMMKDMLSVYFIMGSNNT-------SADPVSVVEKAIEGGATLFQFREKGSGSL-TGEE-RLLFAKRVQDVCRQ : 71
THIE_OCEIH : ~~~~~MKFDKHMLRKYFIMGSQNC-------HRDPREILKEAASAGITAFQYREKGKNSL-TGTA-KVELAKDLKAICHD : 66
THIE_STAAB : ~~~~~~MFNQSYLNVYFICGTSNV-----PSHRTIHEVLEAALKAGITLFQFREKGESAL-KGND-KLVLAKELQHLCHQ : 67
THIE_STACT : ~~~~~~MFQSKDLNVYFICGTQDI-----PEGRTIQEVLKEALEGGITLYQFREKGNGAK-TGQD-KVALAKELQALCKS : 67
                       L  YF6 G  1             6  a   G T 5Q5R7KG   l      4   A      c

                        *       100         *       120         *       140         *       160
THIE_LACLS : AGVSFIVNDDVELARELNADGIHIGQTDESVSKVREKVGQEMWLGLSVTKADELKTAQ-SSGADYLGIGPIYPTNSKNDA : 149
THIE_MANSM : FQVPFIVNDDVELALSIQADGIHVGQKDTAVETILRNTRNKPIIGLSINTLAQALANKDRQDIDYFGVGPIFPTNSKADH : 155
THIE_STRA3 : YQVPFIIDDDIDLVELIDADGLHIGQNDLPVDEARRRLPDKI-IGLSVSTMAEYQKSQ-LSVVDYIGIGPFNPTQSKADA : 142
THIE_LISIN : YQVPFIINDDVALALEIGADGIHVGQNDEEIRQVIASCAGKMKIGLSVHSVSEAEEAERLGSVDYIGVGPIFPTISKADA : 145
THIE_ANOFW : YNIPFIVNDDVDLALALQADGVHVGQEDEVAERVRDRIGDKY-LGVSVHNLNEVKKAL-AACADYVGLGPIFPTVSKEDA : 142
THIE_GEOTN : YGVPFIVNDDVELAIAIDADGVHVGQDDEDARRVREKIGDKI-LGVSAHNVEEARAAI-EAGADYIGVGPIYPTRSKDDA : 149
THIE_BACSU : AGVPFIVNDDVELALNLKADGIHIGQEDANAKEVRAAIGDMI-LGVSAHTMSEVKQAE-EDGADYVGLGPIYPTETKKDT : 149
THIE_BACA2 : AGIPFIINDDVELALRLEADGVHIGQDDADAEETRAAIGDMI-LGVSAHNVSEVKRAE-AAGADYVGMGPVYPTETKKDA : 149
THIE_OCEIH : FQIPFIINDDVDLAKQLDADGIHIGQDDQPVEVVRKQFPNKI-IGLSISTNNELNQSP-LDLVDYIGVGPIFDTNTKEDA : 144
THIE_STAAB : YNVPFIVNDDVSLAKEINADGIHVGQDDAKVKEIAQYFTDKI-IGLSISDLGEYAKSD-LTHVDYIGVGPIYPTPSKHDA : 145
THIE_STACT : YNVPFIVNDDVALAEEIDADGIHVGQDDEAVDDFNNRFEGKI-IGLSIGNLEELNASD-LTYVDYIGVGPIFATPSKDDA : 145
               6pFI61DD6 La   6 ADG6H6GQ D                6G6S      2            DY G6GP   pT 3K Da

                        *       180         *       200         *       220         *       240
THIE_LACLS : AKPIGIKDLR-LMLLENQLPIVGIGGITQDSLTELSAIGLDGLAVISLLTEAENPKKVAQMIRQKITKNG~~~~~~~~~~~ : 218
THIE_MANSM : SPLVGMNFIRQIRQLGIDKPCVAIGGIKEESAAILRRLGADGVAVISAISHSVNIANTVKTLAQK~~~~~~~~~~~~~~~~ : 220
THIE_STRA3 : KPAVGNRTTKAVREINQDIPIVAIGGITSDFVHDIIESGADGIAVISAISKANHIVDATRQLKIEVEKALVNRQKRSDVI : 222
THIE_LISIN : EPVSGTAILEEIRRAGIKLPIVGIGGINETNSAEVLTAGADGVSVISAITRSEDCQSVIKQLKNPGSPS~~~~~~~~~~~ : 214
THIE_ANOFW : KQACGLTMIEHIRAHEKRVPLVAIGGITEQTAKQVIEAGADGIAVISAICRAEHIYEQTKRLYEMVMRAKQKGDR~~~~~ : 217
THIE_GEOTN : NEAQGPGILRHLREQGITIPIVAIGGITADNTRAVIEAGADGVSVISAIASAPEPKAAAAALATAVREANL---R~~~~~ : 221
THIE_BACSU : RAVQGVSLIEAVRRQGISIPIVGIGGITIDNAAPVIQAGADGVSMISAISQAEDPESAARKFREEIQTYKTG--R~~~~~ : 222
THIE_BACA2 : EAVQGVTLIEEVRRQGITIPIVGIGGITADNAAPVIEAGADGVSMISAISQAEDPKAAARKFSEEIRRSKAGLSR~~~~~ : 224
THIE_OCEIH : KTAVGLEWIQSLKKQHPSLPLVAIGGINTTNAQEIIQAGADGVSFISAITETDHILQAVQRL~~~~~~~~~~~~~~~~~ : 206
THIE_STAAB : HTPVGPEMIATFKEMNPQLPIVAIGGINTSNVAPIVEAGANGISVISAISKSENIEKTVNRFKDFFNN~~~~~~~~~~~~ : 213
THIE_STACT : SEPVGPKMIETLRKEVGDLPIVAIGGISLDNVQEVAKTSADGVSVISAIARSPHVTETVHKFLQYFK~~~~~~~~~~~~~ : 212
                G                P V IGGI        6    ga1G6   ISa6

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

              80         *       100         *       120         *       140         *
THIE_LACLS : VSFIVNDDVELARELNADGIHIGQTDESVSKVREKVGQEMWLGLSV-TKADELKTAQSSGADYLGIGPIYPTNSKND :
THIE_MANSM : VPFIVNDDVELALSIQADGIHVGQKDTAVETILRNTRNKPIIGLSINTLAQALANKDRQDIDYFGVGPIFPTNSKAD :
             V FIVNDDVELA   6 ADGIH6GQ D   V   6          6GLS6 T   A   L          DY G6GPIFPTNSK D

                160         *       180         *       200         *       220
THIE_LACLS : AAKPIG---IKDLRLMLLENQLPIVGIGGITQDSLTELSAIGLDGLAVISLLTEAENPKKVAQMIRQKITKNG : 218
THIE_MANSM : HSPLVGMNFIRQIRQLGIDK--PCVAIGGIKEESAAILRRLGADGVAVISAISHSVNIANTVKTLAQK----- : 220
               6G    I4  6R  6  6     P V IGGI 2 S    L   6G DG6AVIS 63    N        6 QK
```

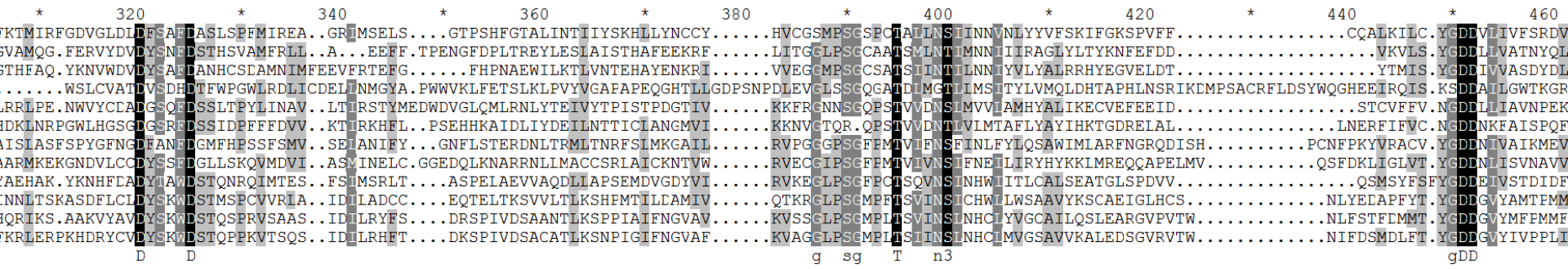
2. Консервативное значит важное

Формулировка МГ: принцип Вальда(Wald)

Примеры

- Глиняная посуда > 20 тыс. лет
  - Пробовали деревянную, pewter – *сплав олова, включающий свинец*, плетённую и др.
- Правые и левые туфли (Древняя Греция ?)
  - До того и в средние века обе туфли были одинаковые

РНК зависимая РНК полимераза (RdRP), консервативные участки



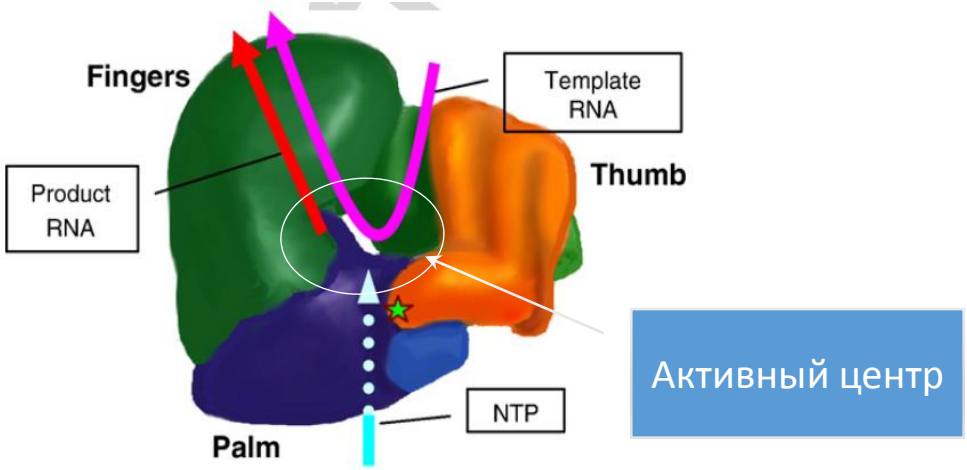


Fig. 1. Schematic architecture of “small” RdRP. The hairpin between the palm and thumb domains is in light blue. The predicted approximate location of MV RdRP Trp460 is marked by star.

Рисунок мой :) хвастаюсь

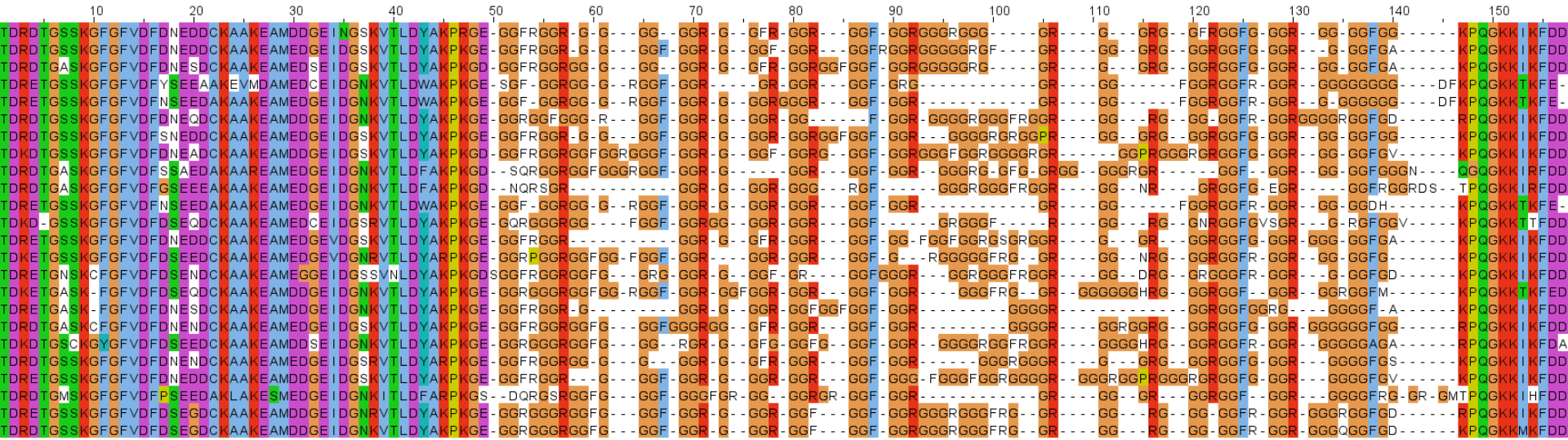
Каталитические мотивы RdRP

	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

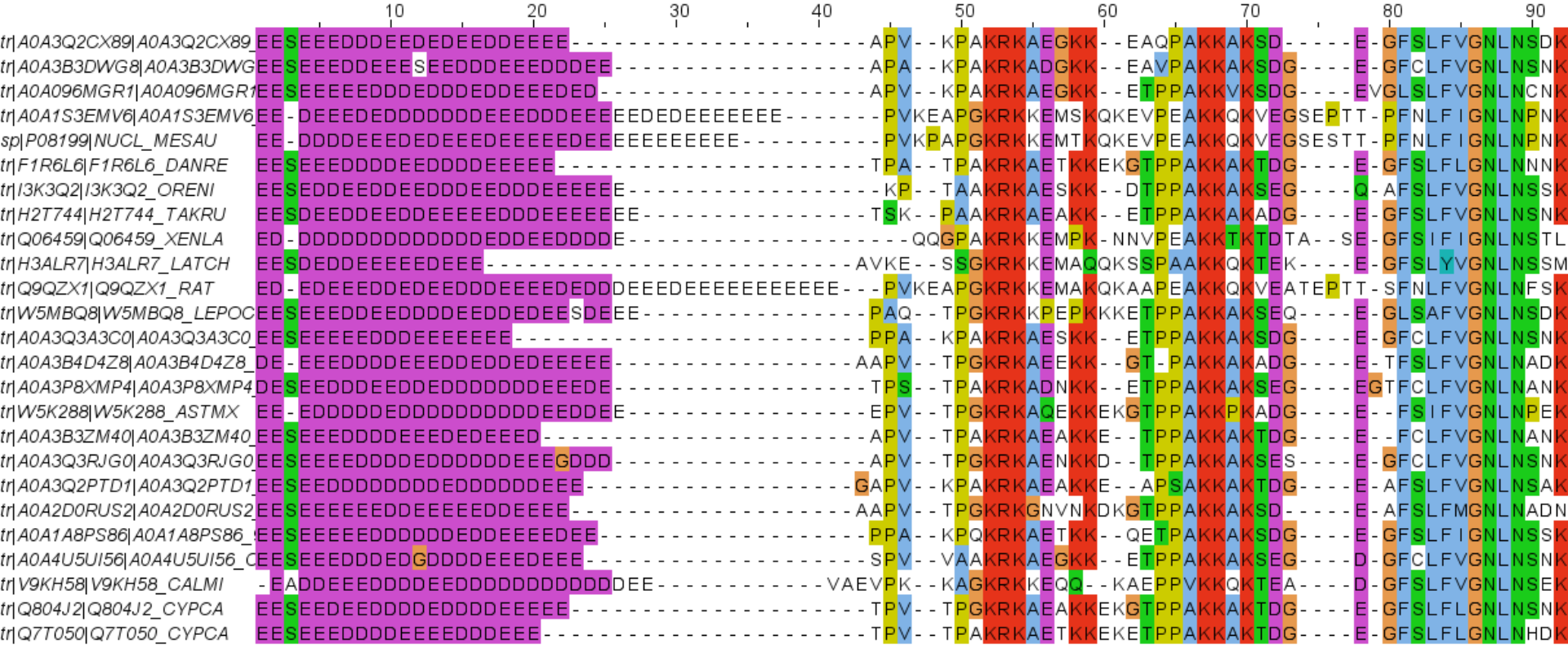
Nucleolin. Глицин-аргинин богатый участок (GAR)



КОНСЕРВАТИВЕН - присутствует в этом месте у всех нуклеолинов. Значит важен для функции. Зачем он – толком не известно. GAR встречаются у многих ДНК и РНК связывающих белков.

Я отредактировал выравнивание вручную чтобы GAR поместился на слайде (в выравнивании, построенном программой, был в два раза длиннее по числу позиций

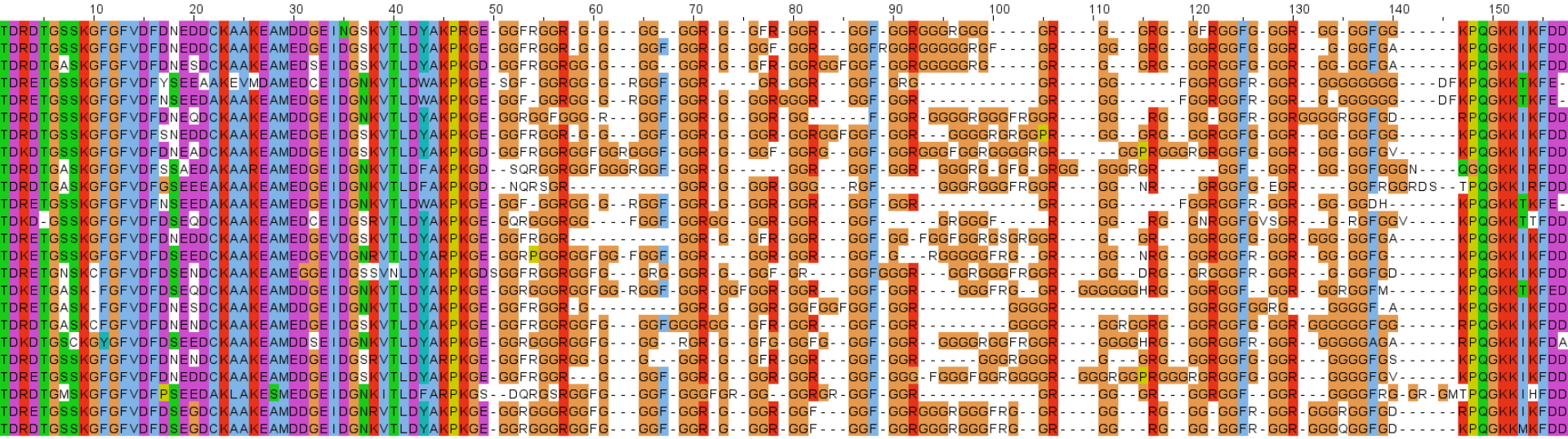
Nucleolin. Кислый участок и сигнал сигнал ядерной локализации



3. Сходные последовательности гомологичны

В выравнивании гомологичных последовательностей гомологичные остатки стоят в одной колонке

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

4. Три типа эволюции белков

1. **Локальные мутации в белке** (из-за мутаций ДНК в кодирующей последовательности) => ДОМЕНЫ

- Этот вид эволюции описывается выравниванием последовательностей
- Выравнивание последовательностей гомологичных белков ныне живущих организмов позволяет реконструировать последовательности у предков! (с некоторой точностью)

2. «Выращивание» коротких последовательностей для определенной функции

- ASN_GLYCOSYLATION, PS00001; N-glycosylation site (PATTERN with a high probability of occurrence!)
  N-{P}-[ST]-{P} N is the glycosylation site

3. Крупные перестройки => DOMAIN SHUFFLING

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

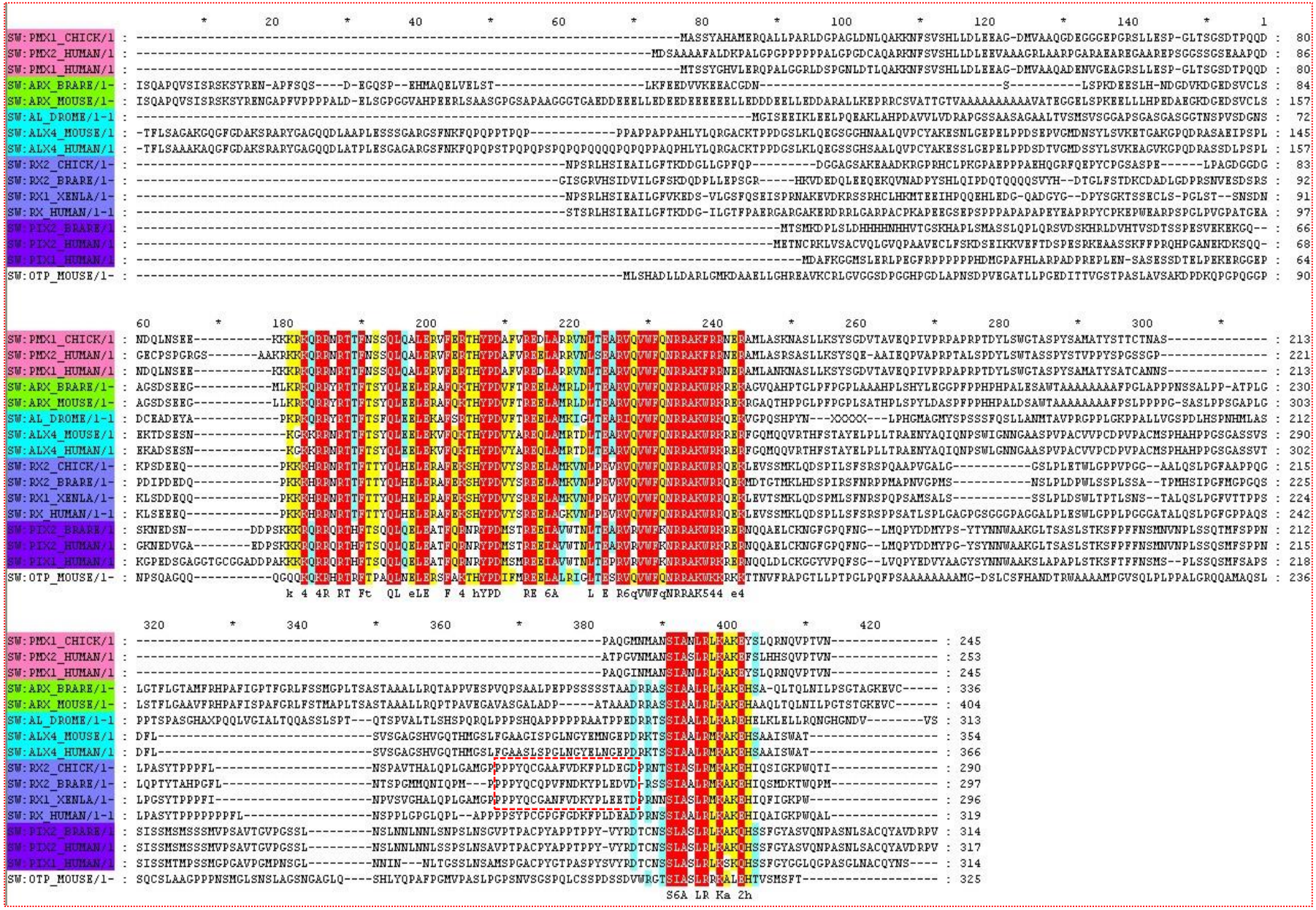
Терминологическая проблема.

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

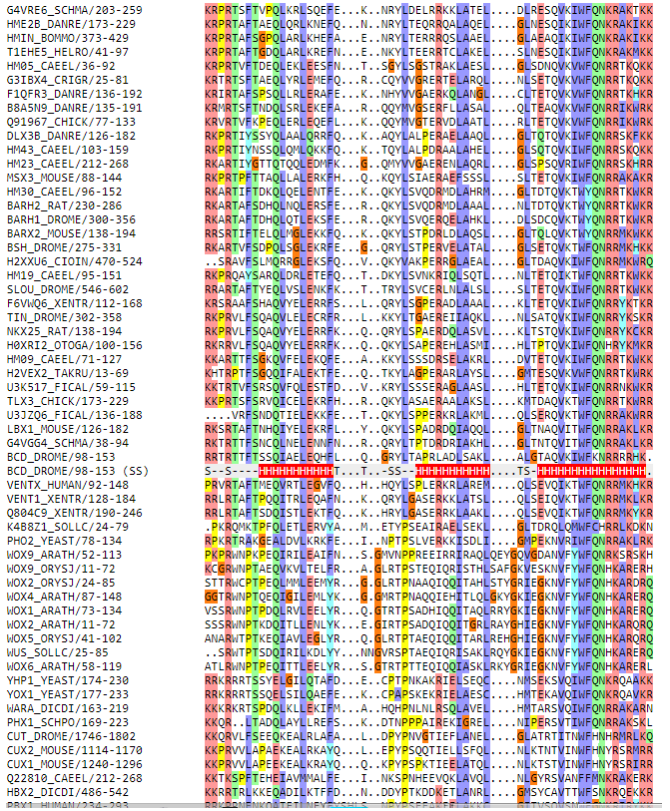
X1WJ92_ACYPI [Acyrthosiphon pisum (Pea aphid)]
Uncharacterized protein (408 residues)

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

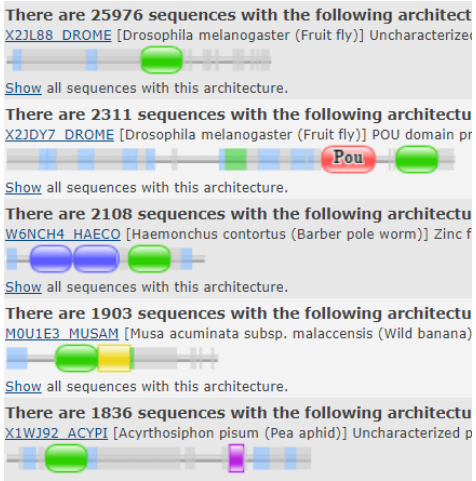
Гомеодомен является ДОМЕНОМ

Доказательство

- Выравнивание, свидетельствующее о гомологичности последовательностей



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



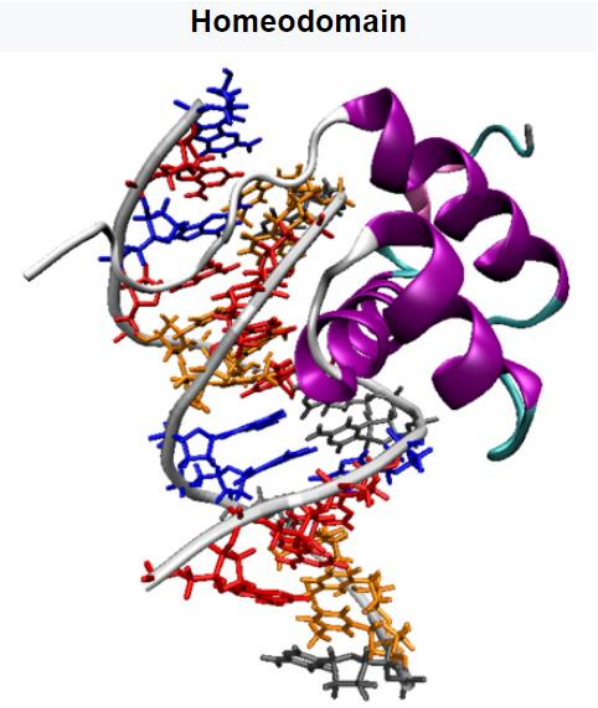
Эволюционные домены

- Имеют определенную функцию (не всегда известна)
  DUF – Domain of Unknown Function

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]

- Часто совпадают со структурными доменами (но не всегда)



5. БД Pfam (http://pfam.xfam.org/)

Protein Families (Pfam)
База семейств доменов
Белков

Одна запись содержит информацию о ДОМЕНЕ, информацию о всех белках, содержащих данный домен, выравнивание доменов из всех белков

Rubens: **Holy Family** with St. Elizabeth

Что посмотрим в Pfam

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

- Type I restriction modification DNA specificity domain
- Methylase_S
- **PF**01420
- Seed – по нему сделан профиль домена
- Full - все находки по профилю в Uniprot (с опозданием по версиям)
- This domain is also known as the target recognition domain (TRD). ….

Смотрим домен methyltransferaseD12

6. Ревизия выравнивания и блоки

Консервативный блок выравнивания (К-блок)

- Подтверждается консервативными или функционально консервативными позиция во всех фрагментах. Назовем «абсолютно консерваттивными (АбКпоз)
- 1я и последняя позиции такие - АбКпоз
- В блоке нет гэпов
- АбКпоз встречаются достаточно часто. Пока на это нет строгих критериев.
- Идущие подряд К-блоки, разделенные короткими неконсервативными линкерами, возможно, с небольшим числом позиций с гэпами, стоит объединить в суперконсервативный блок (СК-блок) и предполагать консервативность ВСЕХ фрагментов СК-блока

Консервативный вертикальный блок

```
                    *       120         *       140         *
A0A1G9TZ02 : -----EVP---D--HDILVGGWPCPSFSI--MGD-----KEG--MDDERG---
A0A1H3E3S2 : --eshPIQR-dE--IDVVIGGPPCKGFSI--AG------HRD--PDDERN---
A0A1I6G129 : ----tEWS---D--ADVVVGGPPCQGFSNlnSTK-----TDE--LDDDRN---
A0A1I6HFR3 : -ligeYLDD--DadATLIAACAPCQPFSP--LNH-----GKE--SSDHAM---
A0A1I6HG27 : ---seLYPD-gA--TKVIAGCAPCQPFSNlnNGT-----DTS--VRDDYG---
A0A1M5MP51 : -----AVVGgdD--VDLLVAGPECTHFSS--ARG-----GKP--VSEQRR---
A0A1M5USI2 : -----AVVGddD--VDLLVAGPECTHFSR--ARG-----TKP--VSDQRR---
A0A256ILS2 : retghGVE---D--VDVVIGGPPCQGFSR--LNNerielDEM--EKDRRN---
A0A256KSV9 : -----SVP---S--HDLLLACWPCPSFSR--MGK-----LDG--LEDERG---
D4H0C8/5-4 : irnefGLEP-gE--VDVIAGCPPCQNFSK--LRD-----TTPwpEDEPKD---
G2MPP5/169 : --iqdAVS---E--LDLLVGGPPCQSLSK--AGY-----RSR--RGDDEDysi
J3ETN5/3-3 : vadlfDSSA--E--ATVIAGCAPCQPFSP--LTH-----GED--SSEHES---
J3JDK0/3-3 : --iaqMYPW--DadLKVIAACAPCQPYST--MGH-----SKG--NTHEDHn--
M0FSM4/5-3 : -----DLGK-aD--VDLVIGGPPCQPFSA--AARra-ggIEG--TESDEG---
V6DPC1/13- : -----KLPD--D--LDLIAGGPPCKGFSS--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

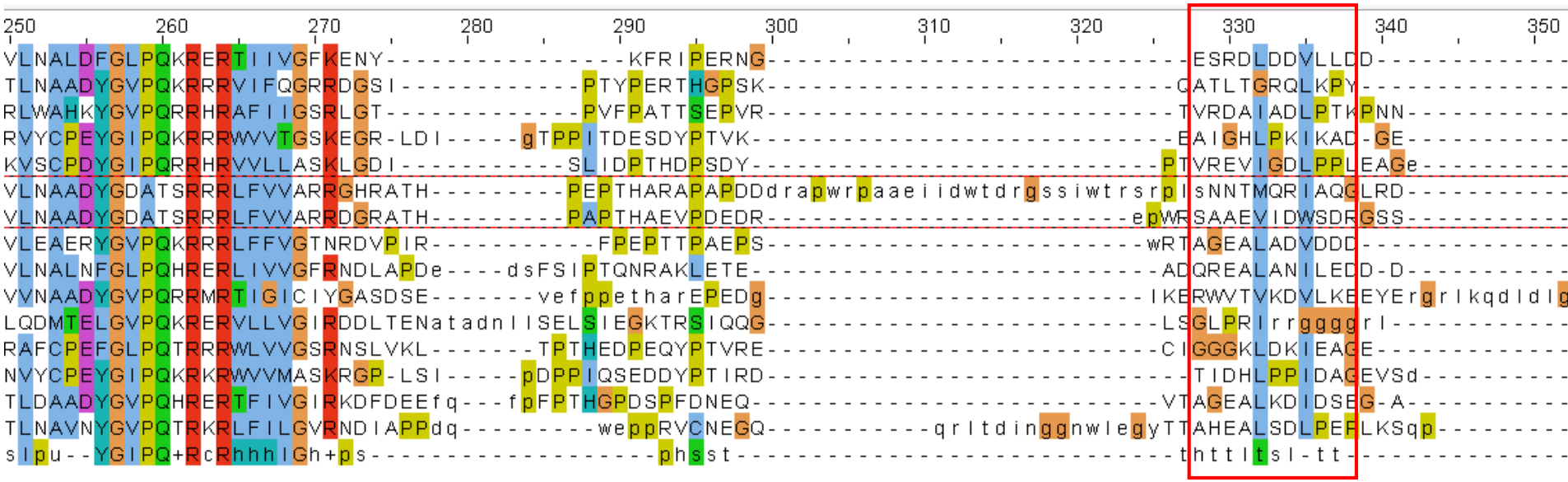
Консервативный НЕ вертикальный блок

```
                 440         *       460         *       480         *       500         *       520         *       540
A0A1G9TZ02 : ------------------------------------------------------SVWHENRSGN------------------------------------------ : 224
A0A1H3E3S2 : ----------------------------------------------IADVDPGESLYESYGDSWR------------------------------------------ : 252
A0A1I6G129 : ------------------------------------------------CWKG--YESGGTDLFGR------------------------------------------ : 263
A0A1I6HFR3 : ------------------------------------------------HKQDSGSTFDSVYGRMEPDE---------------------------------------- : 269
A0A1I6HG27 : ------------------------------------------------HRKESGRSFDSVYGRMEW------------------------------------------ : 258
A0A1M5MP51 : ivpvddledaladrdepflvsstgtaavdggtrmvmgqqsnaraldaD-------------------Repvptiatrgavhfieaqpfvkprnlprgglhtnatyvan : 358
A0A1M5USI2 : alderaepflv---------sttvstaadtgtrmimgggsnaraldaD-------------------Sepvptvatrgavhfieaqpfikprn--------------l : 342
A0A256ILS2 : ---------------------------------------------cdCADTDACPHEPEIVKRYGT------------------------------------- : 272
A0A256KSV9 : ----------------------------------------------AGQVTTRP-YSGTLRASSS------------------------------------------ : 254
D4H0C8/5-4 : ----------------------------------------------HRG-----------------Fdtqa------------------------------------ : 310
G2MPP5/169 : ------------------------------------dtgwdvkynsdGEYSEYVEYDVGTAENKRFGDkyrmlew--------------------------------- : 318
J3ETN5/3-3 : ---------------------------------------------dchKKDTGATYQSVYGRMEPDK------------------------------------- : 274
J3JDK0/3-3 : ----------------------------------------------HRKASGRSYKAPYSRMRPD-E------------------------------------------ : 274
M0FSM4/5-3 : ----------------------------------------------HPEPIFGWRSRFSDYLYKA------------------------------------------ : 271
V6DPC1/13- : ----------------------------------------------TRDGEDVWIPTNHKHQDHSRShrekmaeyelgksg-------------------------- : 286
ConSeq/1-2 : ----------------------------------------------tp-p-h-------t------------------------------------------------ : 187
```

Как бы назвать такие блоки? Пока «не вертикальный К-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.
По возможности, надо избегать того, чтобы внутри него не оказалось К-подблока

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 :(

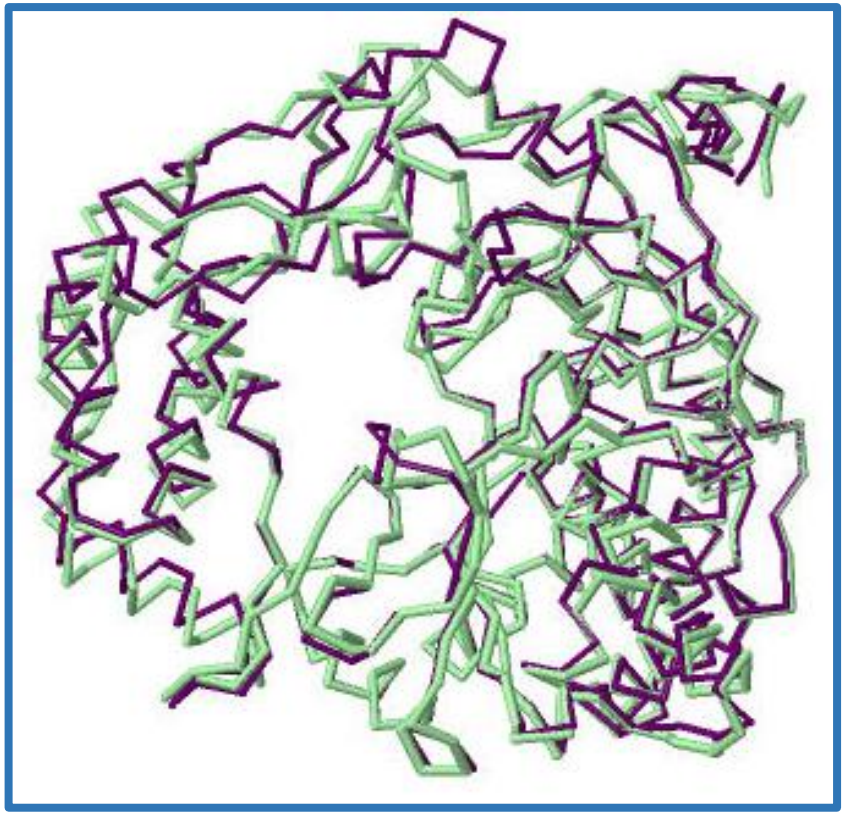
Смотрим выравнивание из MethyltransferaseD12

Другие БД доменов

- PROSITE
- Smart
- Interpro – объединение всех БД белковых доменов

7. Проверка выравнивания по совмещению полипептидных остовов 3Dструктур

Известно, что структура консервативнее последовательностей



Совмещение полипептидных цепей двух RdRP +РНК вирусов.

Не успел посмотреть каких. Выбрал случайно из совмещения пяти структур

КОНЕЦ ПРЕЗЕНТАЦИИ

Страница домена Methylase_S (PF01420)

Family: *Methylase_S* (PF01420)

38 architectures 7284 sequences 3 interactions 2296 species 14 structures

- Summary
- Domain organisation
- Clan
- Alignments
- HMM logo
- Trees
- Curation & model
- Species
- Interactions
- Structures

Jump to... enter ID/acc Go

Summary: Type I restriction modification DNA specificity domain

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

No Wikipedia article | Pfam | InterPro

This tab holds the annotation information that is stored in the Pfam database. As we move to using Wikipedia as our main source of annotation, the contents of this tab will be gradually replaced by the Wikipedia tab.

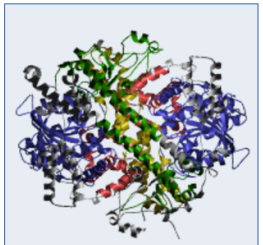
Type I restriction modification DNA specificity domain

Provide feedback

This domain is also known as the target recognition domain (TRD). Restriction-modification (R-M) systems protect a bacterial cell against invasion of foreign DNA by endonucleolytic cleavage of DNA that lacks a site specific modification. The host genome is protected from cleavage by methylation of specific nucleotides in the target sites. In type I systems, both restriction and modification activities are present in one heteromeric enzyme complex composed of one DNA specificity subunit (this family), two modification (M) subunits and two restriction (R) subunits [2].

Literature references

1. Bickle TA, Kruger DH; , Microbiol Rev 1993;57:434-450.: Biology of DNA restriction. PUBMED:8336674 EPMC:8336674
2. Janscak P, Bickle TA; , J Mol Biol 1998;284:937-948.: The DNA recognition subunit of the type IB restriction-modification enzyme EcoAI tolerates circular permutions of its polypeptide chain. PUBMED:9837717 EPMC:9837717



Example structure

PDB entry 2Y7C: Atomic model of the Ocr-bound methylase complex from the Type I restriction-modification enzyme EcoKI (M2S1). Based on fitting into EM map 1534.

View a different structure: 2Y7C

Страница доменных архитектур белков, содержащих **Methylase_S**

Family: *Methylase_S* (PF01420)

38 architectures 7284 sequences 3 interactions 2296 species 14 structures

- Summary
- Domain organisation
- Clan
- Alignments
- HMM logo
- Trees
- Curation & model
- Species
- Interactions
- Structures

Jump to... enter ID/acc Go

Domain organisation

Below is a listing of the unique domain organisations or architectures in which this domain is found. More...

There are 2509 sequences with the following architecture: Methylase_S x 2
D3S0T0_FERPA [Ferroglobus placidus (strain DSM 10642 / AEDII12DO)] Restriction modification system DNA specificity domain protein {ECO:0000313|EMBL:ADC66321.1} (421 residues)
Methylase_S Methylase_S
Show all sequences with this architecture.

There are 1903 sequences with the following architecture: Methylase_S
S0GN70_9PORP [Parabacteroides goldsteinii dnLKV18] Uncharacterized protein {ECO:0000313|EMBL:EOS20219.1} (378 residues)
Methylase_S
Show all sequences with this architecture.

There are 108 sequences with the following architecture: N6_Mtase, Methylase_S
W5WMF1_9PSEU [Kutzneria albida DSM 43870] Uncharacterized protein {ECO:0000313|EMBL:AHI02043.1} (623 residues)
N6_Mtase Methylase_S
Show all sequences with this architecture.

There are 40 sequences with the following architecture: N6_Mtase, Methylase_S x 2
A0A0H1S640_9MOLU [Mycoplasmataceae bacterium RV_VA103A] Uncharacterized protein {ECO:0000313|EMBL:KLL02790.1} (1079 residues)
N6_Mtase Methylase_S Methylase_S
Show all sequences with this architecture.

There are 19 sequences with the following architecture: HSDR_N_2, N6_Mtase, Methylase_S x 2
I3CEN9_9GAMM [Beggiatoa alba B18LD] Type I restriction-modification system methyltransferase subunit {ECO:0000313|EMBL:EIJ42082.1} (1146 residues)
N6_Mtase Methylase_S Methylase_S
Show all sequences with this architecture.

Выравнивания Seed и Full

Seed та выборка последовательностей,
по которой строился HMM-профиль – математическое описание выравнивания, позволяющее искать похожие домены в новых последовательностях

КОНЕЦ ПРЕЗЕНТАЦИИ

Для чего строят множественные выравнивания?

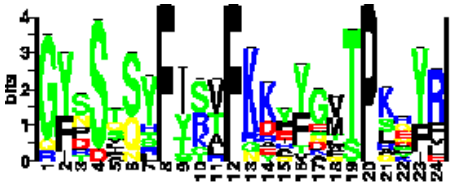
```
                        *        20         *        40         *        60
ROB_ECOLI  : RYQFWHDFLGNAPTIPPVLYGLNETRPSQDKDDEQEVFYTTALAQDQADGYVLTGHPVMLQ :  61
GADX_ECOLI : EW---------TLARIASELLMSPSLLKKKLREE-ETSYSQLLTECRMQ----RALQLIVI :  47
ENVY_ECOLI : YW---------NLRIVASSLCLSPSLLKKKLKNE-NTSYSQIVTECRMR----YAVQMLLM :  47
YDEO_ECOLI : PW---------KLKDICDCLYISESLLKKKLKQE-QTTFSQILLDARMQ----HAKNLIRV :  47
APPY_ECOLI : QW---------HLKDIAELIYTSESLIKKRLRDE-GTSFTEILRDTRMR----YAKKLITS :  47
GADW_ECOLI : RW---------YLRDIAERMYTSESLIKKKLQDE-NTCFSKILLASRMS----MARRLLEL :  47
XYLR_ECOLI : HYIRNHACKGIKVDQVLDAVGISRSNLEKRFKEEVGETIHAMIHAEKLE----KARSLLIS :  57
YDEC_BACSU : NWIHLHYVEKITLEDIAKAGQLSRSECCRYFKRMLNKTPLRYVMDYRIQ----KSLLLLQH :  57
              5                    s 3        e        6           a  66
```

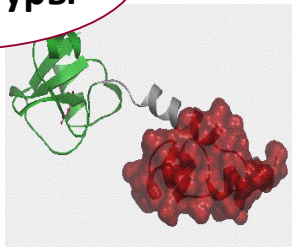
позволяет найти общее

мотивы, паттерны, профили

поиск активного центра

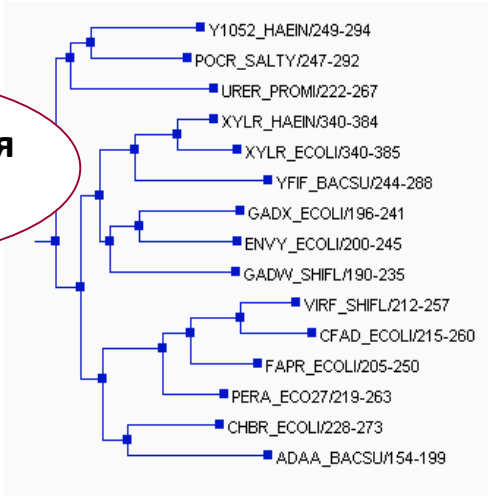
предсказание 3D-структуры





позволяет оценить эволюционные отношения

реконструкция эволюции



Построение множественных выравниваний — необходимый этап решения многих задач молекулярной биологии

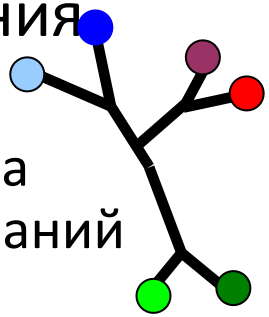
2. Jalview – см. инструкцию на сайте kodomo. В ней есть почти все нужное.

Алгоритмы множественного выравнивания

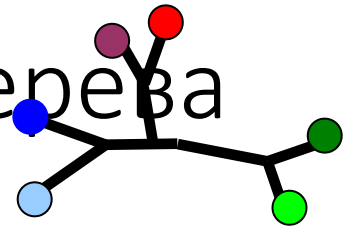
Иерархические и блочные

Иерархическое выравнивание многих последовательностей

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



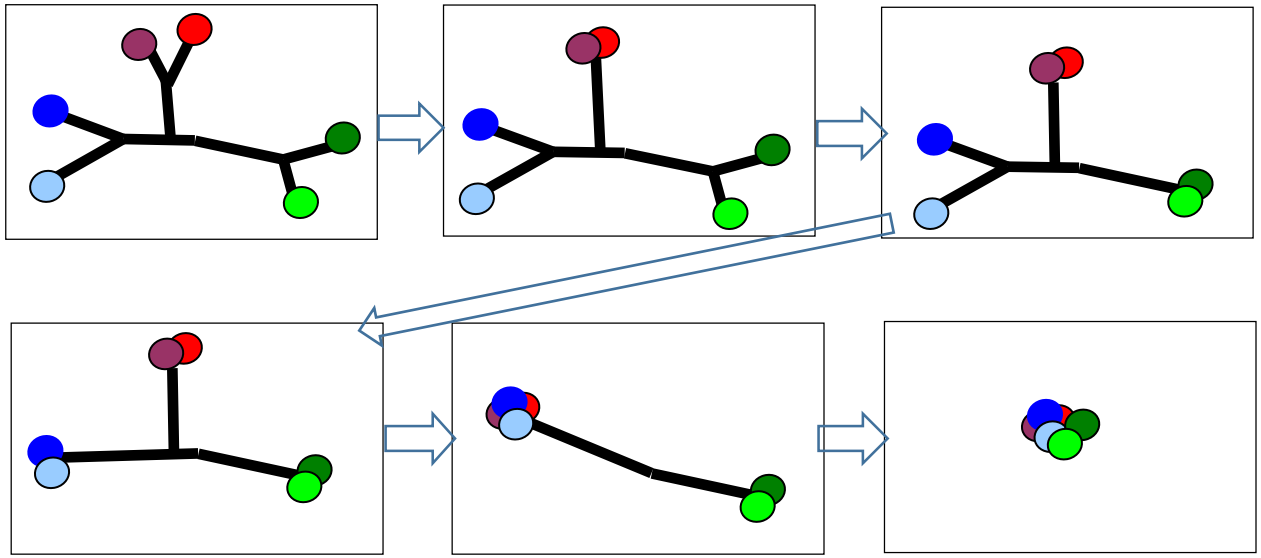
Построение направляющего дерева



- 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

Итерация выравнивания двух выравниваний

- Лист дерева – последовательность или выравнивание нескольких последовательностей



- На каждой итерации для выравнивания выбираются самые близкие “листья” – выравнивания
- Длина ветки до построенного выравнивания равна длине ветки до развилки плюс половина расстояния между листьями

“Рафинирование” выравнивания

- Главная беда иерархического выравнивания – ошибка на первых итерациях никогда не будет исправлена!
- Значит, нужны алгоритмы исправления ошибок – “рафинирования” выравнивания (refinment)
  - На вход подается множественное выравнивание
  - Каким-либо образом выравнивание делится на две части по горизонтали
  - эти части заново выравниваются друг против друга
    - если новое выравнивание лучше (нужен параметр!), то оно принимается, иначе – остается прежнее
  - простейший вариант: каждая последовательность выравнивается против всех остальных пока выравнивание не перестанет улучшаться

Блочное выравнивание

- Идея: найти в последовательностях сходные участки без гэпов. Построить из них блоки
- Найти непротиворечивый набор блоков
- Выровнять участки между блоками

Такой подход реализован в
PSI-BLAST (Position-Specific Iterated BLAST)
для улучшения результатов поиска

Алгоритмы множественного выравнивания

- Не решают формальную задачу – в отличие от алгоритмов парного выравнивания
- Нет гарантии, что построенное выравнивание превосходит все остальные по какому-то параметру, оценивающему качество выравнивания
  ЭВРИСТИЧЕСКИЕ АЛГОРИТМЫ!
- Есть разные параметры качества выравнивания, например, сумма весов парных выравниваний.
  - Построение множественного выравнивания с помощью динамического программирования возможно
  - Число операций $C*n^k$ где k – число последовательностей, n – средняя длина входных последовательностей (геометрическое среднее). Прикиньте сколько времени потребуется, если n = 100 и k = 100

Программы

- Muscle
- Mafft – быстрая
- T-coffee, M-coffee (http://www.tcoffee.org/Projects/tcoffee/)
  - умеет интегрировать разную информацию: парные локальные выравнивания, сходство 3D структур некоторых белков – если известны.
- ClustalW, более современная версия этого агоритма --- ClustalO
- Probcons
- DiAlign – блочное, начинает с парных блоков
- AliBee – блочное
- И много других

конец