Программы множественного выравнивания

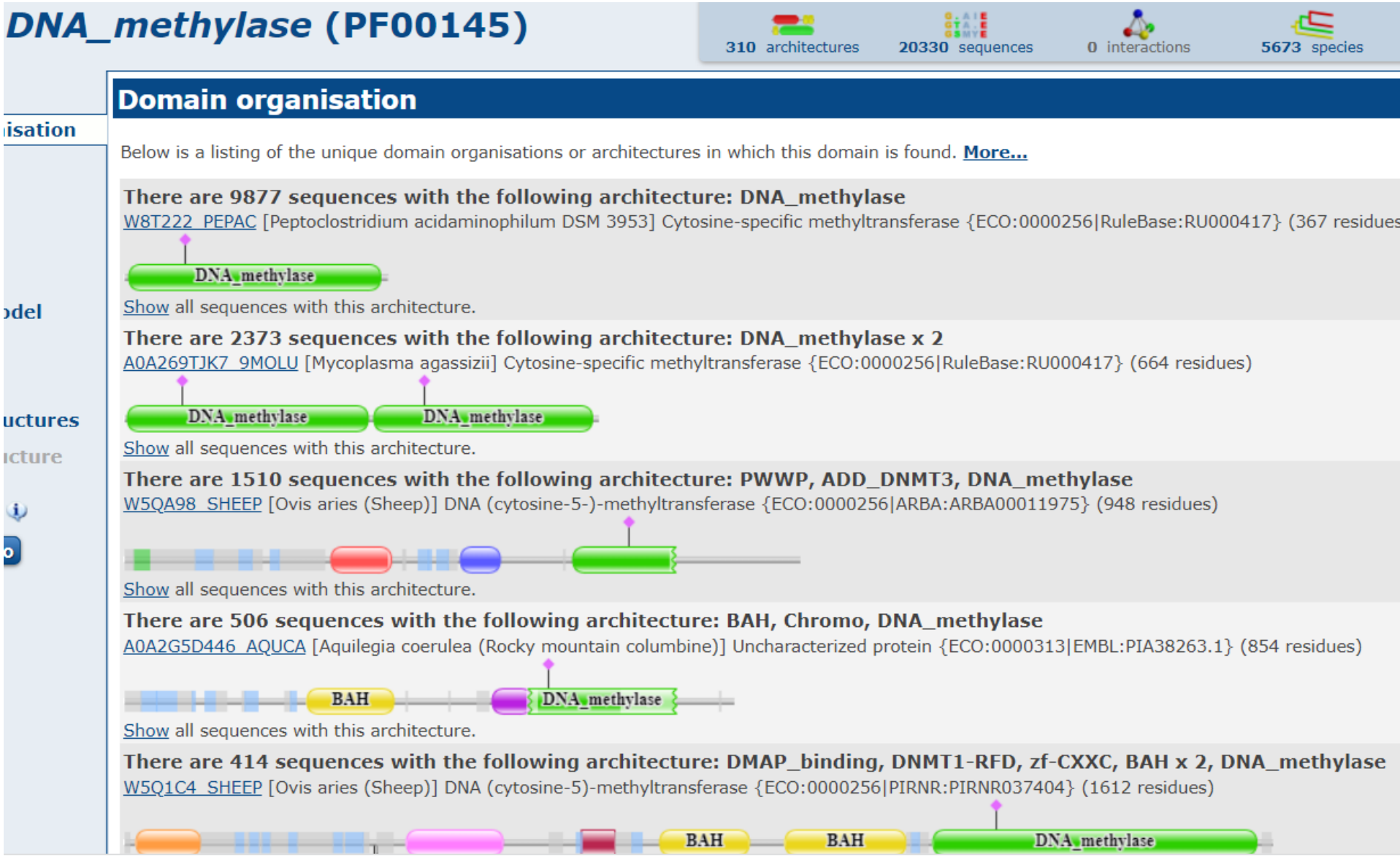
На материале последовательностей гомологичных белков.

MSA – Multiple Sequence Alignment

План

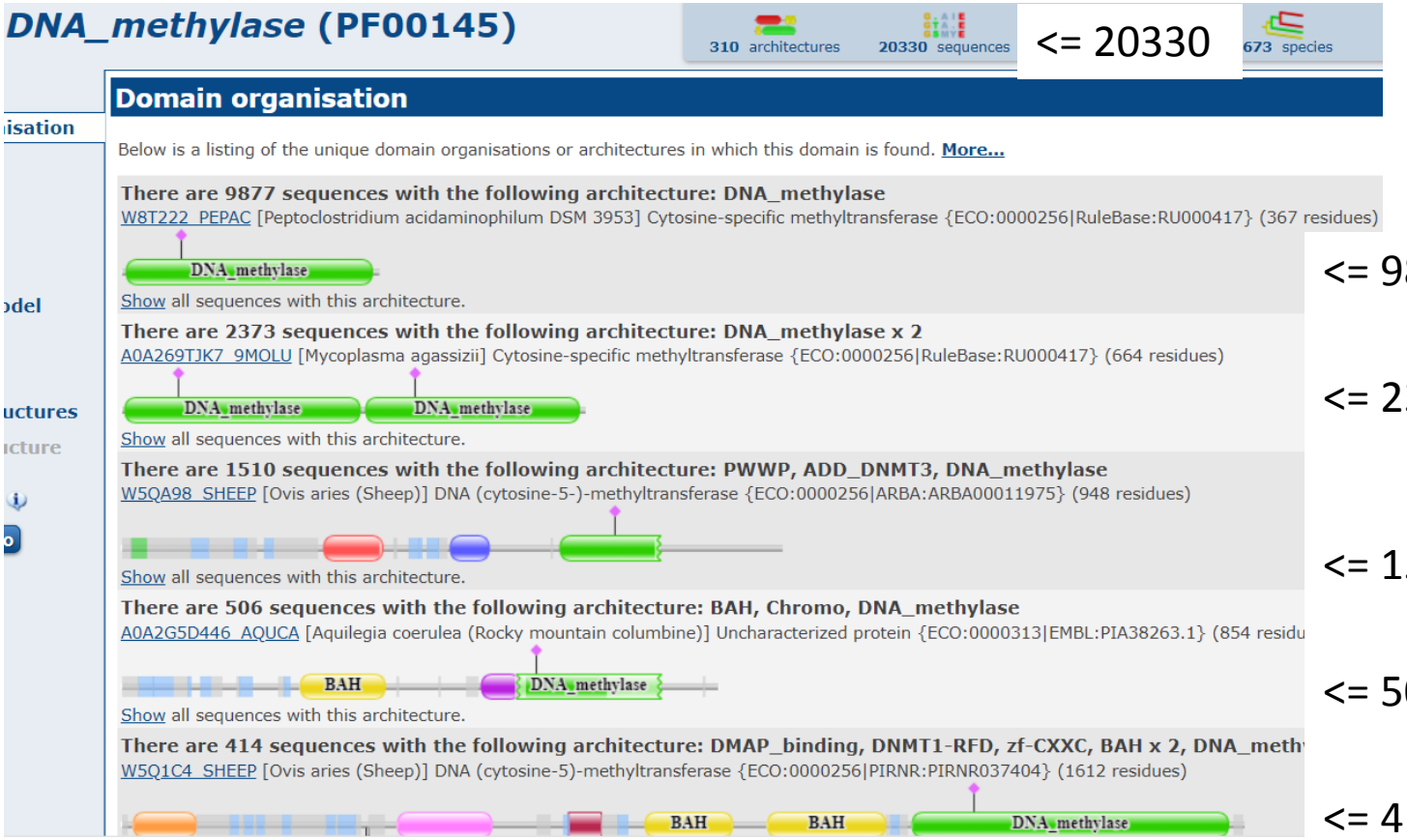
1. +Гомология
2. +Чего хотим от MSA
   - Реконструкцию эволюция
   - Реконструкцию сходства хода полипептидных цепей 3D структур
   - Иногда эволюция ≠ 3D
3. Алгоритмы и программы
   - Прогрессивное выравнивание
   - Итеративное
   - Регрессивное
4. Сравнение выравниваний тех же последовательностей
5. Верификация MSA.
   - 3D совмещение
   - Базы данных (MSA benmarks)
6. Проект блочного MSA

1. Гомология белков?



Гомология белков:

- С одинаковой доменной архитектурой
- Доменов из одного семейства из разных белков



2. Чего хотим в результате MSA гомологичных доменов белка?

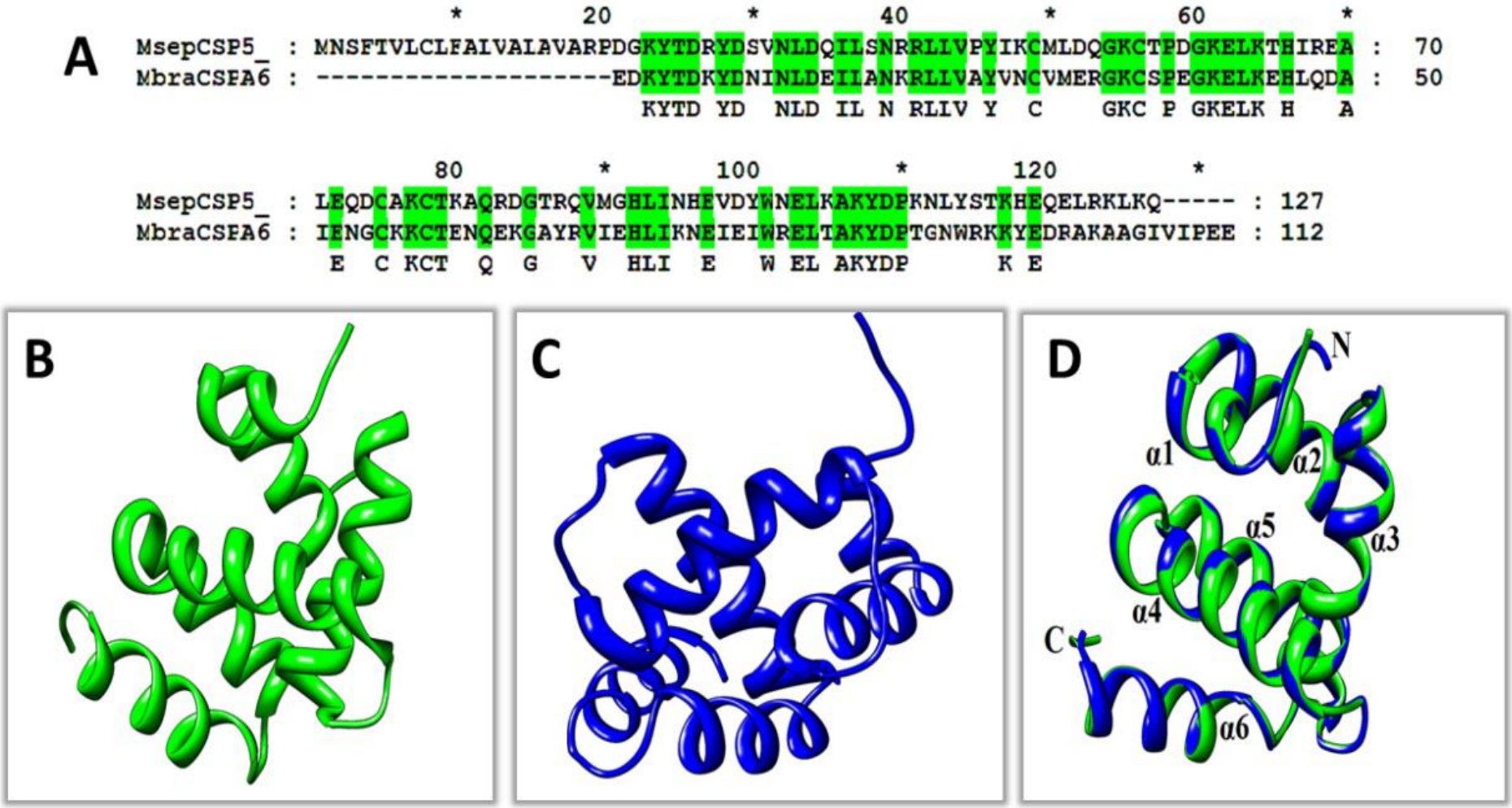
- Семейство доменов – единица локальной эволюции белков (локальная = точечные замены, делеции и вставки; по большей части короткие)
- В эволюции редко, но бывают также крупные перестройки (слияния и разделения белков, дупликации, рекомбинации, крупные делеции, вставки; много примеров циклических перестановок)

1. Хотим от MSA реконструкцию эволюции доменов семейства от общего предка: в колонке выравнивания потомки одного а.к.о. общего предка

Чего хотим в результате MSA гомологичных доменов белка?

- 3D структура белков консервативнее их последовательности.

2. Хотим от MSA чтобы C-alpha атомы а.к.о. из одной колонки при совмещении 3D структур этих белков оказались в одномположении



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

2.3 Хотим в результате MSA гомологичных доменов белка:

1. - реконструкцию эволюции доменов семейства от общего предка: в колонке выравнивания потомки одного а.к.о. общего предка
НЕ ПРОВЕРЯЕМО!

2. - чтобы C-alpha атомы а.к.о. из одной колонки при совмещении 3D структур этих белков оказались в одном положении
ПРОВЕРЯЕМО:

a. Подкопить деньжат на PCA всех белков выравнивания

b**. Проверить на белках с известной 3D**

c. Предсказать 3D структуры всех белков и сравнить их.

БЕДА: основой предсказания служит сходство фрагментов последовательности с фрагментами белков с известной 3D

Хотим в результате MSA гомологичных доменов белка

Хотеть не вредно!

Верно ли, что эволюция = сохранение хода полипептидной цепи в 3D?

Отчасти, да. Убеждают примеры совмещаемых структур со слабым сходством последовательностей.

Отчасти нет. См. ниже.

Chatzou et al., Multiple sequence alignment modeling: methods and applications, 2016

If the alignment is meant to be a structural model, aligned residues should have comparable positions in their respective 2D or 3D structures. If the alignment is functional, as may happen when analyzing genomic data, aligned positions are expected to support similar functions. Even though it is reasonable to expect a significant overlap between these criteria,

It must be stressed that the complexity of evolutionary forces is such that their full agreement cannot be taken for granted.

For instance, two structures may be similar as a consequence of convergent evolution but non-homologous from an evolutionary point of view.

Например, сравнительно длинные вставки в одном и том же месте структуры могу произойти независимо и не быть гомологичными. Возникают потому, что вставки в этом месте не нарушают структуру и функцию белка

ПРИМЕР

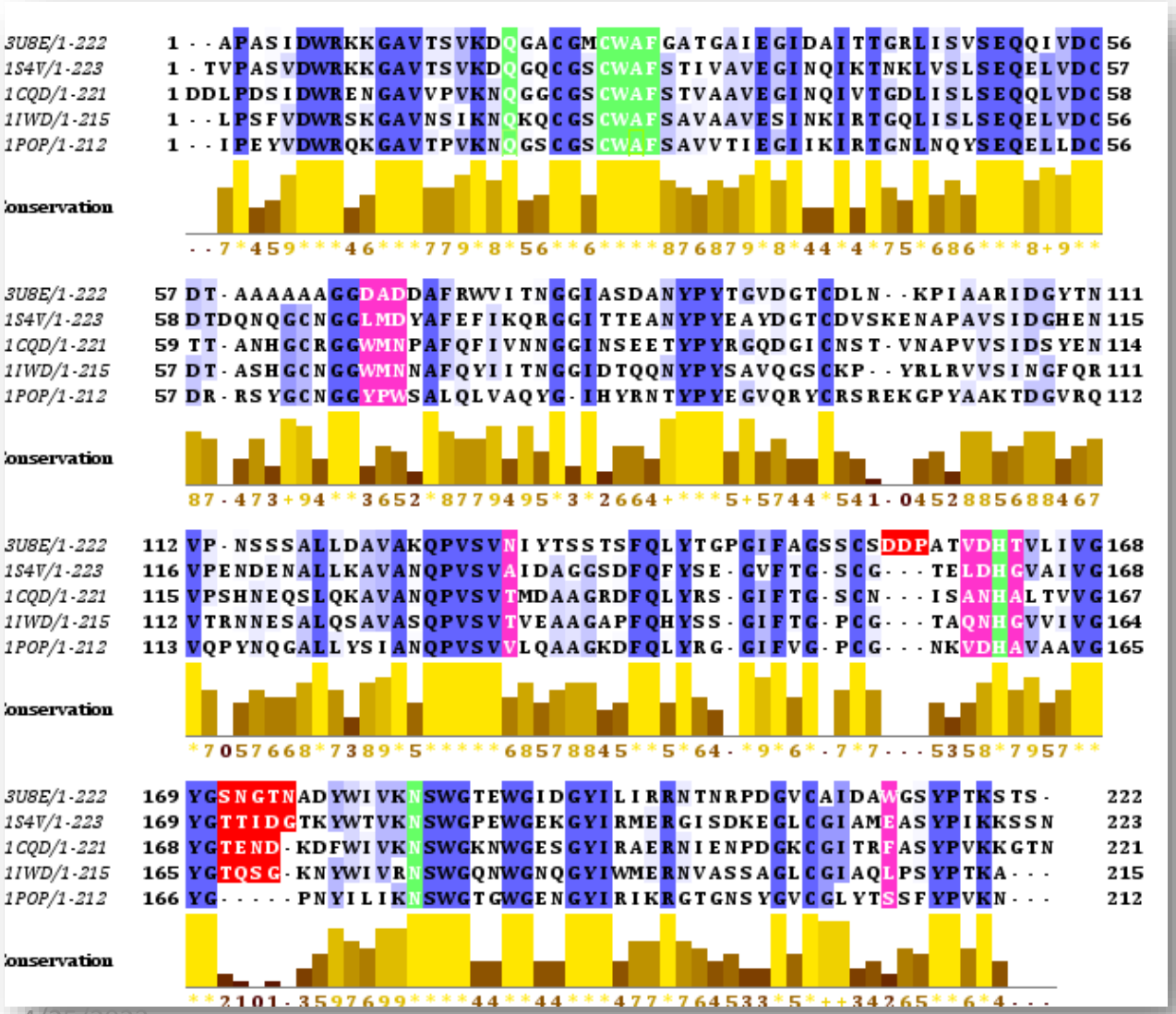
Younas A et al., A chemosensory protein MsepCSP5 involved in chemoreception of oriental armyworm *Mythimna separata*. *Int J Biol Sci*. 2018

Mythimna separata is one of the most serious pests affecting the quality andquantity of crop yield [25, 26]

The binding sites of MsepCSP5 to candidate volatiles were well predicted by three-dimensional structure modeling and molecular docking experiments.

Pursuing further, biological activities of *M. separata* to highly bound compounds elicited strong behavioral responses, such as alcoholic compounds displayed strong attractiveness whereas terpenes showed repellency to *M. separata*

Multiple Sequence Alignment of *Cs*CP against Plant CPs



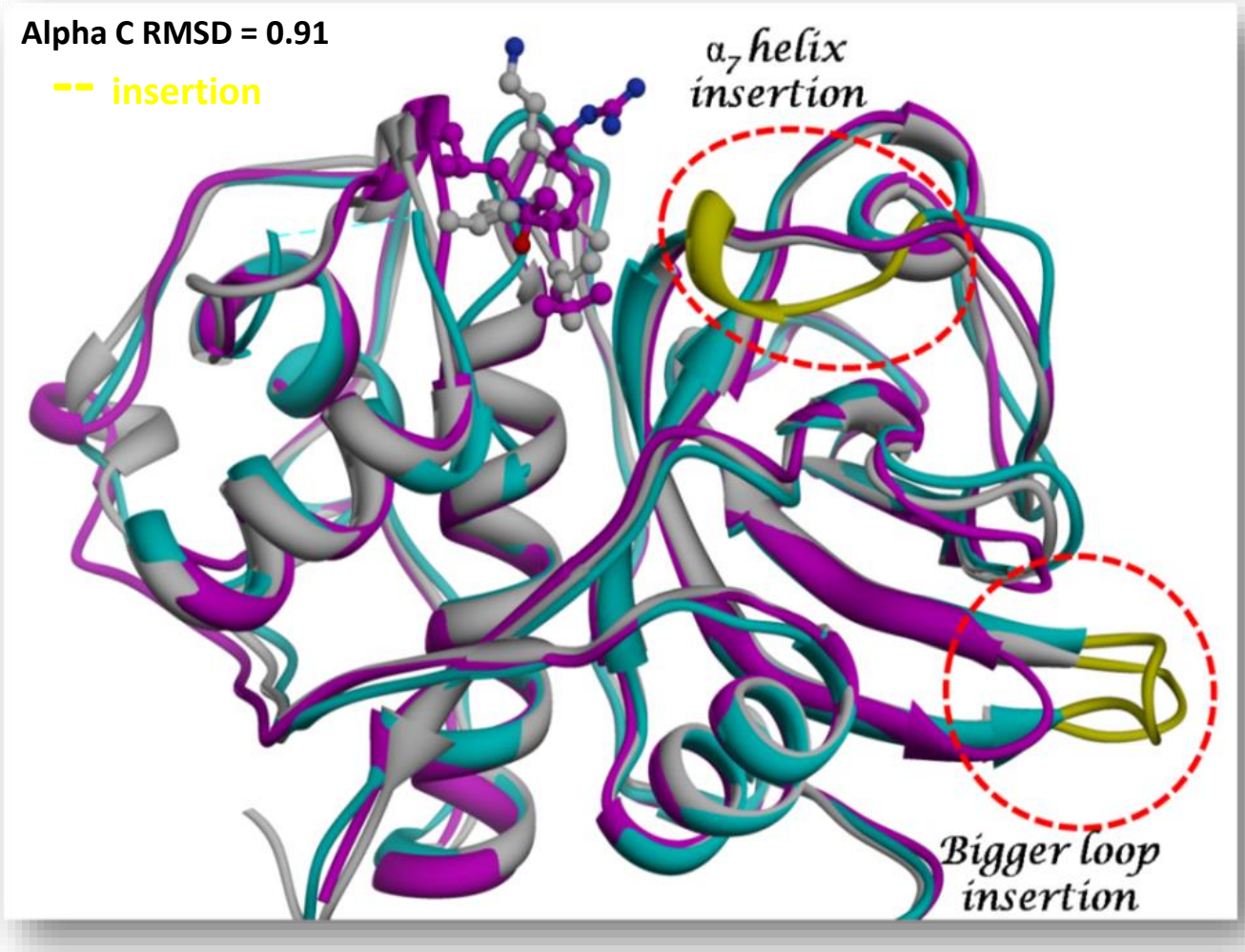
-- **Conserved**

-- **Catalytic**

-- **Active site**

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

Conserved Structure of *Cs*CP Superimposition with Plant CPs

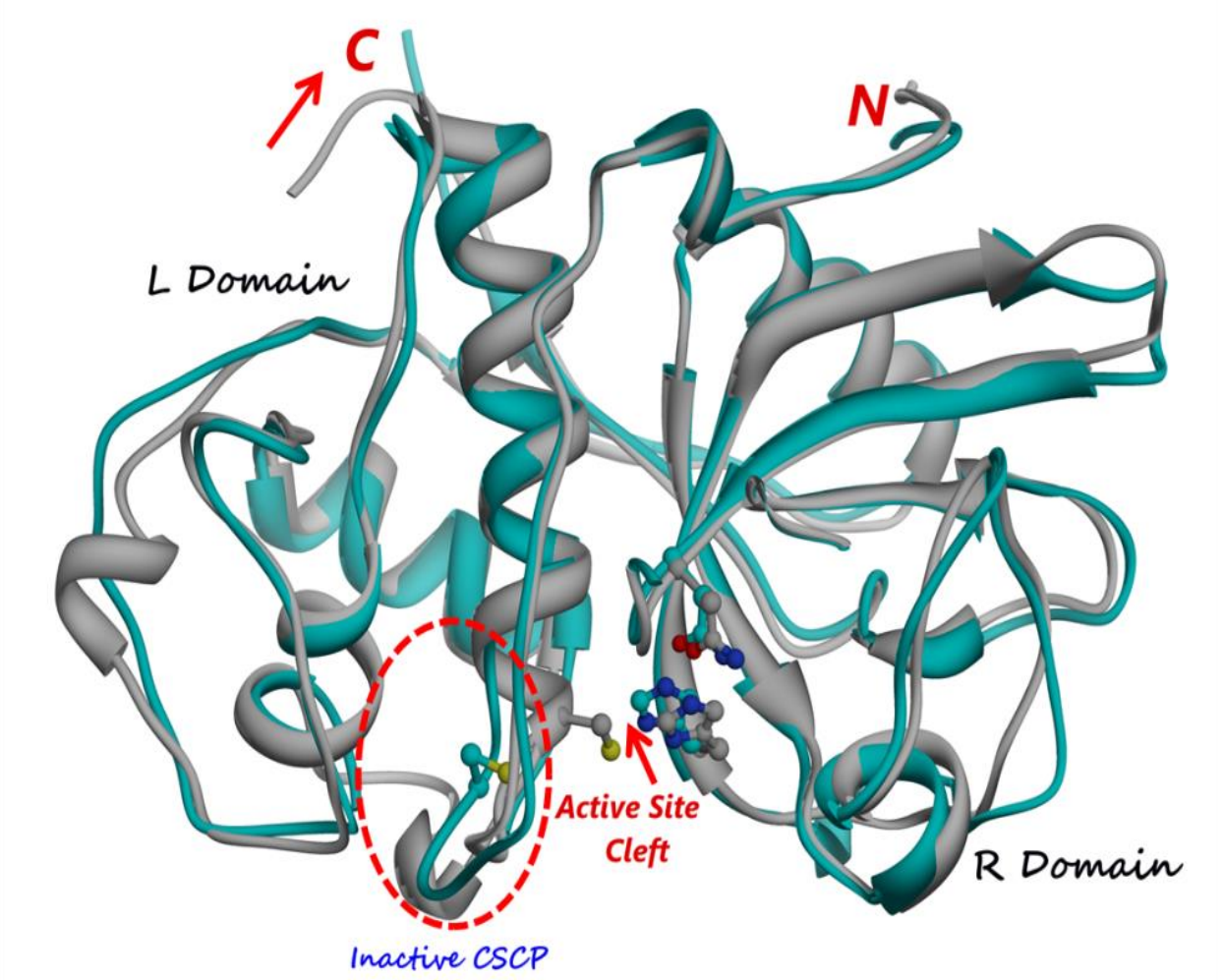


-- *3U8E (Crocus sativus)*

-- 1S4V (*Ricinus communis*)

-- 1POP (*Carica papaya*)

Overall Topology of *Cs*CP



-- Inactive Form

-- Active Form

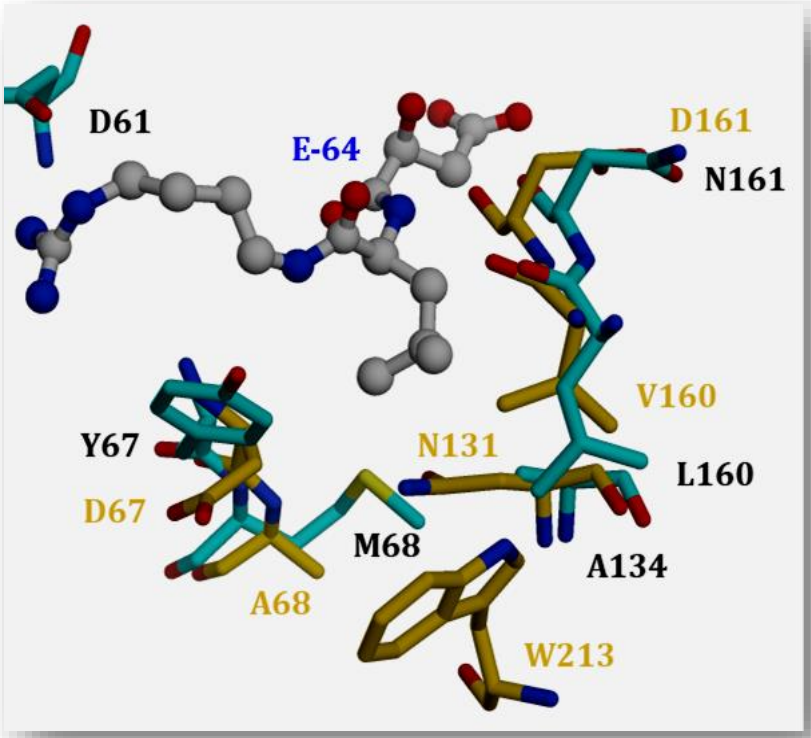
Catalytic Triad:
C25, **H162**, and **N183**
(Active Site Cleft)

Inactive Form:
C25 bridged with C22

Active Form:
C22 bridged with C64
C25 is free

KEY Findings …. Identification of S_2 Pocket Residues of CSCP

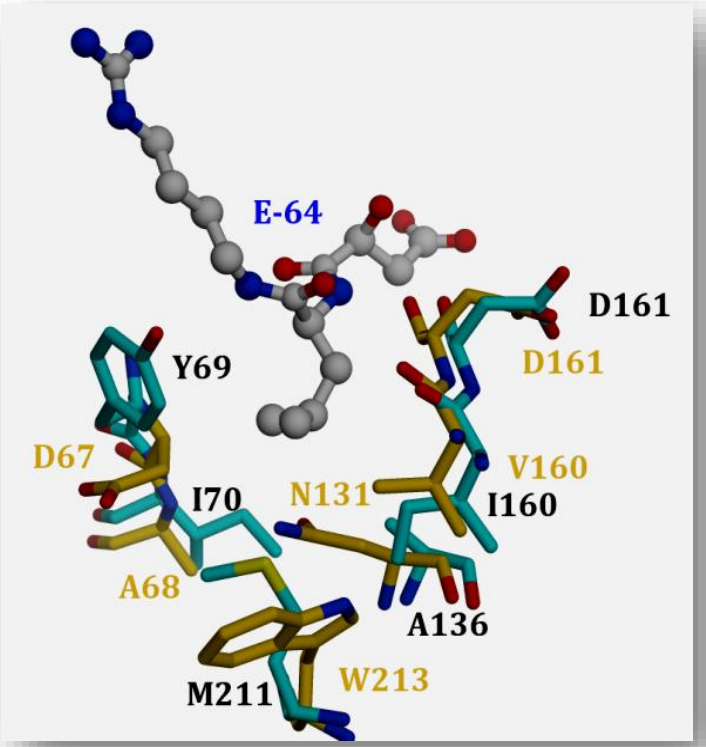
Cathepsin K-E64



-- Others

-- CsCP

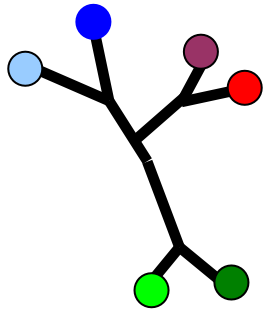
Actinidin-E64



3. Алгоритмы и программы MSA

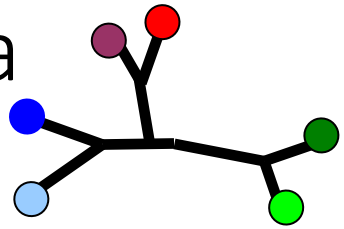
- Прогрессивное выравнивание (ClustalW)
- Итерактивное выравнивание (MUSCLE, PRALINE, IterAlign)
- Методы, основанные на согласованности (MAFFT, ProbCons)
- Методы, использующие структурные данные

Иерархический алгоритм выравнивания многих последовательностей



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

Построение направляющего дерева



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

Книга
Multiple Sequence Alignment Methods and Protocols

Edited by Kazutaka Katoh

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

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

Free

Алгоритмы и программы MSA

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

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

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

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

Алгоритмы и программы MSA

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

One effective solution is to correct this type of mistake is iterative refinement [5–7]. The procedure is:
(i) construct an initial MSA;
(ii) divide the MSA into two groups; (iii) re-align the two groups; repeat (ii) and (iii). This technique is used in Prrn5 (Chapter 5), Clustal Omega (Chapter 1), MAFFT (Chapter 11), and MSAProbs (Chapter 3).

Алгоритмы и программы MSA

Consistency

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

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

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

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

Популярные программы

1. MAFFT
2. Clustal Omega
3. ClustulW
4. Muscle
5. Probcons
6. ……

4. Сравнение двух выравниваний тех же последовательностей

ьььь

Какие выравнивания тех же последовательностей совпадают?

```
     1 2 3 4 5 6 7 8 9 10 11 12 13
Seq1 MKFR-SSHYA-S
Seq2 MKYRRS-HYA-S
Seq3 MEFRRRSHYA-R
```

```
     1 2 3 4 5 6 7 8 9 10 11 12
Seq1 MKF-RSSHYAS
Seq2 MKYRRS-HYAS
Seq3 MEFRRRSHYAR
```

```
     1 2 3 4 5  6 7  8  9 10111213
Seq1 MKFR-SSHYA-S
Seq2 MKYR-RSHYA-S
Seq3 MEFRRRSHYA-R
```

```
     1  2  3  4 5  6 7  8  9 101112
Seq1 -MKFR-SSHYAS
Seq2 -MKYR-RSHYAS
Seq3 -MEFRRRSHYAR
```

Колонка i выравнивания X совпадает с колонкой j выравнивания Y если в них – те же самые остатки; те же самые значит – с теми же номерами, а не с теми же буквами!

ПРИМЕР

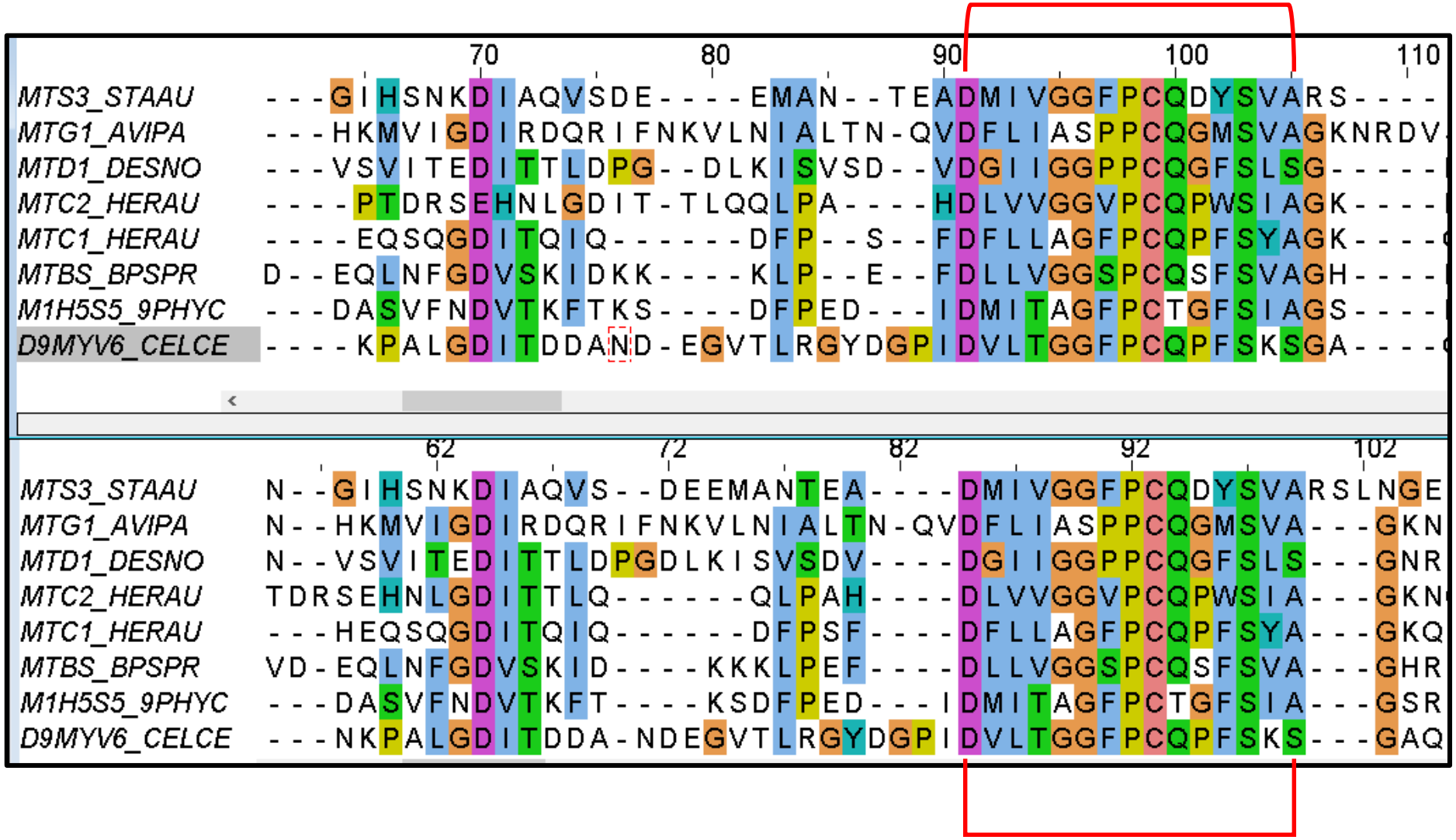
Выравнивание PF00145-seed из PFAM

Прорежённое до 8 посл-й с помощью EDIT => remove redundancy

Выравнивание с помощью MAFFT тех же 8 посл-й

Совпадение выравниваний в позициях 1 – 18. Больше нет (в показанной части)

Продолжение



Take home message

To compare alignments A and B

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

3. Если колонка букв в выравнивании А отличается хотя бы одной буквой от колонки букв в выравнивании В, то они выровнены не одинаково.

Следствие

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

Алгоритм сравнения двух выравниваний A и B тех же n последовательностей

1. Сортировка последовательностей по ID в обеих выравниваниях

2. Для выравнивания A шифруем каждую колонку i вектором A(i)

$(i; A(i)), A(i) = (s_1, s_2, …., s_n)$

n = число последовательностей

s_k зависит от того, стоит ли в колонке i в последовательности s_k буква или гэп (-);

если буква, то s_k = номер этой буквы в последовательности s_k ;

если гэп, то минус номер последней буквы прочтенной в последовательности s_k

3. Аналогично определяем (j, B(j))

Позиция i выравнивания A одинаково выровнена с позицией j выравнивания B если A(i) = B(j)

i может быть равно j, может быть не равно

4. Вычисляем список одинаково выровненных позиций (i,j).

Процент длины этого списка от длины выравнивания – мера совпадения выравниваний.

Порядок позиций (i, j) в одном из выравниваний описывает расположение совпадающих частей выравнивания.

Готовые программы для сравнения выравниваний тех же последовательностей

- Нашел одну
  VerAlign: a multiple sequence alignment assessment tool

5. Верификация MSA

- По структуре
  - BaliBase
  - Программа FATCAT множественного совмещения структур белков. Не нашел доступного сервиса.
  - Реализация FATCAT в PDB для парных сравнений работает.
- Базы эталонных выравниваний для сравнения программ MSA
  - OxBanch не удалось использовать

Le et al., 2016

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

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

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

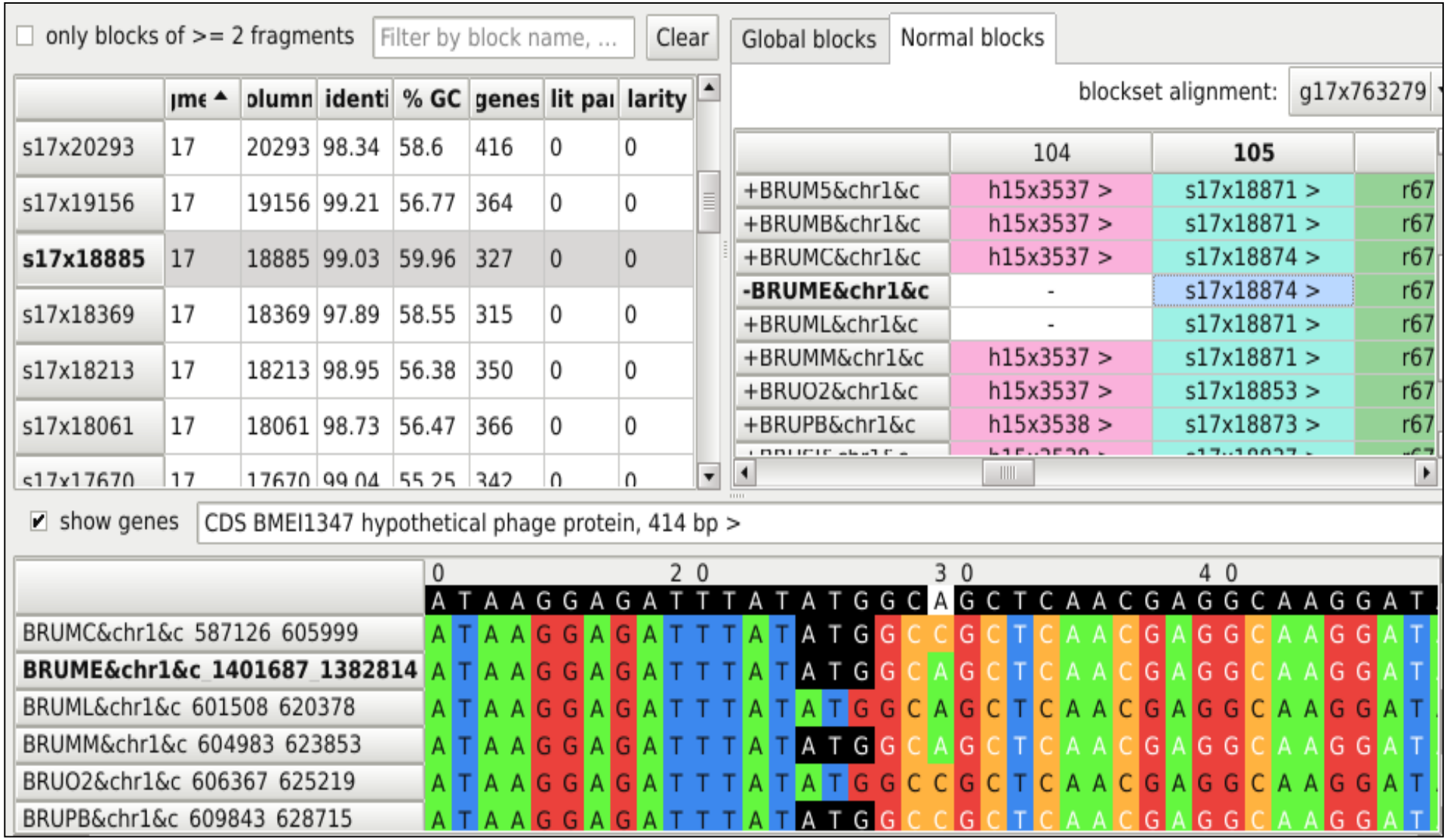
5. Проект блочного MSA

- Определение. MSA = набор блоков достоверного выравнивания во всех или части последовательностей. ААл
- Этап I: во входном выравнивании найти все блоки достоверного выравнивания.
- Этап II: найти все блоки достоверного выравнивания во входном множестве последовательностей
- Этап III: создать сервис для блочного MSA

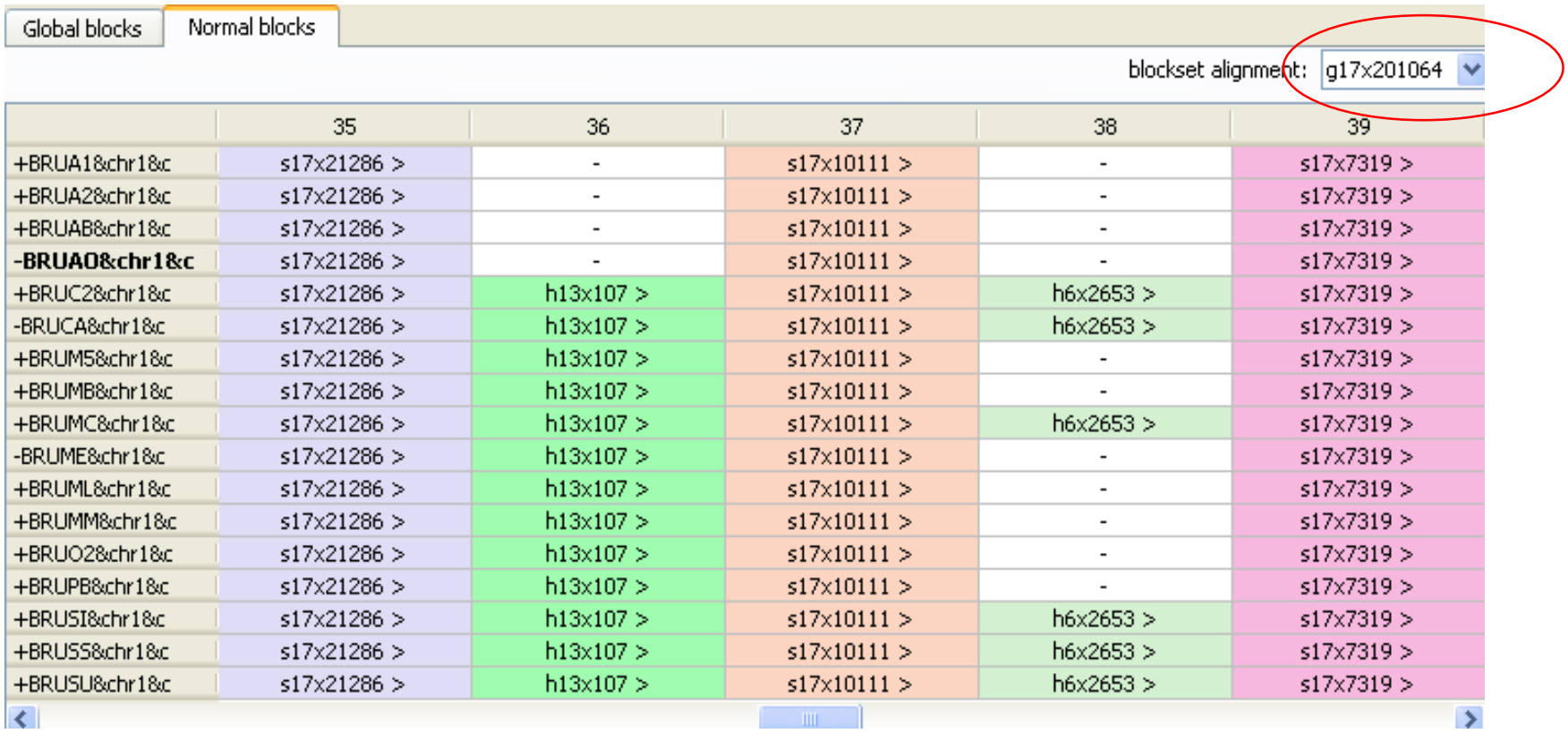
Пример: блочное выравнивание геномов

- Борис Нагаев NPGexplorer

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



Каждая ДНК представляется последовательностью блоков



Global blocks | Normal blocks

blockset alignment: g17x201064

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

Про типы блоков следующий слайд

КОНЕЦ ПРЕЗЕНТАЦИИ

ВЫРАВНИВАНИЕ