



# Алгоритмы и программы множественного выравнивания Сравнение MSA

# Цель множественного выравнивания последовательностей гомологичных доменов или белков

- ❖ Реконструкция эволюции белков от общего предка. В колонке **правильного** выравнивания стоят аминокислотные остатки (а.о.), унаследованные от а.о. общего предка всех этих белков.

*Наследуются кодоны, это подразумевается.*

- ❖ Основа для установления гомологии по сходству последовательностей.

*Сходство неравномерно на разных участках.*

В какой мере выравнивание правильно — практически не проверяемо

(не считая парочки долговременных экспериментов по эволюции в лаборатории: Ленски — *E.coli*, Кондрашов — *Schizophyllum commune*, ...)

На практике стремятся к этой цели, но следует понимать, что она недостижима.

Разные программы множественного выравнивания используют разные алгоритмы и разные эвристики.

# Как проверить правильность множественного выравнивания белков?

Искать блоки множественного выравнивания, построенного программой, которые соответствуют эволюционному?

Крайне маловероятно, чтобы у нескольких белков было столько одинаковых букв на одинаковых расстояниях друг от друга по случайным причинам. Значит, скорее всего вы этом месте выравнивания правильное.

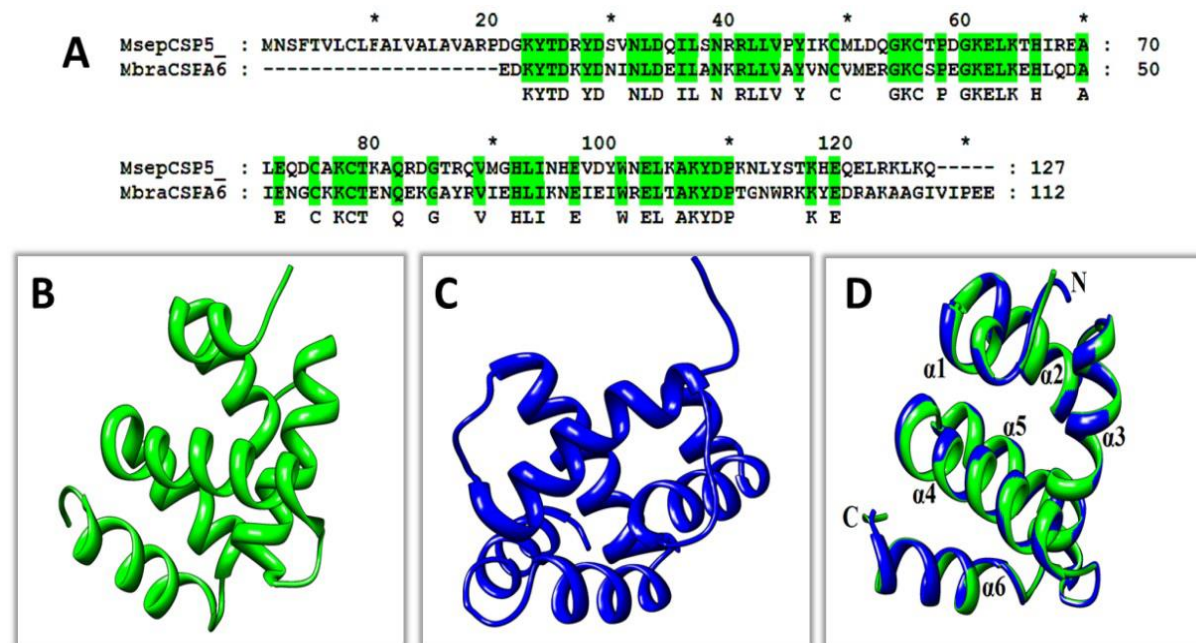
The image shows a multiple sequence alignment of ten protein sequences. The sequences are: K9Z0V0\_CYAAP/42-80, K9SST0\_9SYNE/42-80, K9SGD2\_9CYAN/42-80, K9Y0S0\_STAC7/42-80, PSBE\_RIPO1/42-80, PSBE\_SYNY3/42-80, PSBE\_GLOC7/42-80, PSBE\_OSTTA/42-79, K9W859\_9CYAN/42-80, and PSBE\_SYNE7/43-81. The alignment is shown with columns of amino acids. A red dashed line is drawn above the alignment, indicating a conserved region. The amino acids in this region are: L, A, Y, D, V, F, G, T, P, R, P, D, E, Y, F, T, Q, T, R, Q, E, L, P, I, I, N, D. The amino acids in the region below the red dashed line are: L, A, Y, D, I, F, G, T, P, R, P, N, E, Y, Y, T, N, D, R, Q, E, A, P, I, L, R, D. The amino acids in the region below the blue dashed line are: L, A, Y, D, V, F, G, T, P, R, P, N, E, Y, Y, T, L, D, R, S, D, A, P, V, L, L, D. The amino acids in the region below the green dashed line are: L, A, Y, D, V, F, G, T, P, R, P, N, E, Y, F, T, Q, D, R, L, E, L, P, I, I, N, D. The amino acids in the region below the orange dashed line are: L, A, Y, D, V, F, G, T, P, R, P, D, Q, Y, F, T, Q, E, R, Q, E, L, P, I, I, S, D. The amino acids in the region below the purple dashed line are: L, A, Y, D, A, F, G, T, P, R, P, D, E, Y, F, T, Q, T, R, Q, E, L, P, I, L, Q, E. The amino acids in the region below the cyan dashed line are: L, A, Y, D, V, F, G, T, P, R, P, D, E, Y, Y, T, Q, E, R, L, E, L, P, I, L, K, D. The amino acids in the region below the red dashed line are: L, A, Y, D, V, F, G, T, P, R, P, N, E, Y, F, T, E, E, R, Q, E, L, P, L, I, S, D. The amino acids in the region below the blue dashed line are: L, A, Y, D, V, F, G, T, P, R, P, N, E, Y, F, T, Q, D, R, V, E, A, P, I, L, S, D. The amino acids in the region below the yellow dashed line are: L, A, Y, D, A, F, G, T, P, R, P, N, E, Y, F, T, Q, D, R, T, E, V, P, I, V, S, D.

Sequence	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30					
K9Z0V0_CYAAP/42-80	L	A	Y	D	V	F	G	T	P	R	P	D	E	Y	F	T	Q	T	R	Q	E	L	P	I	I	N	D								
K9SST0_9SYNE/42-80	L	A	Y	D	I	F	G	T	P	R	P	N	E	Y	Y	T	N	D	R	Q	E	A	P	I	L	R	D								
K9SGD2_9CYAN/42-80	L	A	Y	D	V	F	G	T	P	R	P	N	E	Y	Y	T	L	D	R	S	D	A	P	V	L	L	D								
K9Y0S0_STAC7/42-80	L	A	Y	D	V	F	G	T	P	R	P	N	E	Y	F	T	Q	D	R	L	E	L	P	I	I	N	D								
PSBE_RIPO1/42-80	L	A	Y	D	V	F	G	T	P	R	P	D	Q	Y	F	T	Q	E	R	Q	E	L	P	I	I	S	D								
PSBE_SYNY3/42-80	L	A	Y	D	A	F	G	T	P	R	P	D	E	Y	F	T	Q	T	R	Q	E	L	P	I	L	Q	E								
PSBE_GLOC7/42-80	L	A	Y	D	V	F	G	T	P	R	P	D	E	Y	Y	T	Q	E	R	L	E	L	P	I	L	K	D								
PSBE_OSTTA/42-79	L	A	Y	D	V	F	G	T	P	R	P	N	E	Y	F	T	E	E	R	Q	E	L	P	L	I	S	D								
K9W859_9CYAN/42-80	L	A	Y	D	V	F	G	T	P	R	P	N	E	Y	F	T	Q	D	R	V	E	A	P	I	L	S	D								
PSBE_SYNE7/43-81	L	A	Y	D	A	F	G	T	P	R	P	N	E	Y	F	T	Q	D	R	T	E	V	P	I	V	S	D								

**А если белки накопили очень много различий?**

# А если белки накопили очень много различий?

- ❖ 3D-структура белков консервативнее их последовательности!
- ❖ Совмещение структур
  - ✓ C $\alpha$ -атомы а.о. из одной колонки при совмещении 3D структур этих белков оказываются в одном положении



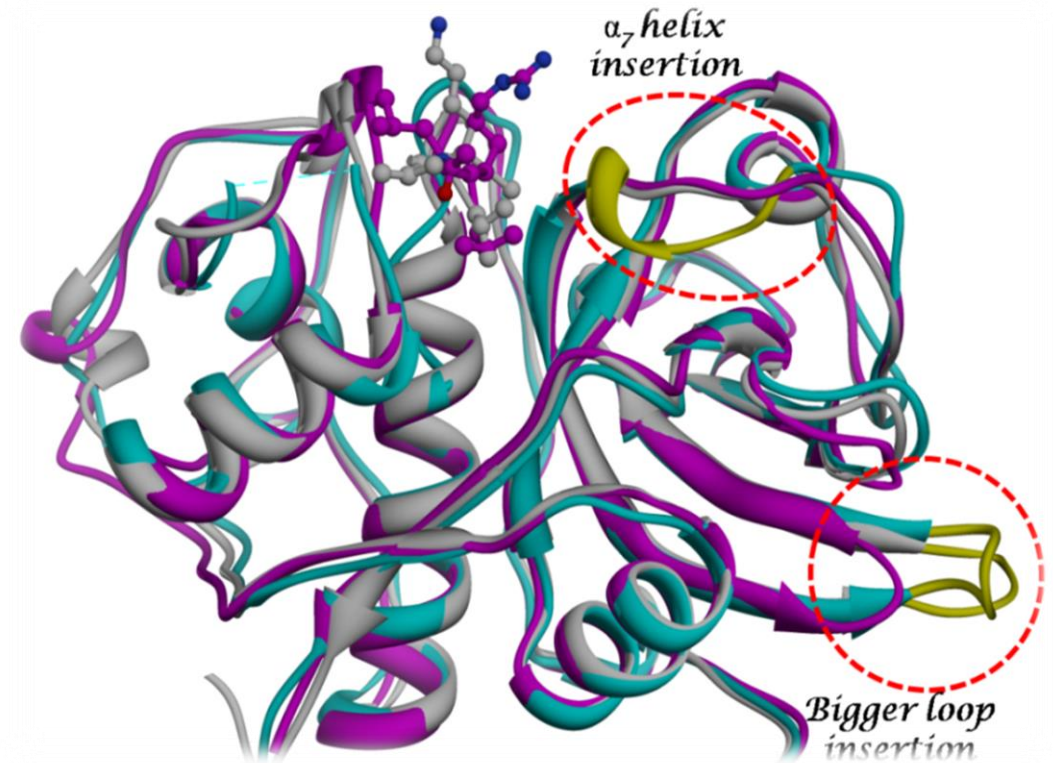
# Совмещение структур белков

- ❖ Соответствие выравнивания совмещению 3D-структур можно проверить:
    - ✓ Накопить денег и сделать рентгеноструктурный анализ всех белков выравнивания
    - ✓ **Проверить на белках с известной структурой**
    - ✓ Предсказать 3D структуры всех белков и сравнить
- Основой предсказания служит сходство фрагментов последовательности с фрагментами белков с известной 3D-структурой  
(даже при использовании *AlphaFold*, в конечном счёте)

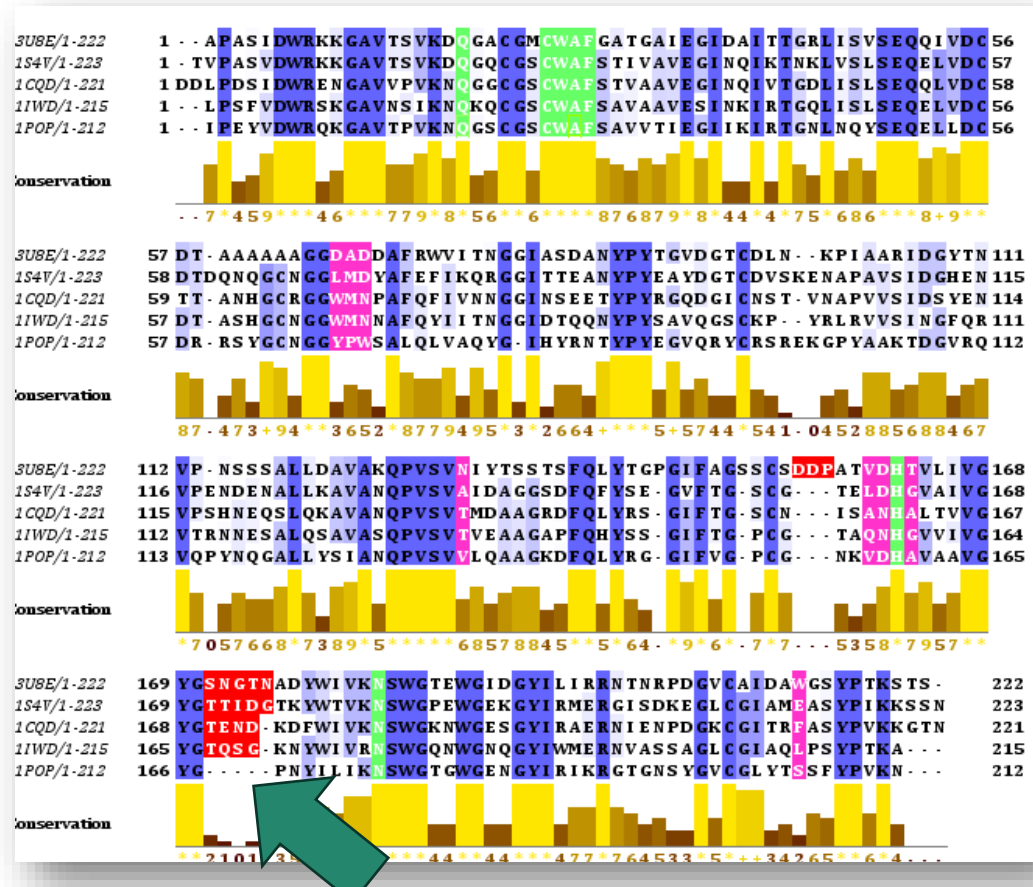
❖ Верно ли, что эволюция = сохранение хода полипептидной цепи в 3D?

# Эволюция = сохранение хода полипептидной цепи в 3D?

- ❖ Отчасти, да – есть много примеров совмещаемых структур со слабым сходством последовательностей
- ❖ Но сравнительно длинные вставки в одном и том же месте структуры могут произойти независимо и не быть ГОМОЛОГИЧНЫМИ



# Выравнивание по совмещению структур папаин-подобных цистеиновых протеаз растений



- Conserved
- Catalytic
- Active site

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

В этих колонках мы не уверены, что амк. остатки гомологичны

# Совмещение структур белков

- ❖ PDBeFold – и парное, и множественное совмещение

<https://www.ebi.ac.uk/msd-srv/ssm/cgi-bin/ssmserver>

- ❖ TM-Align, FATCAT и др. – парное сравнение, веб-сервис доступен на сайте RSCB PDB

<https://www.rcsb.org/alignment>

- ❖ Caretta – множественное совмещение

<https://github.com/TurtleTools/caretta>

# **Алгоритмы и программы множественного выравнивания**

# Динамическое программирование

Алгоритм Нидлмана — Вунша для **парного** выравнивания

Время работы и требуемая память пропорциональны произведению длин последовательностей (пишут “сложность  $O(nm)$ ”)

match = 1    mismatch = -1    gap = -1

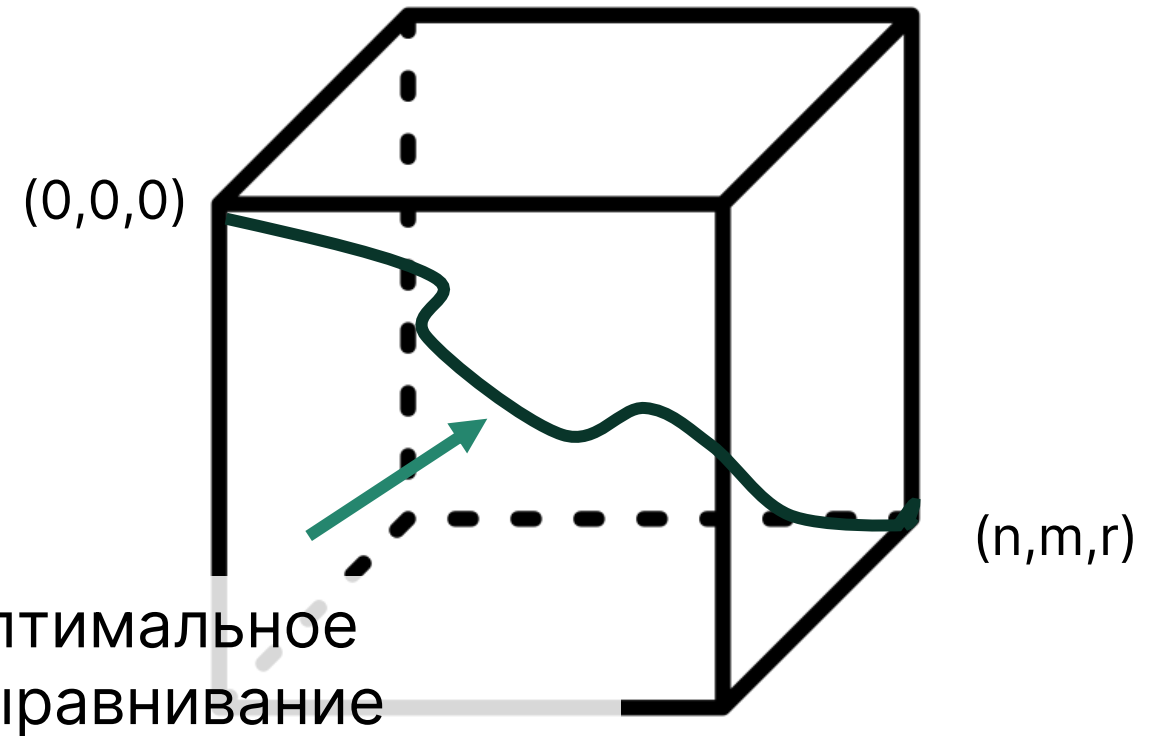
		G	C	A	T	G	C	G	
		0	-1	-2	-3	-4	-5	-6	-7
G		-1	1	0	-1	-2	-3	-4	-5
A		-2	0	0	1	0	-1	-2	-3
T		-3	-1	-1	0	2	1	0	-1
T		-4	-2	-2	-1	1	1	0	-1
A		-5	-3	-3	-1	0	0	0	-1
C		-6	-4	-2	-2	-1	-1	1	0
A		-7	-5	-3	-1	-2	-2	0	0

# Динамическое программирование

Множественное для трех последовательностей

match = 1    mismatch = -1    gap = -1

		G	C	A	T	G	C	G	
		0	-1	-2	-3	-4	-5	-6	-7
G		-1	1	0	-1	-2	-3	-4	-5
A		-2	0	0	1	0	-1	-2	-3
T		-3	-1	-1	0	2	1	0	-1
T		-4	-2	-2	-1	1	1	0	-1
A		-5	-3	-3	-1	0	0	0	-1
C		-6	-4	-2	-2	-1	-1	1	0
A		-7	-5	-3	-1	-2	-2	0	0



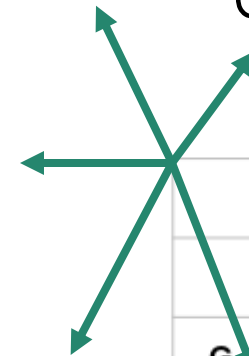
# Динамическое программирование

Множественное выравнивание  
Сложность  $O(n_1 * n_2 * \dots * n_k)$

match = 1    mismatch = -1    gap = -1

		G	C	A	T	G	C	G
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

match = 1    mismatch = -1    gap = -1



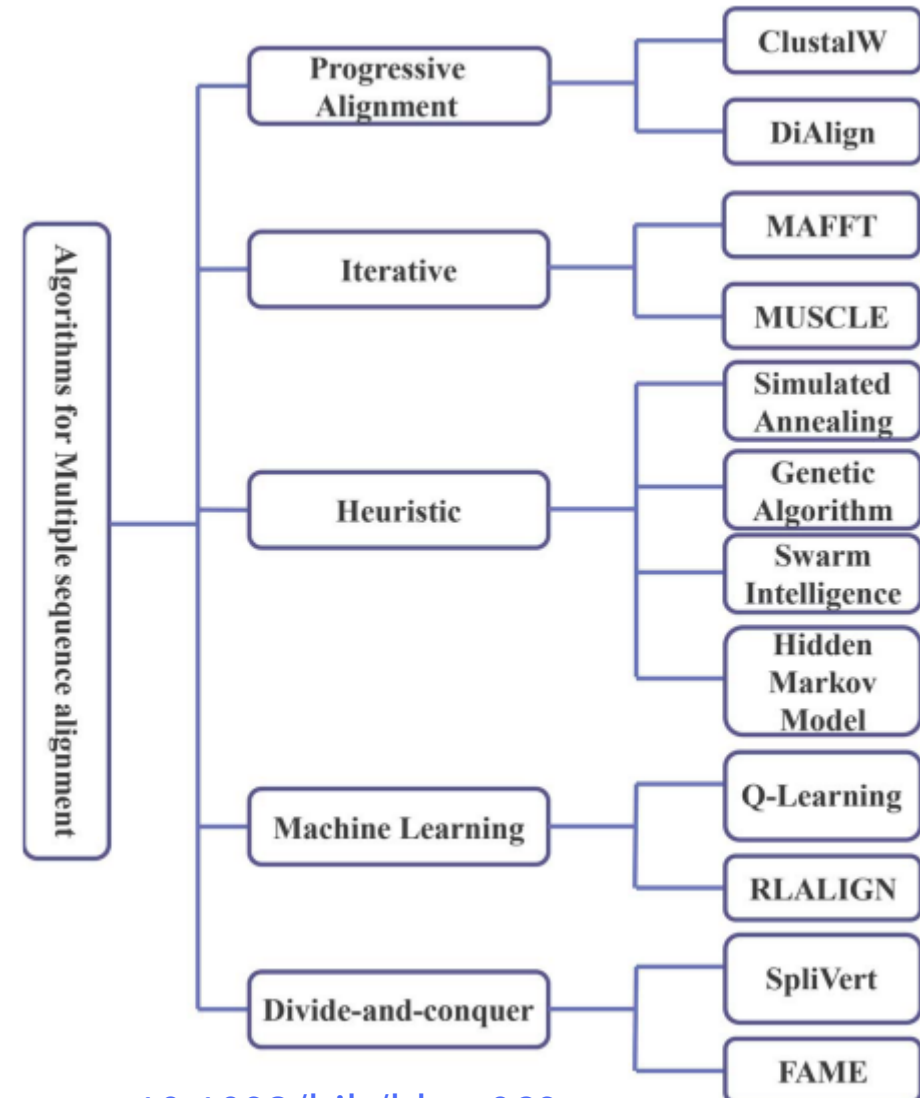
		G	C	A	T	G	C	G
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

# Нахождение оптимального выравнивания для большого числа последовательностей практически неосуществимо

Выход — эвристические алгоритмы

# Программы множественного выравнивания

- ❖ Программ много, и все они используют разные **эвристические** алгоритмы
- ❖ В литературе встречаются разные их классификации (например, [https://link.springer.com/protocol/10.1007/978-1-0716-1036-7\\_17](https://link.springer.com/protocol/10.1007/978-1-0716-1036-7_17))

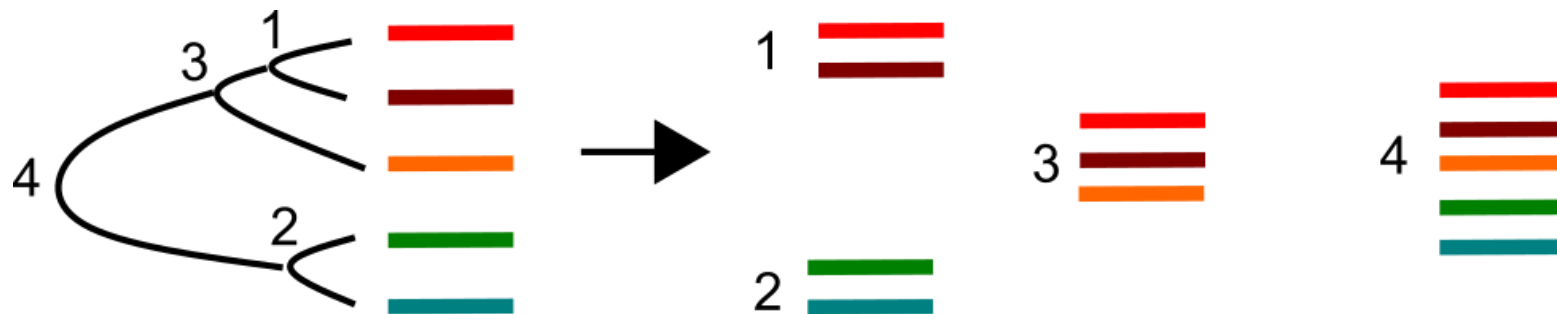


# Список популярных программ MSA

Program	Год	В Jalview	На kodomo	Website
ClustalO	2011	Yes	No	<a href="https://www.ebi.ac.uk/jdispatcher/msa/clustalo">https://www.ebi.ac.uk/jdispatcher/msa/clustalo</a>
ClustalW	1994	Yes	Yes	<a href="http://www.clustal.org/">http://www.clustal.org/</a>
MAFFT	2002	Yes	Yes	<a href="https://mafft.cbrc.jp/alignment/software/">https://mafft.cbrc.jp/alignment/software/</a>
MSAProbs	2010	Yes	No	<a href="http://msaprobs.sourceforge.net/">http://msaprobs.sourceforge.net/</a>
MUSCLE	2004	Yes	Yes	<a href="http://www.drive5.com/muscle/">http://www.drive5.com/muscle/</a>
ProbCons	2005	Yes	No	<a href="http://probcons.stanford.edu/">http://probcons.stanford.edu/</a>
T-Coffee	2000	Yes	No	<a href="https://tcoffee.crg.eu/">https://tcoffee.crg.eu/</a>
Glprobs	2015	Yes	No	<a href="https://sourceforge.net/projects/glprobs/">https://sourceforge.net/projects/glprobs/</a>
DiAlign	1998	No	Yes	<a href="http://dialign.gobics.de/">http://dialign.gobics.de/</a>
FAME	2020	No	No	<a href="http://github.com/naznoosh/msa">http://github.com/naznoosh/msa</a>
Kalign	2005	No	No	<a href="https://www.ebi.ac.uk/Tools/msa/kalign/">https://www.ebi.ac.uk/Tools/msa/kalign/</a>
NX4	2019	No	No	<a href="https://www.nx.io">https://www.nx.io</a>
PRANK	2008	No	Yes	<a href="http://wasabiapp.org/software/prank/">http://wasabiapp.org/software/prank/</a>
Probalign	2006	No	No	<a href="http://probalign.njit.edu/standalone.html">http://probalign.njit.edu/standalone.html</a>

# Прогрессивное выравнивание

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

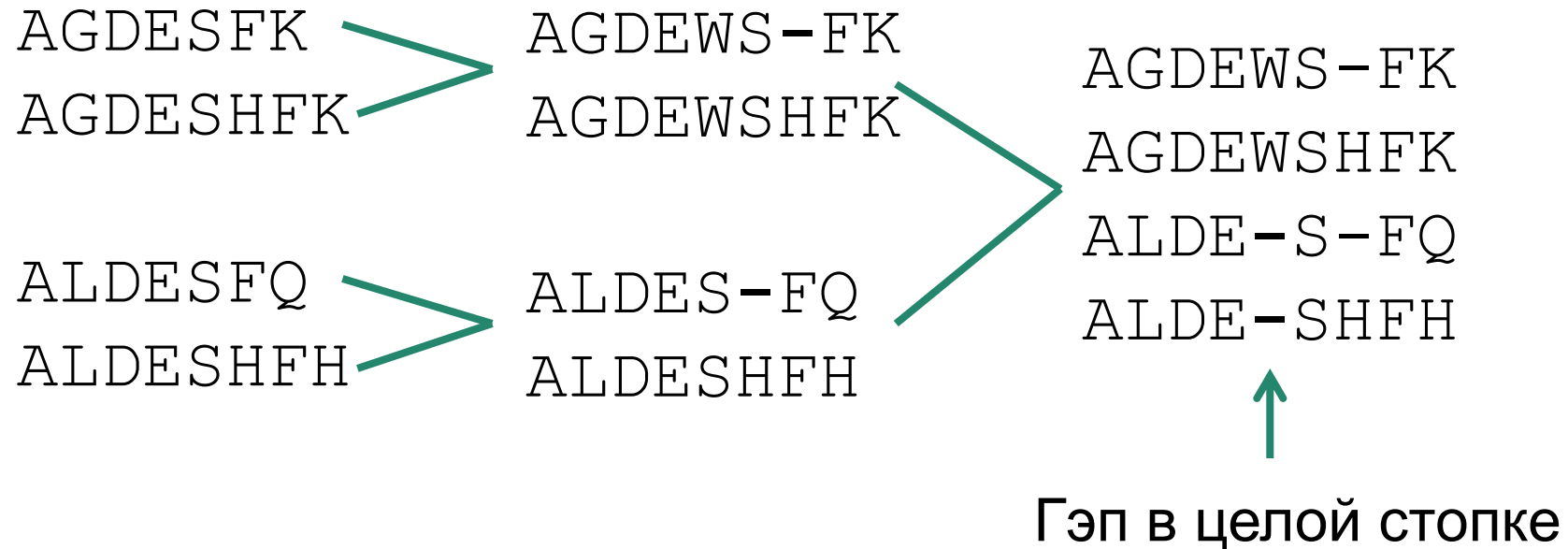


# Построение направляющего дерева

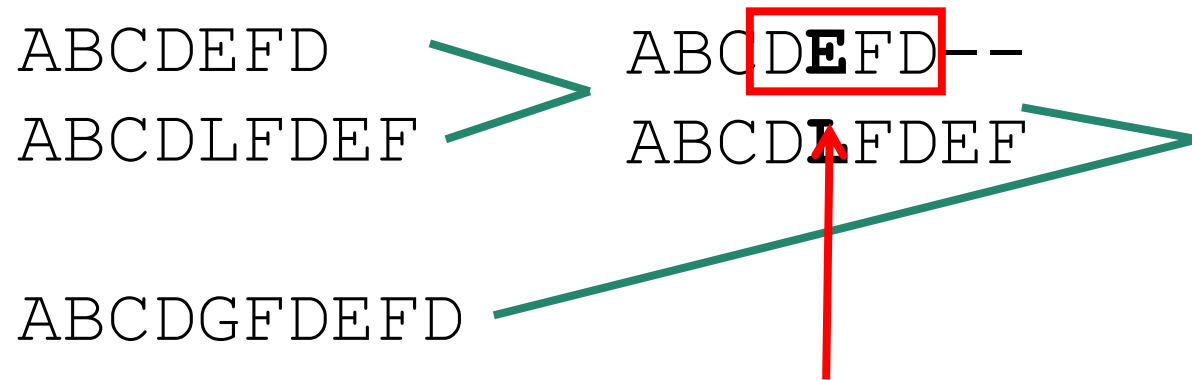
- ❖ Для ВСЕХ ПАР последовательностей строится парное выравнивание.
- ❖ Вес парного выравнивания пересчитывается в расстояние между последовательностями:
  - ✓ чем больше вес, тем меньше расстояние;
  - ✓ расстояние между совпадающими последовательностями равно 0.
- ❖ Получается матрица расстояний между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



# Пример прогрессивного выравнивания



# Проблема: если ошиблись, то ошибка никуда не денется



Но уже поздно  
передвигать этот блок

**Получилось**

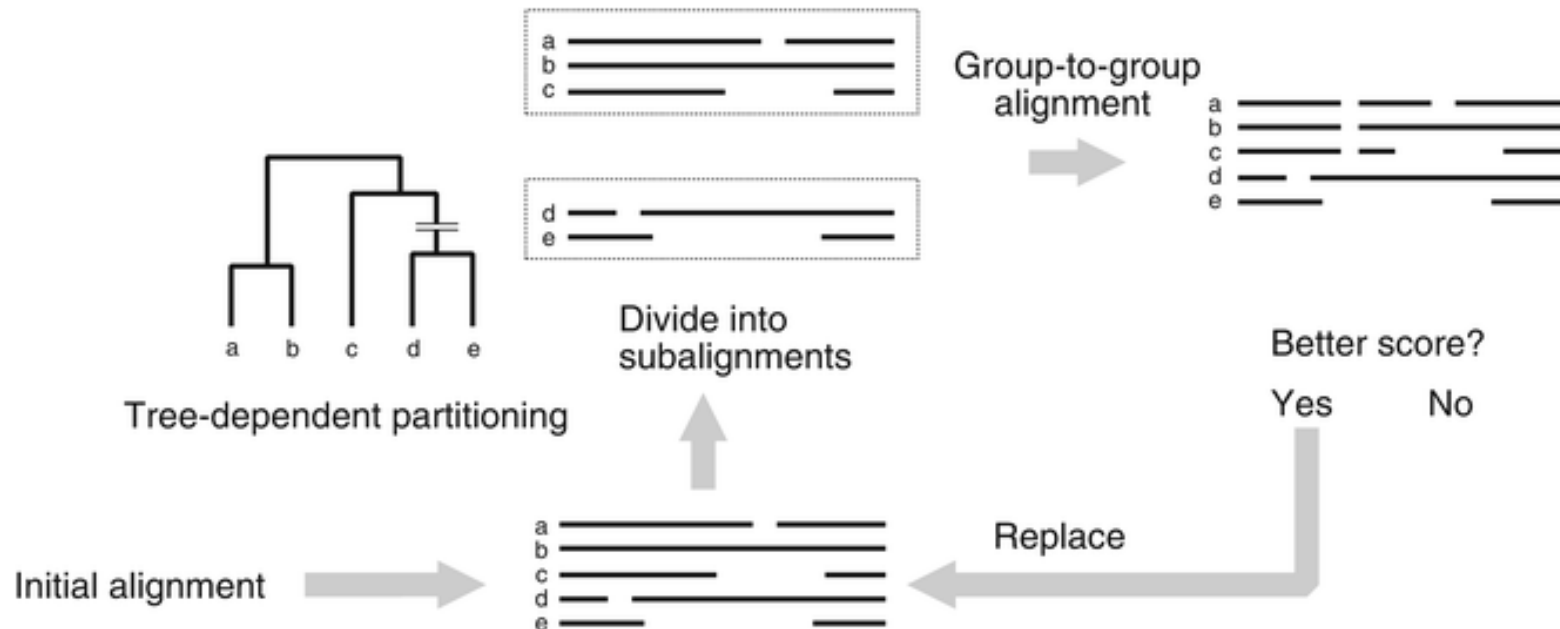
ABCDEFD---  
ABCDLFDEF-  
ABCDGFDEFD

**А хотелось бы так:**

ABC---DEFD  
ABCDLFDEF-  
ABCDGFDEFD

# Итеративное улучшение (refinement) выравнивания

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



# 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

# Список популярных программ множественного выравнивания

- ❖ Прогрессивное выравнивание (ClustalW)
- ❖ Прогрессивное выравнивание → итеративное улучшение (Muscle, PRALINE, IterAlign, MAFFT)
- ❖ Методы, учитывающие согласованность попарных сравнений (PROBCONS, Prank, Muscle5)
- ❖ Методы, использующие структурные данные (3D-Coffee)

# А какая программа лучше?

- ❖ Все алгоритмы множественного выравнивания эвристические

# Как сравнить программы?

- ❖ Базы данных идеальных выравниваний!  
Обычно 3D-наложение структур + ручная проверка экспертами
- ❖ BAliBASE (Benchmark Alignment dataBASE) – наиболее признанная
  - ✓ 10 наборов референсных выравниваний, покрывающих разные особенности (высокая вариабельность, повторы и т д)
- ❖ Есть и другие
  - ✓ SABMark(SequenceAlignmentBenchMark)
  - ✓ OXBench(OXfordBenchmark)
  - ✓ SMART(SimpleModularArchitectureResearchTool)
  - ✓ PREFAB(ProteinREFerenceAlignmentBenchmark)
  - ✓ и др.

# Результаты сравнения программ на BaliBase

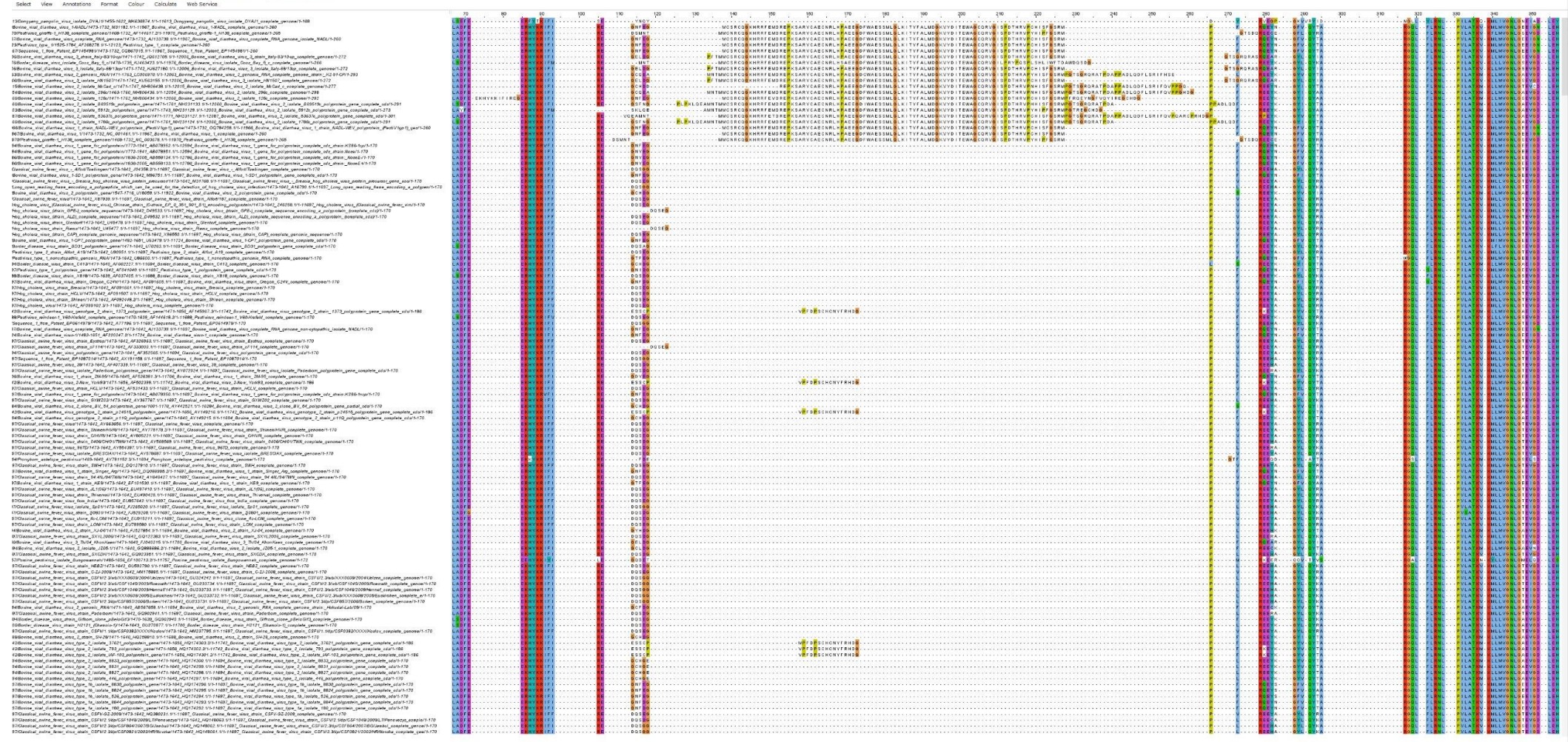
Program	R1-1	R1-2	R2	R3	R4	R5	Avg Score	Tot time(s)
MSAProbs [122]	0.441	<b>0.865</b>	<b>0.464</b>	<b>0.607</b>	<b>0.622</b>	<b>0.608</b>	<b>0.607</b>	12382.00
Probalign [84]	<b>0.453</b>	0.862	0.439	0.566	0.603	0.549	0.589	10095.20
MAFFT [14]	0.439	0.831	0.450	0.581	0.605	0.591	0.588	1475.40
Probcons [17]	0.417	0.855	0.406	0.544	0.532	0.573	0.558	13086.30
Clustal Omega [123]	0.358	0.789	0.450	0.575	0.579	0.533	0.554	539.91
T-Coffee [71]	0.410	0.848	0.402	0.491	0.545	0.587	0.551	81041.50
Kalign [72]	0.365	0.790	0.360	0.476	0.504	0.435	0.501	<b>21.88</b>
MUSCLE [15]	0.318	0.804	0.350	0.409	0.450	0.460	0.475	789.57
FSA [136]	0.270	0.818	0.187	0.259	0.474	0.398	0.419	53648.10
DiAlign [12]	0.265	0.696	0.292	0.312	0.441	0.425	0.415	3977.44
PRANK [137]	0.223	0.680	0.257	0.321	0.360	0.356	0.376	128355.00
ClustalW [11]	0.227	0.712	0.220	0.272	0.396	0.308	0.374	766.47

- Колонки 2-7 – разные наборы тестовых выравниваний.
- Указан TC (total column) вес сравнения с референсным выравниванием.
- Значения от 0 до 1. Чем больше – тем более похоже выравнивание на референсное.
- Последняя колонка – время работы программы



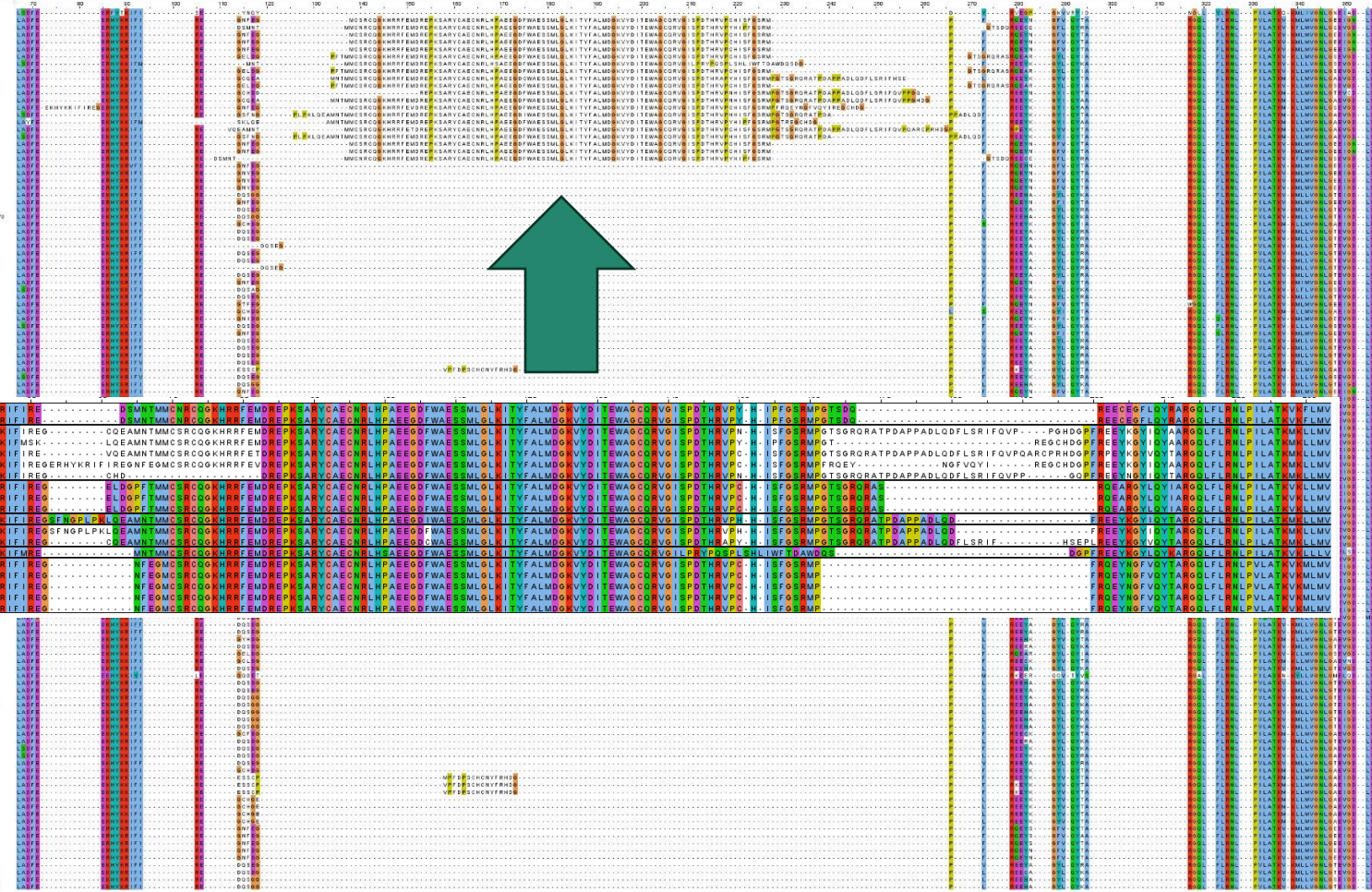
# Выравнивание тех же последовательностей с помощью MUSCLE

logy/Flaviviridae/pesti\_alignment/Type5insertion\_allseq\_muscle.fasta



logy/Flaviviridae/pesti\_alignment/TypeInsertion/alseq\_muscle.fasta

Select View Annotations Format Colour Calculate Web Service



# **Сравнение разных выравниваний одних и тех же последовательностей**

Построенных разными алгоритмами множественного выравнивания

# Какие выравнивания тех же последовательностей совпадают?

1

Seq1	MKFR-SSHYA-S
Seq2	MKYRRS-HYA-S
Seq3	MEFRRRSHYA-R

2

Seq1	MKF-RSSHYAS
Seq2	MKYRRS-HYAS
Seq3	MEFRRRSHYAR

4

Seq1	MKFR-SSHYA-S
Seq2	MKYR-RSHYA-S
Seq3	MEFRRRSHYA-R

3

Seq1	-MKFR-SSHYAS
Seq2	-MKYR-RSHYAS
Seq3	-MEFRRRSHYAR

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

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

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

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

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

```
Seq1 MKFR-SSHYA-S  
Seq2 MKYRRS-HYA-S  
Seq3 MEFRRRSHYA-R
```

```
Seq1 MKF-RSSHYPAS  
Seq2 MKYRRS-HYPAS  
Seq3 MEFRRRSHYPAR
```

Вместо аминокислоты укажем номер остатка в последовательности:

```
Seq1 1234-5678 9 -10  
Seq2 123456-78 9 -10  
Seq3 123456789 10-11
```

```
Seq1 123-45678 9 10  
Seq2 123456-78 9 10  
Seq3 123456789 10 11
```

Seq1 MKFR-SSHYA-S  
Seq2 MKYRRS-HYA-S  
Seq3 MEFRRRSHYA-R

Seq1 MKF-RSSHYPAS  
Seq2 MKYRRS-HYPAS  
Seq3 MEFRRRSHYPAR

Вместо аминокислоты укажем номер остатка в последовательности:

Seq1 123**4**-5678 9 -10  
Seq2 123**45**6-78 9 -10  
Seq3 123**45**6789 10-11

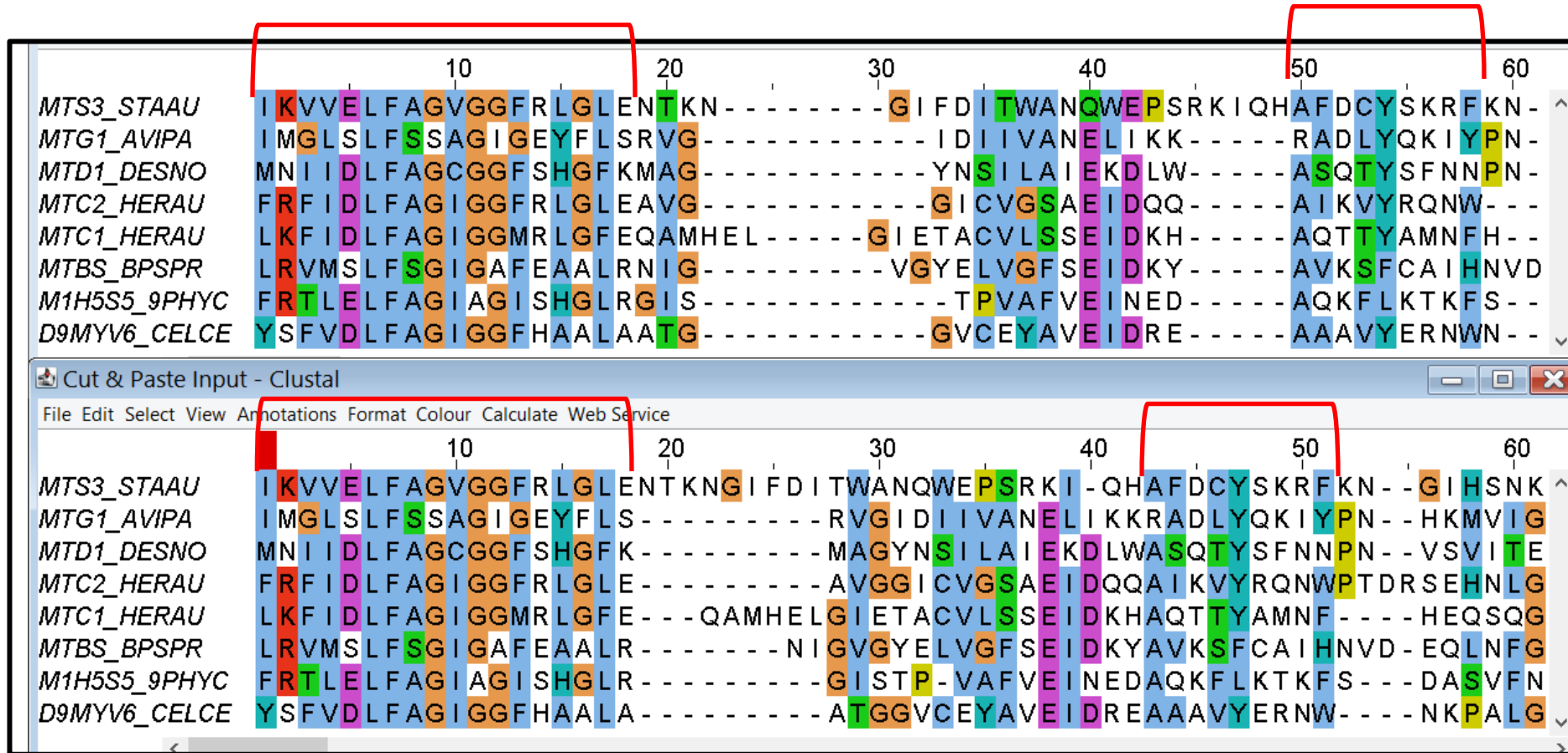
Seq1 123-**4**5678 9 10  
Seq2 123**45**6-78 9 10  
Seq3 123**45**6789 10 11

# Точное сравнение двух выравниваний

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

# Выравнивания совпадают на двух выделенных участках

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

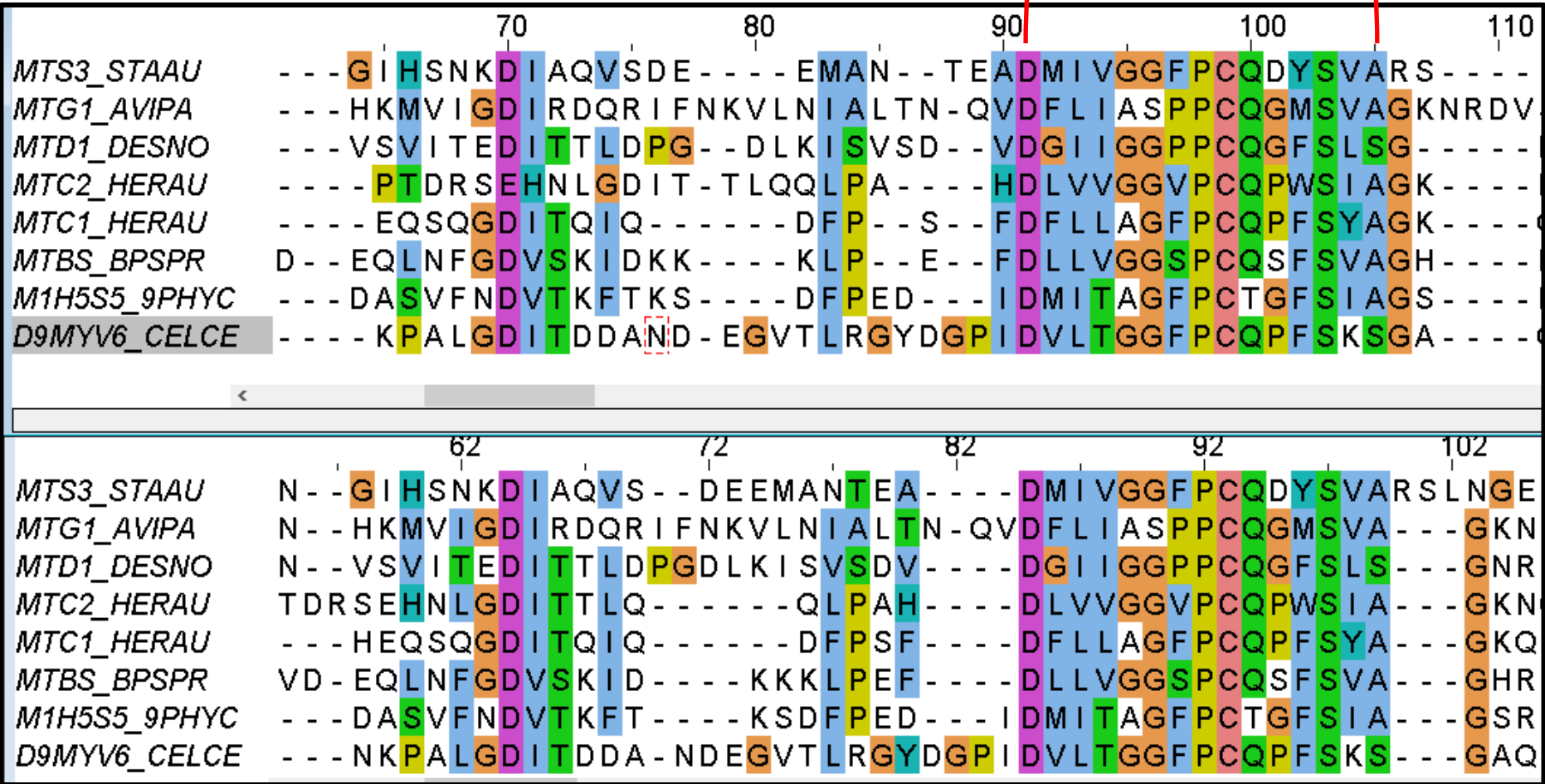




# Продолжение. Разметьте самостоятельно!

			70		80		90		100		110																																						
<i>MTS3_STAAU</i>	- - -	G	I	H	S	N	K	D	I	A	Q	V	S	D	E	- - -	E	M	A	N	-	T	E	A	D	M	I	V	G	G	F	P	C	Q	D	Y	S	V	A	R	S	- - -							
<i>MTG1_AVIPA</i>	- - -	H	K	M	V	I	G	D	I	R	D	Q	R	I	F	N	K	V	L	N	I	A	L	T	N	-	Q	V	D	F	L	I	A	S	P	P	C	Q	G	M	S	V	A	G	K	N	R	D	V
<i>MTD1_DESNO</i>	- - -	V	S	V	I	T	E	D	I	T	T	L	D	P	G	- -	D	L	K	I	S	V	S	D	- -	V	D	G	I	I	G	G	P	P	C	Q	G	F	S	L	S	G	- - -						
<i>MTC2_HERAU</i>	- - -	P	T	D	R	S	E	H	N	L	G	D	I	T	-	T	L	Q	Q	L	P	A	- - -	H	D	L	V	V	G	G	V	P	C	Q	P	W	S	I	A	G	K	- - -							
<i>MTC1_HERAU</i>	- - -	E	Q	S	Q	G	D	I	T	Q	I	Q	- - -	-	D	F	P	- -	S	- -	F	D	F	L	L	A	G	F	P	C	Q	P	F	S	Y	A	G	K	- - -										
<i>MTBS_BPSPR</i>	D - -	E	Q	L	N	F	G	D	V	S	K	I	D	K	- - -	K	L	P	- -	E	- -	F	D	L	L	V	G	G	S	P	C	Q	S	F	S	V	A	G	H	- - -									
<i>M1H5S5_9PHYC</i>	- - -	D	A	S	V	F	N	D	V	T	K	F	T	K	S	- - -	D	F	P	E	D	- -	I	D	M	I	T	A	G	F	P	C	T	G	F	S	I	A	G	S	- - -								
<i>D9MYV6_CELCE</i>	- - -	K	P	A	L	G	D	I	T	D	D	A	N	D	-	E	G	V	T	L	R	G	Y	D	G	P	I	D	V	L	T	G	G	F	P	C	Q	P	F	S	K	S	G	A	- - -				
< <span style="display: inline-block; width: 100px; height: 10px; background-color: #ccc;"></span>																																																	
			62		72		82		92		102																																						
<i>MTS3_STAAU</i>	N - -	G	I	H	S	N	K	D	I	A	Q	V	S	- -	D	E	E	M	A	N	T	E	A	- - -	D	M	I	V	G	G	F	P	C	Q	D	Y	S	V	A	R	S	L	N	G	E				
<i>MTG1_AVIPA</i>	N - -	H	K	M	V	I	G	D	I	R	D	Q	R	I	F	N	K	V	L	N	I	A	L	T	N	-	Q	V	D	F	L	I	A	S	P	P	C	Q	G	M	S	V	A	- - -	G	K	N		
<i>MTD1_DESNO</i>	N - -	V	S	V	I	T	E	D	I	T	T	L	D	P	G	D	L	K	I	S	V	S	D	V	- - -	D	G	I	I	G	G	P	P	C	Q	G	F	S	L	S	- - -	G	N	R					
<i>MTC2_HERAU</i>	T	D	R	S	E	H	N	L	G	D	I	T	T	L	Q	- - - - -	Q	L	P	A	H	- - -	D	L	V	V	G	G	V	P	C	Q	P	W	S	I	A	- - -	G	K	N								
<i>MTC1_HERAU</i>	- - -	H	E	Q	S	Q	G	D	I	T	Q	I	Q	- - - - -	D	F	P	S	F	- - -	D	F	L	L	A	G	F	P	C	Q	P	F	S	Y	A	- - -	G	K	Q										
<i>MTBS_BPSPR</i>	V	D	-	E	Q	L	N	F	G	D	V	S	K	I	D	- - -	K	K	K	L	P	E	F	- - -	D	L	L	V	G	G	S	P	C	Q	S	F	S	V	A	- - -	G	H	R						
<i>M1H5S5_9PHYC</i>	- - -	D	A	S	V	F	N	D	V	T	K	F	T	- - -	K	S	D	F	P	E	D	- -	I	D	M	I	T	A	G	F	P	C	T	G	F	S	I	A	- - -	G	S	R							
<i>D9MYV6_CELCE</i>	- - -	N	K	P	A	L	G	D	I	T	D	D	A	-	N	D	E	G	V	T	L	R	G	Y	D	G	P	I	D	V	L	T	G	G	F	P	C	Q	P	F	S	K	S	- - -	G	A	Q		

Если в двух блоках без гэпов из разных выравниваний есть хотя бы одна одинаково выровненная позиция, то блоки выровнены одинаково



# Как сравнить два множественных выравнивания

1. Отсортировать последовательности по ID в обоих выравниваниях.
2. Если колонка (номеров) букв в выравнивании A отличается хотя бы одним номером от колонки букв в выравнивании B, то они выровнены не одинаково

## 3. Алгоритм сравнения

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

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

$s_1$  – номер буквы в первой последовательности

$s_2$  – номер буквы во второй последовательности

.....

Если в последовательности  $i$  стоит гэп, то  $s_i = \text{"-"}$

2) Если  $SI(N) = SII(N')$ , то колонка  $N$  выравнивания I выровнена одинаково с колонкой  $N'$  выравнивания II

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

# Готовые программы для сравнения выравниваний

- ❖ VerAlign: a multiple sequence alignment assessment tool  
(<https://www.ibi.vu.nl/programs/veralignwww/>)
- ❖ AlignStat: A tool for the statistical comparison of alternative multiple sequence alignments  
(<https://github.com/TS404/AlignStat?tab=readme-ov-file>)

# VerAlign – программа для ручного сравнения – раскрашивает колонки второго выравнивание цветами первого

```

GVLAGLUNDPS-----CKYAYGQNRK----TKFLEKSLSEVDGKELNALYFNNQ--HKILLYSCAFCCQDESSQYIK
EIVAAVDNWRP-----AINTYQGNF-----THPIHELDLAQIDAAVSLIKTHS--PELILGGFFCQDESSAG--
NSILALEKDLW-----ASQTYGFNNPN--VSVITEDITTLDPG--DLKISVSD--VDGILGGFFCQGFSSLSG--
ICVGSAAEIDQQ-----AIKVYQGNW-----PTDRSEHNLGDIIT-TLQQLFA---HDLVYGGVFCQPFWSIAGK
ACVLSSEIDKH--AQT--TYAMNFH-----EQSQGDITQIQ-----DFP--S--FDFELAGEFPCQPFSSYAGK
ELVGFSEIDKY-----AVKSFCAIHNVN--EQQLNFGDVSVIDKDK--KLP--E--FDLLVGGSPCQSFSSVAGH
KCVFSSSEWDKY--AAQ--TYFANYG-----EKPHGDITKINEN--DIP--D--QDVLAGEFPCQPFSSNIGK
ECVLSSEIDKK--ACE--TYALNFK-----EEPQGDITHEIT-----SFP--E--FDFELAGEFPCQPFSSYAGK
ETVWANEYDKN-----AAITYQSNFKN--KLIIDDIRNIKVE--DVP--D--FDVLVSGFPCFSSVAGY
VCVASAAEIDQQ-----AIKVYQGNW-----PTDGVVDHNLGDIIT-AIQQLFA---HDVLVGGVFCQPFWSIAGK
TPVAVVEINED-----AQKFLTKKFS--DASVFNDVTKFTKS--DFPED--IDMITAGEFPCFSSIAGS
ECVYSNEWDKY--AQE--VYEMNFG-----EKPEGDITQVNEK--TIP--D--HDILCAGFPCQAFSSISGK
VCEYAVEIDRE-----AAAVYRNWN--KFPALGDITDDAND--EGVTLRGYDGPIDVLVGGFPCQPFSSKSGA
NVVFSSEWDKF--AQK--TYFANYG-----DFPDGDITKIDEK--DIP--D--HEILVGGFPCVAFSSQAGL
EHAWANDIDEW-----ACETFSTNI--CPDRPDSVVC--GDVRELDIKSLGEEKFG--EIDAFVGGFPCNDYSSIVGE

```

```

FEI--PAANEYD---K-----TIWATFRANHPKT--HLIEGDIRKI-----KE-----E-D---F---
PVSNGYWKRRKKD---D-----ELKIYNAIKLSEK--EGNIFDIRDL-----YK-----R-T---L---
-----FENRADKLGQ---K-----KLKDMYANKLNK--NFGDIRSI-----DP-----K-K---L---
VKS--VFSSEID---K-----FAIKTYANFGD--E-PHGDITKI-----DE-----K-D---I---
LST--YGAVEID---K-----NAAETLRINRPKW--KVIENDIEFI-----AD-----NLDEFI---D---
FDI--TWANQWE---PSRRIQHAFDCYSKREKNGI--H-SNKDIAQV-----SD-----E-E---M---
VRC--VFSSEID---K-----YAVQTYGANHGG--T-VCGDITQT-----DV-----A-D---I---
FSH--VALIEIE---P-----SACQTLRINRPDW--NVIEGDVRIEF-----QG-----E-G---Y---
GKC--VFSSEID---P-----FAKFTYTFNEGV--V-PFGDITKV-----EA-----T-T---I---
FNI--VFANDNW---K-----GCWKTFRKNGHI--KINKKPIEWL-----KP-----S-E---I---
FEI--CAAFENW---E-----KAIEIYNNFESH--PIYNIDLRE-----KE-----AV-E-K---IKK---
FRI--ICANEYD---K-----SIWKTYSNHSA--KLIKGDISKI-----SS-----D-E---F---

```

# Практическое задание в классе

- ❖ Дано: два выравнивания pf00145\_seed-reduced.fasta и pf00145\_seed-tcoffee-reduced.fasta одних и тех же последовательностей
- ❖ Найти
  - ✓ блок одинаково выровненных колонок в двух выравниваниях
  - ✓ участок, на котором выравнивания полностью различны
  - ✓ результаты сохранить в гугл-форме

# Участок несовпадающих выравниваний – между двумя совпадающими блоками

