Описание мотива в белках паттерном

Выбрали белки с мнемоникой PGK, которая отвечает фосфоглицераткиназе. Это ключевой фермент гликолиза (класс трансфераз), катализирующий обратимый перенос фосфатной группы от 1,3-бисфосфоглицерата на АДФ, образуя АТФ и 3-фосфоглицерат.

Выбрали и скачали в Swiss-Prot 10 последовательностей белка у разных бактерий: PGK_ECOLI, PGKT_THEMA , PGK_GEOSE, PGK_BACSU, PGK_SYNFM, PGK_RUTMC, PGK_CHLPD, PGK_ACIBT, PGK_FRATW, PGK_MYCGI

В программе Jalview сделали выравнивание алгоритмом Muscle:
выравнивание

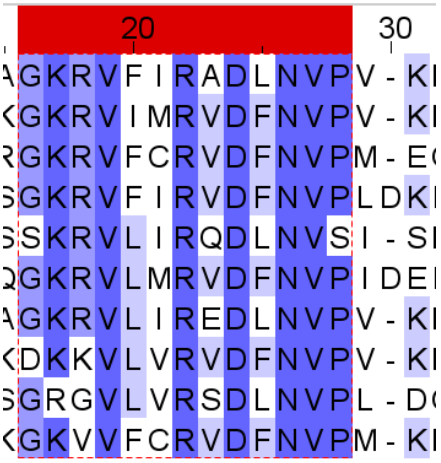

Рис. 1 Позиции в выравнивании, рассматриваемые в качестве мотива: 16-28 (консервативный участок без гэпов длиной 13 позиций)

Состсавим паттерн:
[GSD]-[KR]-[RKGV]-V-[LFI]-x-R-x-D-[FL]-N-V-[PS]

Посчитаем кол ичество находок с такой мнемоникой в файле /P/y24/term4/bacteria-sw.fasta:

```
grep "^>.*PGK_" /P/y24/term4/bacteria-sw.fasta | wc -l
```

612 находок

В UniProtKB по запросу (taxonomy_id:2) AND (id:PGK_*) получаем 611 находок.

Запустим fuzzpro

```
fuzzpro -sequence /P/y24/term4/bacteria-sw.fasta -pattern
"[GSD]-[KR]-[RKGV]-V-[LFI]-x-R-x-D-[FL]-N-V-[PS] " -outfile
motifs_rpa.txt
```

С помощью grep посмотрим количество находок:

```
grep -c 'HitCount' motifs_rpa.txt
```
430

Посмотрим количество правильных находок:
```
grep -c '# Sequence: PGK_' motifs_rpa.txt
```
427

То есть у нас 3 ложноположительные находки и 184 ложноотрицательные находки, что связано со строгостью паттерна.

Ослабим паттерн:
[GSD]-[KR]-x-V-x-x-R-x-D-[FL]-N-V-[PS]
473 находки, из которых 470 с нужной мнемоникой. Дальше ослаблять нет смысла, тк паттерн станет нерепрезентативным.

Поиск мотивов в белках программой MEME и поиск этих мотивов в банке

Возьмем те же белки, которые вы использовали для создания паттерна. Найдем в них мотивы программой MEME с опциями: последовательности аминокислотные, по одному представителю мотива на последовательность, минимальная длина 8, максимальная длина 15, до трёх мотивов.

```
meme pgk.fasta -protein -mod oops -nmotifs 3 -minw 8 -maxw 15 -oc meme_out
```

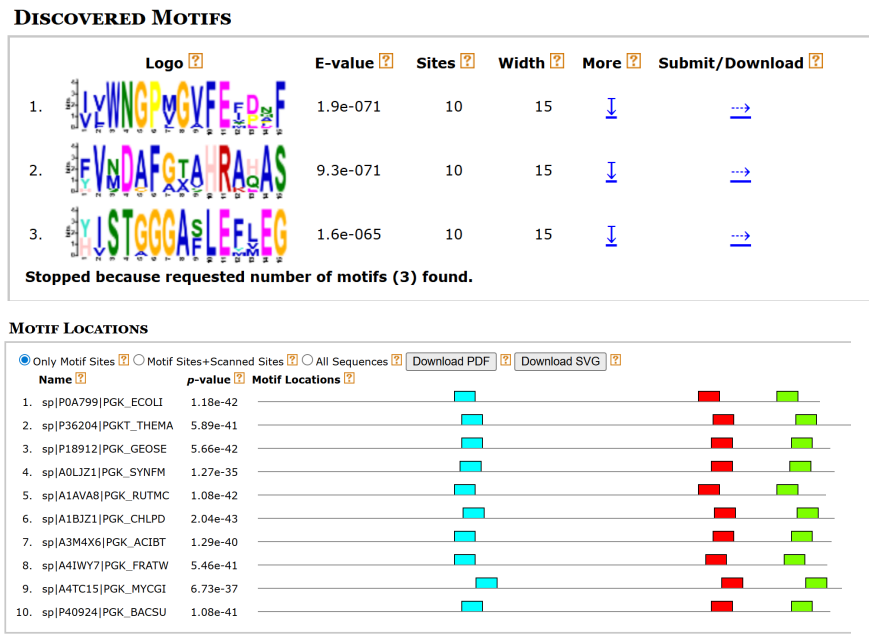

Рис 2. Выдача MEME

Дальше запустим MAST

```
mast meme_out/meme.txt /P/y24/term4/bacteria-sw.fasta -oc mast_out
```

Рис3. Выдача MAST.

У нас получилось так, что определенный мной мотив совершенно не совпал с програмными :)

На странице с выдачей MAST указано, что 611 последовательнонстей имеет E-value < 10. А также что показанные совпадения мотивов имеют p-value для позиции меньше 0,0001.

Поиск последовательности Шайна — Дальгарно в геноме своего прокариота

Последовательность Шайна — Дальгарно (SD-последовательность) — это участок на мРНК прокариот (обычно AGGAGG), расположенный за 6–10 нуклеотидов до старт-кодона AUG. Она обеспечивает инициацию трансляции, связываясь с комплементарным 3'-концом 16S рРНК, что позволяет рибосоме правильно определить место начала синтеза белка.

В первом семестре работали с *Natribaculum luteum*, будем работать с его геномом. Запустим программу fuzznuc с параметром -complement Y (поиск как на прямой, так и на обратной цепи).

```
fuzznuc ~/term1/genome/GCF_023008545.1_ASM2300854v1_genomic.fna
-pattern 'A-G-G-A-G-G' -complement Y -outfile SD.fuzznuc
```

Было найдено 3358 хитов, из которых 2011 на прмой цепи, 1347 на оюбратной.
`grep -c '+ pattern' SD.fuzznuc`

Найдем число находок, ожидаемое по случайным причинам. Для это при помощи программы infoseq был определен GC-состав генома:

```
infoseq ~/term1/genome/GCF_023008545.1_ASM2300854v1_genomic.fna
-only -pgc
```

(-pgc гц состав, -only выводит только таблицу)
64.19%

Отсюда частоты нуклеотидов: p(G)=p(C)=0.32, p(A)=p(T)=0.18. Тогда вероятность появления паттерна AGGAGG:

P(AGGAGG) = p(A)*p(G)*p(G)*p(A)*p(G)*p(G) = 3.4*10^-4

Тогда ожидаемое число находок = L*P*2, где L- длина генома в нуклеотидах (3656043), а 2 - учет прямой и обратной цепей. Получаем 2484.

```
infoseq -sequence
~/term1/genome/GCF_023008545.1_ASM2300854v1_genomic.fna -only
-length
```

При больших n и маленьких p биномиальное распределение становится Пуассоновским. При λ = 2484 и наблюдаемом числе сайтов k = 3358 z критерий составляет :

z ≈ (3358 - 2484) / 49.84 ≈ 17.53

Поскольку z >> 1.96 (при стандартном пороге alpha = 0.05), мы должны отвергнуть нулевую гипотезу о том, что наблюдаемое число сайтов последовательности Шайна-Дальгарно соответствует случайному ожиданию. Наблюдаемое число сайтов значимо превышает то, которое стоило бы ожидать при случайном распределении нуклеотидов, что говорит о том, что это сигнал, а не случайный набор букв.

Просмотрев 15 случайных находок и сравнив их с координатами кодирующих последовательностей из файла с геномной таблицей, поняла, что ни одна из них по сути не является настоящим сайтом Шайно-Дальгарно, поскольку они не лежат за 20 нуклеотидов от кодирующих участков.