

Signals 3

1. Консервативные мотивы в выравнивании

Будем рассматривать архитектуру PF12171. Это семейство цинковых пальцев, связывающих двухцепочную РНК. Seed- выравнивание данной архитектуры содержит 108 последовательностей и выглядит следующим образом (раскраска Clustal):

```
G I W C Q Y C Q - K N Y S K Q T V Y D A H L N S - - K G H K
A F Y C E V C Q - K F F G K I T V F E A H K K S - - K A H N
S F Y C Q L C Q - K G Y S R M N D Y E A H L S S Y D H S H K
E V Y C E A C D - K L F A K D S V Y K G H L S G - - K K H Q
E V Y C S V C Q - K V F A K I S V Y Q G H L N G - - K K H K
G F W C S A C K - K D F A K E S V Y T A H L T G - - K K H K
P L Y C K S C D - K H F T K E S V F T A H L T G - - K K H K
S L Y C E A C A - K K F S N K A V Y D G H L L G - - K K H K
E L F C K F C D - K Q F T N Q A V Y D N H L P G - - K K H K
G I Y C P F C S - R W F K T S S V F E S H L V G - - K I H K
N D Y C K L C D - A S F S P A V A Q A H Y Q G - - K N H A
D R Y C G L C A - A W F N N P L M A Q Q H Y D G - - K K H K
Q L M C A L C D - A P V K N A L L W O T H V L G - - K Q H K
K L V C V L C R - E H I K T E A L W D G H I R N - - T A H R
H L S C V L C N - V Q V K S E L L W P T H V L G - - K P H K
I Y Y C L Y C D - K M F N S A H I Y D S H R T G - - K A H S
E W F C D Y C G - R I V D N A N V F K T H C A T - - K K H K
Q F N C L L C H - R N F S N A S V M D Y H F K T - - K K H K
V H I C L T C N - K E F K E A Q F L H R H Y S T - - K Y H R
Q F Y C L H C D - R Y F S N V S V R D D H F K T - - K K H K
Q F T C L N C D - A R F A T A E V Q R E H Y K T - - D W H R
C L R C T L C D - L A F K D S L T L V O H L S S - - E A H R
G Y Y C K I C D - C T L K D S Q T Y L D H I N G - - K N H N
G F Y C D V C D - V S F K D S N S M L A H V N T - - K R H N
S F Y C P V C D - I Q F S D S Q A A E A H K A S - - R Q H K
A F Y C P S C N - V Y C S D S R T A A L H R S S - - L K H K
G F F C P I C S - L F Y S G E K A M T N H C K S - - T R H K
Q Y Y C Q D C D - I S M N S I K Q M E Q H M T S - - A R H R
Q Y Y C Q P C N - M M M N H E S T L Q Q H F I G - - K K H L
Q F Y C S M C N - V G A G E E M E F R Q H L E S - - K Q H K
P F Y C D L C E - Y Q T N T R Y S F L R H K K S - - K K H Q
N F Y C H L C G - R H S N G D R Q W E Q H I S S - - E R H K
K F E C R I C D - K E F N S D G T Y Q Q H L T S - - K K H K
L F Y C V E C D - R N F P S Q K D Q L T H I A S - - K L H K
H F Y C C E C D - R H F I T E K V L M E H K R S - - N P H R
M H F C I I C D - R Y F I S A E A L E T H K A T - - G K H K
Q I I C K I C D - R M F I S E E A L R K H E K E - S K L H Q
Q F F C I F C D - K Y F I D Q I T L D L H K K Q - - K P H K
V S Y C V H C K G K T F L N Q E S V D K H L S S - - K N H L
K F K C I L C P K K I I N E S D L D K H I K S - - K Q H L
Q N Y C I H C S - K H F V T N E D L Q S H I K G - - K P H K
Q F Y C L Q C A - R Y F I D D K S L K D H I K S - - R V H K
L H R C L A C A - R Y F I D S T N L K T H F R S - - K D H K
Q Y Y C I P C A - R Y F N N D E S L K T H Q T T - - K V H K
Q Y Y C V S C A - R Y F V N E E S I K K H Q V S - - K Q H K
```

Выберем из этого выравнивания мотив [YFAWLV]..H..[GTS]K.HK - сразу описываю его в Jalview-формате. Его концевая часть - K.HK - известна как дилизиновый мотив и в случае нахождения на С-конце обычно исполняет сигнальную функцию для Однако обсуждаемая архитектура не у всех белков seed-выравнивания располагается на С-конце.

[Поиск](#) по выравниванию этого паттерна даёт 35 находок, и все они находятся только в той колонке, откуда паттерн был выделен исходно. Поэтому паттерн действительно специфичный.

Запишем паттерн в формате PROSITE: [YFAWLV]-x-x-H-x-x-[GTS]-K-x-HK. Поиск по с такой маской по Swiss-Prot даёт 50 находок, большинство из которых являются ДНК-связывающими белками, однако попадают белки, связывающиеся с тРНК и даже несколько протеасомных белков.

2. Мотив, специфичный для клады

С помощью Neighbour Joining я построил в Jalview дерево для seed-выравнивания, откуда далее выделил кладу из 8 последовательностей. Выравнивание, которое ей соответствует, выглядит следующим образом:

```

A8BB65_GIAIC/58-84  M H F C I I C D - R Y F I S A E L E T H K A T - - G K H K
G4VJB8_SCHMA/51-77  Q F F C I F C D - K Y F I D Q I T L D L H K K Q - - K P H K
Q4N239_THEPA/59-86  V S Y C V H C K G K T F L N Q S V D K H L S - - K N H L
Q5CUH6_CRYPI/74-101 K F K C I L C P K K I I I N E S D L D K H I K S - - K Q H L
ZN593_DICDI/59-85   Q N Y C I H C S - K H F V T N E D L Q S H I K G - - K P H K
A7SMU1_NEMVE/58-84 Q F Y C L Q C A - R Y F I D D K S L K D H I K S - - R V H K
ZN593_HUMAN/60-86  L H R C L A C A - R Y F I D S T N L K T H F R S - - K D H K
Q23MN1_TETTS/59-85 Q Y Y C I P C A - R Y F N N D S L K T H Q T - - K V H K

```

Можно выделить следующий паттерн: C[VLI][IFHA]C[DKPSA][KR][YTHI][FI][VLI] - - по нему среди всего сид-выравнивания находятся 5 последовательностей и только из этой кледы - значит, паттерн для данной кледы специфичен.

3. PSI-BLAST

Будем смотреть на белок с AC Q7VDL2 .

В качестве порога на e-value берём умолчательный - 5e-3 .

И вот что получается:

Итерация	# выше порога	АС худшей из выше	<-eval	АС лучшей из ниже	<-eval
1	141	Q9AG20	0.005	A8GFG7	0.005
2	185	B6JKX0	7e-8	A7H8E6	0.014
3	185	Q9ZM51	2e-12	A7H8E6	0.013

Видим, что наши результаты стабилизировались на третьей итерации.

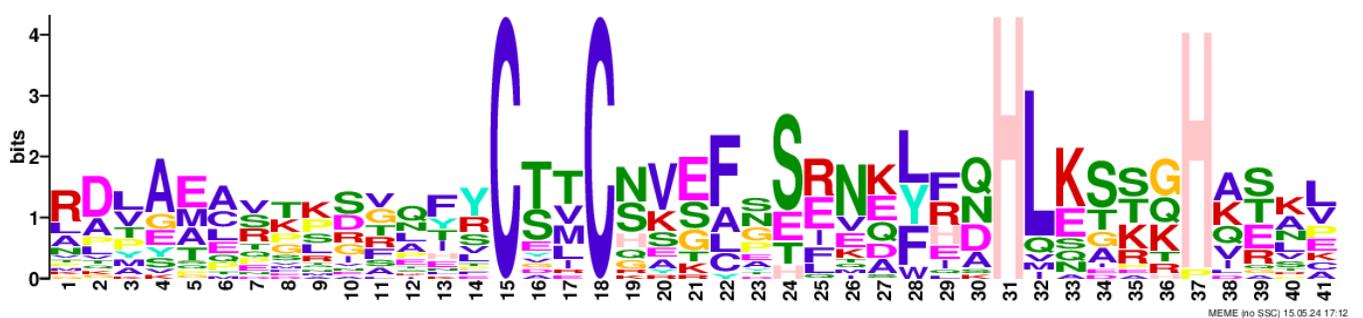
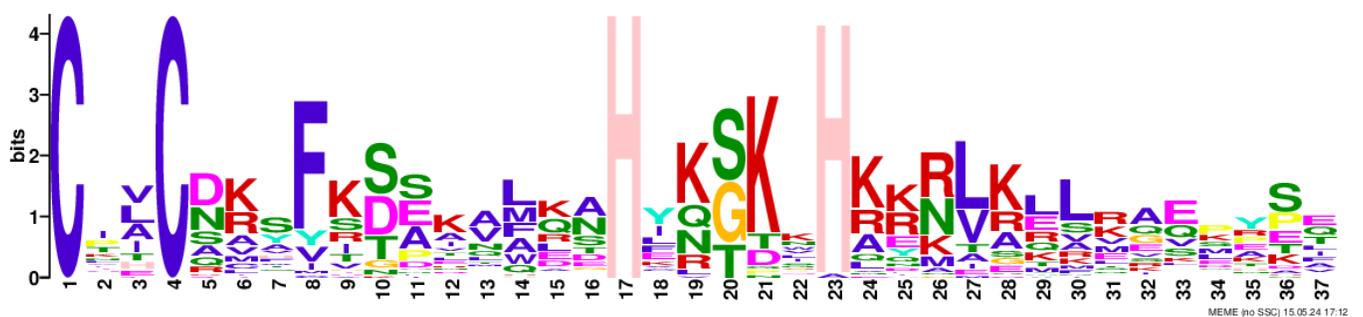
4. MEME

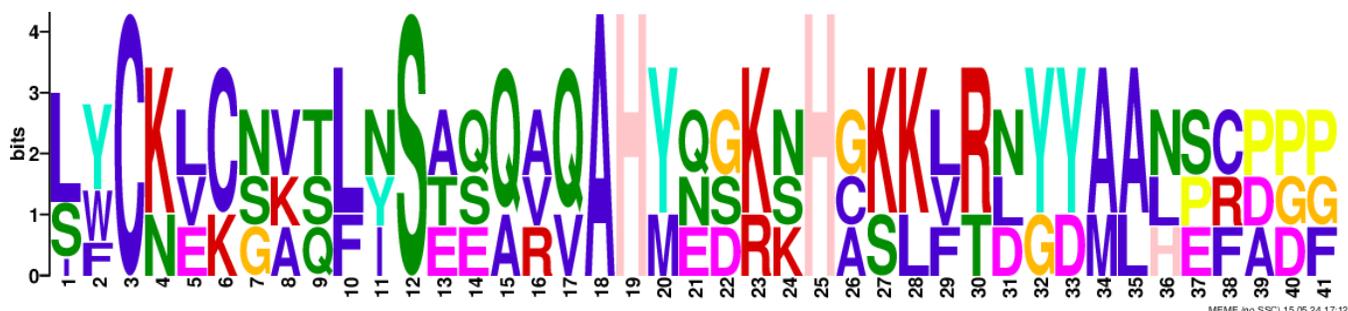
Получив с InterPro только те последовательности (их получилось 45) из нашего seed-выравнивания из заданий 1-2, что относятся к Swiss-Prot, запускаем MEME для поиска мотивов:

```
meme train.fa -dna -nmotifs 3 -minw 6 -maxw 50
```

Полную выдачу можно найти [здесь](#).

Получаем 3 лого:





Видим, что все три лого почти целиком покрывают seed-выравнивание и потомк содержат конструкции, на которые мы уже обращали внимание выше - например (в Jalview-нотации), C...C или KKNK. Приведём некоторые описательные характеристики этих лого, а потом завершим задание общим выводом.

Лого	e-value	# сайтов	длина
1	1.1e-433	44	37
2	7.2e-166	22	41
3	6.8e-152	10	41

Теперь, вместе с данными о e-value, можно делать вывод, что самым адекватным среди рассмотренных является первый паттерн, поскольку он встречается во всех последовательностях (это вывод не из таблицы, а из картинки про распределение паттернов по последовательностям, которая есть в html-отчёте MEME в папке, на которую дана ссылка выше) и обладает примерно в 10^{300} раз меньшим p-value, чем остальные паттерны.

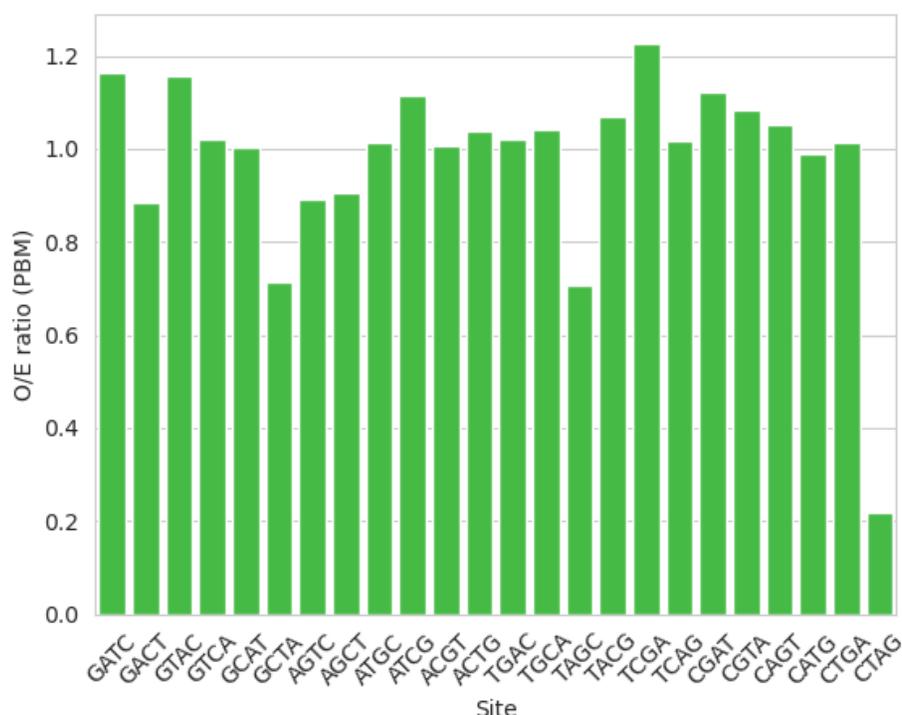
5. Представленность GATC

Обратимся к старому знакомому: *Rhodococcus fascians* D188!

С помощью программы `cbcalc` рассчитаем отношения реальных частот перестановок строки GATC к ожидаемым:

```
cbcalc -s sites.txt R_fasc.fna -o out.tsv
```

Посмотрим теперь на барплот для каждого из сайтов:



По всей видимости, среди рассмотренных четырёхмеров CTAG встречается реже всех. Возможно, он несёт у *R. fascians* сигнальную функцию. Также вероятно, что

частота СТАГ снижена в силу того, что он содержит кодон TAG (а у *R. fascians* это стоп-кодон).