Практикум 8.

1.

Выбранная мнемоника – RL1 (Large ribosomal subunit protein uL1). Это белок, участвующий в образовании структуры рибосомы: он связывает 23S-рРНК очень близко к ее 3’-концу.
В файле bacteria-sw.fasta 787 белков с такой мнемоникой (grep '|RL1_’ bacteria-sw.fasta | sort -u | wc -l).
Идентификаторы выбранных белков:
RL1_ECOLI
RL1_HELPY
RL1_THET8
RL1_BACSU
RL1_GEOSE
RL1_STRGR
RL1_POLAQ
RL1_PSEAE
Выравнивание проводилось командой mafft RL1.fasta > RL1_aln.fasta.

Ссылка на файл с выравниванием.

Участок, выбранный для профиля – позиции выравнивания с номерами 129-149. (Рис.1.)
Паттерн –
[LVI]-G-[QR]-[VI]-L-G-P-[KR]-G-L-[ML]-P-[ND]-P-K-[VTA]-G-T-V-[TG]-[PMF] .

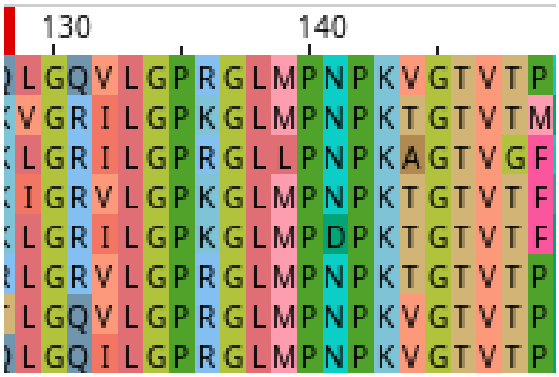

Рис.1. Изображение выбранного для профиля участка выравнивания.
Поиск белков по профилю производился командой:

```
fuzzpro -sequence /P/y24/term4/bacteria-sw.fasta -pattern
[LVI]-G-[QR]-[VI]-L-G-P-[KR]-G-L-[ML]-P-[ND]-P-K-[VTA]-G-T-V-[TG
]-[PMF] -outfile report
```

Всего найдено последовательностей – 306. Из них с верной мнемоникой – тоже 306: grep 'RL1_' report | sort -u | wc -l . Ложноположительных результатов нет. Ложноотрицательных результатов 787 - 306 = 581.
Попробуем ослабить регулярку: x-G-x-[VI]-L-G-P-[KR]-G-L-[ML]-P-x

С ней найдено 513 последовательностей. Из них с верной мнемоникой – тоже 513 Ложноотрицательных результатов – 787 - 513 = 375. Уже лучше. Но всё равно большая частота отвержений. Вывод: паттерны далеко не самый мощный способ искать нужные белки.

2.

Запустим программу meme на выбранных в самом начале последовательностях:

```
meme RL1.fasta -protein -mod oops -nmotifs 3 -minw 8 -maxw 15
```

Выдача meme в html.

В результате было найдено 3 мотива, присутствующих во всех выбранных белках.

Далее был выполнен поиск белков из SwissProt, содержащих найденные мотивы, при помощи программы mast:

```
mast ./meme_out/meme.html /P/y24/term4/bacteria-sw.fasta
```

Выдача mast в html.

Всего найдено последовательностей – 795. Из них с верной мнемоникой – 786. Ложноположительных результатов – 9. Ложноотрицательных результатов 787 - 786 = 1.

Вывод: хоть появилось 9 ложноположительных результатов, ложноотрицательных практически нет: 375 против 1. MEME + MAST является гораздо более эффективной связкой для поиска нужных белков.

3. Поиск последовательности Шайна — Дальгарно в геноме *Paracidovorax avenae*.

В качестве консенсусной возьмем последовательность AGGAGG.
Поиск выполнялся командой:

fuzznuc GCF_003029985.1_ASM302998v1_genomic.fna -pattern 'A-G-G-A-G-G' -complement Y -outfile sd.fuzznuc
Опция -complement Y нужна для поиска как на обратной, так и на прямой цепях.
Обнаружено – 864 находки на прямой цепи и 1740 на обратной.

GC-состав генома данной бактерии: 62.86%. Длина – 5594294 п.о.

p(A) = p(T) = 0.37/2 = 0.185.
p(G) = p(C) = 0.63/2 = 0.315.

p(AGGAGG) = (0.185)^2 * (0.315)^4 = 0.000336966

Expected number of findings = p(AGGAGG) *len(genome)*2 = 3770 – ожидается случайных находок.

Воспользуемся Z-тестом:
real = 864 +1740 = 2604
Z = (real-exp)/sqrt(exp).
Z = −18.99. |Z| = 18.99. |Zcrit| для alpha = 0.05 = 1.96. => Различие имеется для сколько угодно малого уровня значимости.

Однако, воспользовавшись геномной таблицей я выяснил, что ни одна из 12 случайных находок не находится в правильной позиции относительно старт-кодона какого-либо CDS. Вероятнее всего, *Paracidovorax* имеет другую последовательность Шайна-Дальгарно.