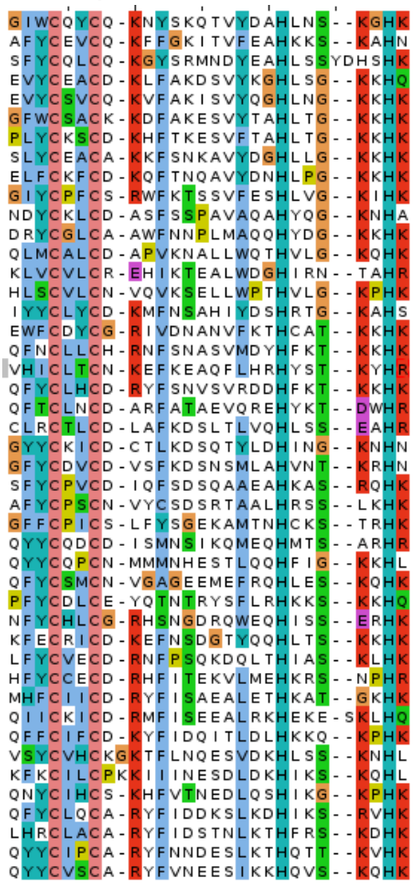
Signals 3

1. Консервативные мотивы в выравнивании

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



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 попадаются белки, связывающиеся с тРНК и даже несколько протеасомных белков.

2. Мотив, специфичный для клады

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

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 клады - значит, паттерн для данной клады 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	AC худшей из выше	<-eval	AC 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`

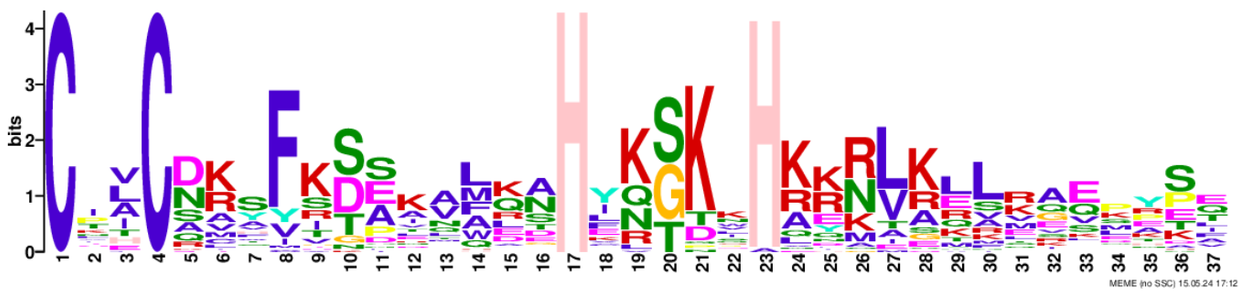
Видим, что наши результаты стабилизировались на третьей итерации.

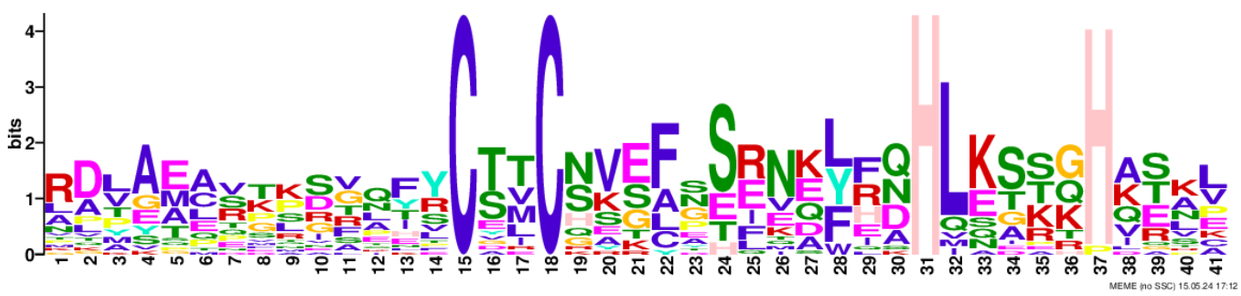
4. MEME

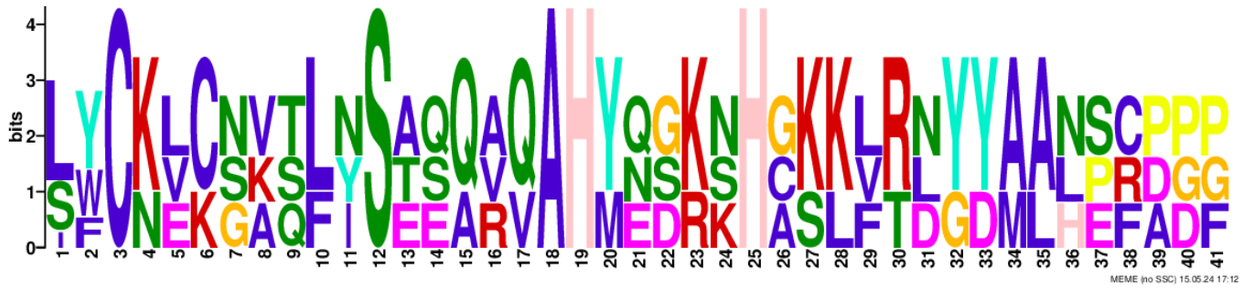
Получив с InterPro только те последовательности (их получилось 45) из нашего seed-выравнивания из заданий 1-2, что относятся к Swiss-Prot, запускаем MEME для поиска мотивов:

```
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 накрывают seed-выравнивание и потомк содержат конструкции, на которые мы уже обращали внимание выше - например (в Jalview-нотации), `C..C` или `KKHK`. 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

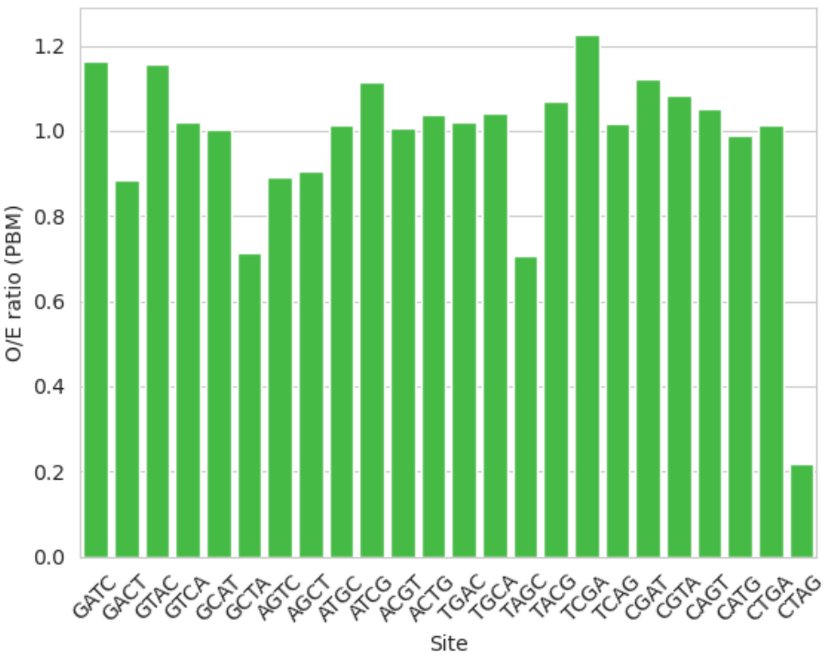
Теперь, вместе с данными о e-value, можно делать вывод, что самым адекватным среди рассмотренных является первый паттерн, поскольку он встречается во всех последовательностях (это вывод не из таблицы, а из картинки про распределение паттернов по последовательностям, которая есть в html-отчёте MEME в 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 CTAG снижена в силу того, что он содержит кодон TAG (а у *R. fascians* это стоп-кодон).