

Домашнее Задание №4

ФИО: Мотузенко Кристина Сергеевна

1. Посчитать IQ бактерии

1.1. Выберите бактерию

Для анализа была выбрана бактерия *Mycobacterium avium*.

1.2. Скачайте последовательности всех белков своей бактерии

Последовательность всех белков скачали с помощью текстового браузера elinks.

The screenshot shows an FTP directory listing for the path `ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/Mycobacterium_avium/representative/GCF_000696715.1_Ma...`. The listing includes various files such as `GCF_000696715.1_Ma..._assembly_report.txt`, `GCF_000696715.1_Ma..._assembly_stats.txt`, `GCF_000696715.1_Ma..._cds_from_genomic.fna.gz`, `GCF_000696715.1_Ma..._feature_count.txt.gz`, `GCF_000696715.1_Ma..._feature_table.txt.gz`, `GCF_000696715.1_Ma..._genomic.fna.gz`, `GCF_000696715.1_Ma..._genomic.gbff.gz`, `GCF_000696715.1_Ma..._genomic.gff.gz`, `GCF_000696715.1_Ma..._genomic.gtf.gz`, `GCF_000696715.1_Ma..._protein.faa.gz`, `GCF_000696715.1_Ma..._protein.gpff.gz`, `GCF_000696715.1_Ma..._rna_from_genomic.fna.gz`, `GCF_000696715.1_Ma..._translated_cds.faa.gz`, `GCF_000696715.1_Ma..._assembly_wgsmaster.gbff.gz`, `README.txt`, `annotation_hashes.txt`, `assembly_status.txt`, and `md5checksums.txt`.

Below the listing, a dialog box titled "Save to file" is shown, with the file path `./GCF_000696715.1_Ma..._protein.faa.gz` entered. The dialog has "OK" and "Cancel" buttons.

1.3. Скачайте выравнивания-затравки для всех нужных доменов

	PF00015_seed	07.04.2022 15:16	Файл "FASTA"	2 КБ
	PF00069_seed	07.04.2022 15:17	Файл "FASTA"	17 КБ
	PF00211_seed	07.04.2022 15:17	Файл "FASTA"	6 КБ
	PF00512_seed	08.04.2022 18:52	Файл "FASTA"	35 КБ
	PF00563_seed	08.04.2022 18:53	Файл "FASTA"	21 КБ
	PF00990_seed	08.04.2022 18:55	Файл "FASTA"	9 КБ
	PF01295_seed	08.04.2022 18:57	Файл "FASTA"	15 КБ
	PF01928_seed	08.04.2022 18:57	Файл "FASTA"	16 КБ
	PF01966_seed	08.04.2022 18:58	Файл "FASTA"	37 КБ
	PF02518_seed	08.04.2022 18:58	Файл "FASTA"	267 КБ
	PF07536_seed	08.04.2022 18:59	Файл "FASTA"	4 КБ
	PF07568_seed	08.04.2022 18:59	Файл "FASTA"	7 КБ
	PF07730_seed	08.04.2022 19:00	Файл "FASTA"	18 КБ

1.4. Установите HMMER

Установили на ubuntu.

1.5. Напишите скрипт, который запустит hmmer на всех выравниваниях

Был написан bash-скрипт:

```
for I in *.fasta; do ~/hmmer-3.3.2/src/hmmbuild ~/profiles/$i.hmm $i;
done
```

```
kristi@DESKTOP-RMMSQNG: /mnt/c/jupyter/applied_bioinformatics/hw5/seed
(base) kristi@DESKTOP-RMMSQNG:~$ cd /mnt/c/jupyter/applied_bioinformatics
(base) kristi@DESKTOP-RMMSQNG:/mnt/c/jupyter/applied_bioinformatics/hw5/seed$ for i in *.fasta; do /mnt/c/jupyter/Applied_bioinformatics/hw5/hmmer-3.3.2/src/hmmbuild /mnt/c/jupyter/Applied_bioinformatics/hw5/profiles/$i.hmm $i;done
# hmmbuild :: profile HMM construction from multiple sequence alignments cd seed
# HMMER 3.3.2 (Nov 2020); http://hmmer.org/
# Copyright (C) 2020 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
#
# input alignment file:      PF00015_seed.fasta
# output HMM file:          /mnt/c/jupyter/Applied_bioinformatics/hw5/profiles/PF00015_seed.fasta.hmm
#
# -----
# idx name      nseq alen mlen eff_nseq re/pos description
# -----
# 1 PF00015_seed      9  192  172    1.11  0.591
#
# CPU time: 0.08u 0.02s 00:00:00.10 Elapsed: 00:00:00.17
# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.3.2 (Nov 2020); http://hmmer.org/
# Copyright (C) 2020 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
#
# input alignment file:      PF00069_seed.fasta
# output HMM file:          /mnt/c/jupyter/Applied_bioinformatics/hw5/profiles/PF00069_seed.fasta.hmm
#
# -----
# idx name      nseq alen mlen eff_nseq re/pos description
# -----
# 1 PF00069_seed     38  419  263    2.63  0.590
#
# CPU time: 0.13u 0.02s 00:00:00.15 Elapsed: 00:00:00.20
# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.3.2 (Nov 2020); http://hmmer.org/
# Copyright (C) 2020 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
#
# input alignment file:      PF00211_seed.fasta
# output HMM file:          /mnt/c/jupyter/Applied_bioinformatics/hw5/profiles/PF00211_seed.fasta.hmm
#
# -----
# idx name      nseq alen mlen eff_nseq re/pos description
# -----
# 1 PF00211_seed     19  276  183    1.94  0.590
#
# CPU time: 0.08u 0.01s 00:00:00.09 Elapsed: 00:00:00.13
# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.3.2 (Nov 2020); http://hmmer.org/
# Copyright (C) 2020 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
#
# input alignment file:      PF00512_seed.fasta
# output HMM file:          /mnt/c/jupyter/Applied_bioinformatics/hw5/profiles/PF00512_seed.fasta.hmm
#
# -----
# idx name      nseq alen mlen eff_nseq re/pos description
# -----
# 1 PF00512_seed    265  109   66   30.57  0.850
#
# CPU time: 0.04u 0.01s 00:00:00.05 Elapsed: 00:00:00.08
# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.3.2 (Nov 2020); http://hmmer.org/
# Copyright (C) 2020 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
#
# -----
```














Получили файлы:

PF00015_seed.fasta.hmm	08.04.2022 19:04	Файл "HMM"	80 КБ
PF00069_seed.fasta.hmm	08.04.2022 19:04	Файл "HMM"	121 КБ
PF00211_seed.fasta.hmm	08.04.2022 19:04	Файл "HMM"	85 КБ
PF00512_seed.fasta.hmm	08.04.2022 19:04	Файл "HMM"	31 КБ
PF00563_seed.fasta.hmm	08.04.2022 19:04	Файл "HMM"	108 КБ
PF00990_seed.fasta.hmm	08.04.2022 19:04	Файл "HMM"	75 КБ
PF01295_seed.fasta.hmm	08.04.2022 19:04	Файл "HMM"	274 КБ
PF01928_seed.fasta.hmm	08.04.2022 19:04	Файл "HMM"	83 КБ
PF01966_seed.fasta.hmm	08.04.2022 19:04	Файл "HMM"	54 КБ
PF02518_seed.fasta.hmm	08.04.2022 19:04	Файл "HMM"	52 КБ
PF07536_seed.fasta.hmm	08.04.2022 19:04	Файл "HMM"	39 КБ
PF07568_seed.fasta.hmm	08.04.2022 19:04	Файл "HMM"	36 КБ
PF07730_seed.fasta.hmm	08.04.2022 19:04	Файл "HMM"	32 КБ

Далее запускаем скрипт уже с нашим .faa файлом с белками нашей бактерии:

```
for I in *.hmm; do ~/hmmer-3.3.2/src/hmmsearch $i ~/hw5/$ma_protein.faa
> res_$i.txt; done
```

Получили файлы (названия файлов выглядят не очень, но разобраться можно):

 res_PF00015_seed.fasta.hmm	09.04.2022 12:56	Файл "TXT"	4 КБ
 res_PF00069_seed.fasta.hmm	09.04.2022 12:56	Файл "TXT"	22 КБ
 res_PF00211_seed.fasta.hmm	09.04.2022 12:56	Файл "TXT"	20 КБ
 res_PF00512_seed.fasta.hmm	09.04.2022 12:56	Файл "TXT"	17 КБ
 res_PF00563_seed.fasta.hmm	09.04.2022 12:56	Файл "TXT"	8 КБ
 res_PF00990_seed.fasta.hmm	09.04.2022 12:56	Файл "TXT"	13 КБ
 res_PF01295_seed.fasta.hmm	09.04.2022 12:56	Файл "TXT"	2 КБ
 res_PF01928_seed.fasta.hmm	09.04.2022 12:56	Файл "TXT"	3 КБ
 res_PF01966_seed.fasta.hmm	09.04.2022 12:56	Файл "TXT"	14 КБ
 res_PF02518_seed.fasta.hmm	09.04.2022 12:56	Файл "TXT"	35 КБ
 res_PF07536_seed.fasta.hmm	09.04.2022 12:56	Файл "TXT"	3 КБ
 res_PF07568_seed.fasta.hmm	09.04.2022 12:56	Файл "TXT"	3 КБ
 res_PF07730_seed.fasta.hmm	09.04.2022 12:56	Файл "TXT"	8 КБ

1.6. Проанализируйте результаты, посчитайте число сигнальных белков

Весь анализ проводили вручную.

Тип ферментов	Домен 1	Домен 2
Histidine kinases	<p>phosphoacceptor domain: HisKA [Pfam:PF00512]</p> <p>WP_230587751.1 WP_023866367.1 WP_196244515.1 WP_003878943.1 WP_095764104.1 WP_033729966.1 WP_009975281.1 WP_023866309.1 WP_009978773.1 WP_023866422.1 WP_023866818.1 WP_011723897.1</p> <p>HisKA_2 [Pfam:PF07568]</p> <p>WP_011725921.1</p> <p>HisKA_3 [Pfam:PF07730]</p> <p>WP_011726582.1 WP_023864974.1 WP_031344159.1 WP_023864970.1</p> <p>HWE_HK [Pfam:PF07536]</p> <p>WP_011725921.1</p> <p>$\sum = 18$</p>	<p>ATPase domain: HATPase_c [Pfam:PF02518]</p> <p>WP_230587751.1 WP_023866367.1 WP_196244515.1 WP_003878943.1 WP_095764104.1 WP_033729966.1 WP_009975281.1 WP_023866309.1 WP_009978773.1 WP_023866422.1 WP_023866818.1 WP_011723897.1</p> <p>WP_011725921.1</p> <p>WP_011726582.1 WP_023864974.1 WP_031344159.1 WP_023864970.1</p> <p>WP_011725921.1</p>

Methyl-accepting chemotaxis proteins	Methyl-accepting protein (MCP) domain: [Pfam:PF00015] -	
Ser/Thr/Tyr kinases	Ser/Thr/Tyr kinase (STYK) domain: [Pfam:PF00069] WP_009974173.1 WP_033729082.1 WP_080710794.1 WP_023865713.1 WP_033730043.1 WP_023865452.1 WP_023865708.1 WP_033730321.1 WP_033729883.1 WP_159105796.1 WP_023862161.1 $\Sigma = 11$	
Diguanylate cyclases	GGDEF domains: [Pfam:PF00990] WP_230587770.1 WP_023866398.1 WP_023864771.1 WP_031348914.1 WP_023867208.1 WP_033730287.1 WP_031348840.1 $\Sigma = 7$	
Adenylate cyclases	AC1 domains: [Pfam:PF01295], 0 AC2 domains: [Pfam:PF01928], WP_033729611.1 or AC3 domains: [Pfam:PF00211] WP_023866177.1 WP_023865579.1 WP_023866202.1	

	WP_023865352.1 WP_003873209.1 WP_033729391.1 WP_010948854.1 WP_009975646.1 WP_033730454.1 WP_023864866.1 WP_033729880.1 WP_033729583.1 $\Sigma = 13$	
Predicted c-di-GMP phosphodiesterases	EAL domains: [Pfam:PF00563], WP_023866398.1 WP_230587770.1 WP_031348914.1 HD-GYP domain: [Pfam:PF01966] WP_023867189.1 WP_003872507.1 WP_009977758.1 WP_196244571.1 WP_023865145.1 WP_005116484.1 WP_003876695.1 $\Sigma = 9$	

В сумме 58 сигнальных белков.

Всего 4457.

1.7. Посчитайте IQ

Разделите число сигнальных белков на общее число белков и запишите результат в таблицу.

$$IQ = \frac{58}{4457} = 0.013$$

Посчитаем по формуле из статьи:

```
In [9]: n = 58
```

```
In [10]: L = 4457
```

```
In [11]: IQ = 5*(10**4)*((n-5)**(1/2))*(L**(-1))
```

```
In [12]: IQ
```

```
Out[12]: 81.67051704375721
```