Meme-Fimo (Signals-2)

Выбор и предобработка генома бактерии

Я выбрал геном E. Coli, штамм DSM-30083. Fasta-файл с её геномом был отправлен в недры веб-сервиса Operon mapper, выявившего в геноме 2454 оперона.
Опероны, найденные маппером, были профильтрованы по колонке `function`, далее из них были извлечены последовательности их промоторов. Это было сделано с помощью небольшой модификации скрипта, созданного Георгием Муравьёвым.
Ключевые слова, по которым производилась фильтрация:

```
key_words = [polymerase, gyrase, ligase, ATP_synthase]
```

Соответственно, ключевые слова в скрипте Георгия были заменены на представленные.
Таким образом, было получено 3 файла: `train.fa`, `test.fa`, `negative.fa` - файлы для обучения, тестирования и негативного контроля соответственно. Посмотреть их содержание можно, обратившись к директории.

Запуск MEME

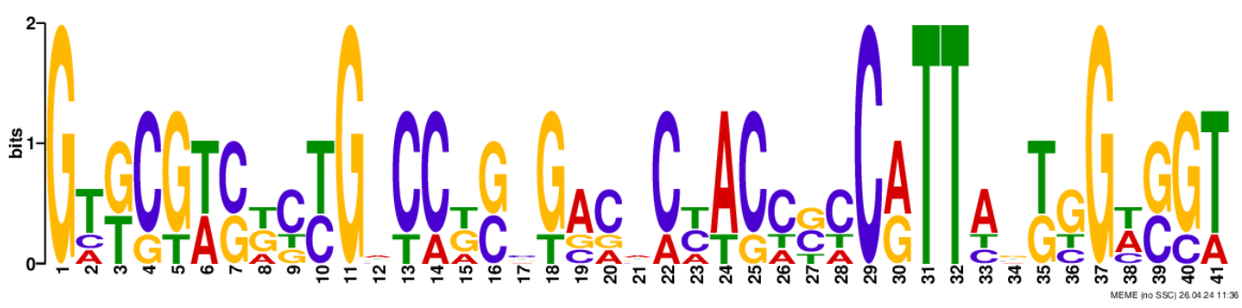
```
meme train.fa -dna -nmotifs 3 -minw 6 -maxw 50
```

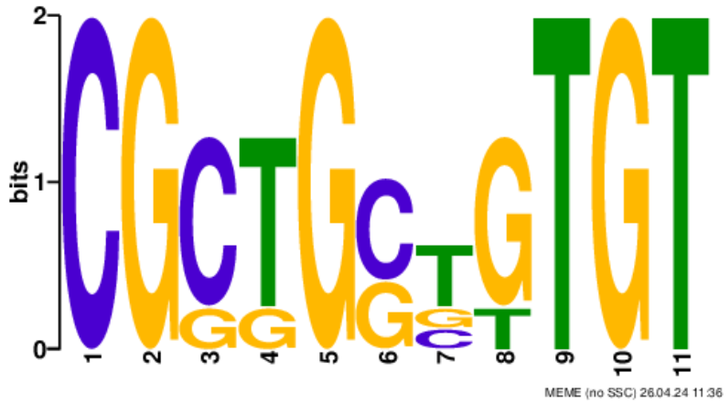
Параметр	Значение	Значение
`-nmotifs`	3	Сколько мотивов найти
`-minw`	6	Минимальная ширина мотива
`-maxw`	50	Максимальная ширина мотива
`-dna`		ДНКовый алфавит

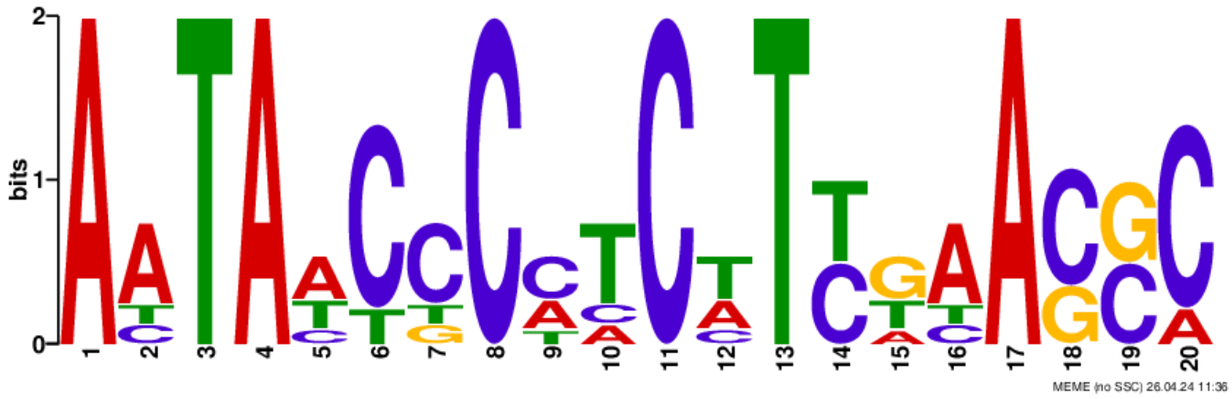
Остальные параметры были выбраны по умолчанию.
Полученные находки:

Мотив	Количество сайтов с ним	E-value
GTKCGWSDCYGNCCDSNGACNCHACCBCCRTTANKGGHSGT	5	$6.7 * 10^{-1}$
CGCTGSTGTGT	5	$1.5 * 10^{-3}$
AATAWCCCMTCWTYKAASSC	6	$1.1 * 10^{-3}$

А вот и лого:







Наименьшим e-value обладает третий мотив. Выпишем для него регулярное выражение в Jalview-нотации:
A[ATC]TA[ATC][CT][CTG]C[CAT]{2}C[CAT]T[CT][GTA][CAT]A[CG]{2}[CA]

Выдачу `MEME` я поместил в директорию `meme_out`; приятного просмотра!

Запуск FIMO

Поиск мотива в положительном и отрицательном контролях был выполнен с помощью команд:

```
fimo --norc -motif AATAWCCCMTCWTYKAASSC -thresh 0.001 ./meme_out/meme.txt
test.fa

fimo --norc -motif AATAWCCCMTCWTYKAASSC -thresh 0.001 ./meme_out/meme.txt
negative.fa
```

Параметр	Значение	Значение
`--nnorc`		Только по одной цепи
`-treshold`	6	Порог на e-value

Чтобы нам было удобно анализировать результаты запусков с разными значениями порога, посмотрим, сколько всего последовательностей в файлах, на которых мы будем запускать fimo:

`test.fa`	`negative.fa`
9606	118

Теперь, опять-таки, таблично, опишем, как количество находок для каждого из файлов зависит от порога на e-value:

e-value threshold	`test.fa`	`negative.fa`
0.1	40341	481

e-value threshold	`test.fa`	`negative.fa`
0.0225	9397	100
0.01	4253	45
0.007	3090	30
0.005	2219	22
0.001	456	9
0.0001	64	6

Из таблицы нам видно, что наиболее оптимальным порогом на e-value из рассмотренных является 0.001, поскольку при нём возникает адекватное (когда мы нашли что-то промоторное в том, что промотором не является), а находок среди последовательностей, которые были отобраны как промоторы, достаточно. Почему достаточно? - спросите вы, - у нас ведь в положительном контроле более 9000 промоторных последовательностей! Тут нужно обратить внимание на результаты запуска `MEME`: в мотив, который мы искали при запуске `FIMO`, вносят вклад всего 6 последовательностей из 118 в `train.fa` (5%). Поэтому, если процедура получения `train.fa` и `test.fa` идентична, можно ожидать, что рассмотренному мотиву среди 9606 последовательностей теста тоже около 5%. При пороге на evalue мы это и видим: получается 456 находок, чуть меньше ожидаемых 480.

Находки `FIMO` при пороге 0.001 лежат в директориях: `fimo_test_out` и `fimo_negative_out`.