Национальный исследовательский университет «Высшая школа экономики»

(НИУ ВШЭ)

Факультет компьютерных наук

Магистерская программа: Анализ данных в биологии и медицине

Отчет к домашнему заданию №5

по курсу:

«Практическая биоинформатика»

Преподаватель: Коновалов Дмитрий Львович

Выполнила: Осинцева Екатерина Дмитриевна

Группа: мАДБМ21

Москва 2022

Цель работы*:* оценить IQ бактерии по доле сигнальных белков среди всех белков.

Количество кодируемых в геноме бактерии сигнальных белков (или их долю в общем наборе белков) можно использовать как меру способности организма приспосабливаться к различным условиям, то есть как меру «бактериального IQ» [1]. В статье Михаила Гальперина предложено рассметривать в качестве основных сигнальных белков 6 типов ферментов (Таблица 1).

Чтобы белок в геноме анализируемой бактерии был признан гистидин киназой, необходимо, чтобы он содержал 2 домена: как АТФазный, так и фосфоакцепторный домен (любой из четырех).

Таблица 1 – Основные типы сигнальных белков, свидетельствующие об «интеллекте» бактерии

Тип ферментов	Домен 1	Домен 1
Histidine kinases	HisKA [Pfam:PF00512] HisKA_2 [Pfam:PF07568] HisKA_3 [Pfam:PF07730] HWE_HK [Pfam:PF07536]	ATPase domain: HATPase_c [Pfam:PF02518]
Methyl-accepting chemotaxis proteins	Methyl-accepting protein (MCP) domain: [Pfam:PF00015]	
Ser/Thr/Tyr kinases	Ser/Thr/Tyr kinase (STYK) domain: [Pfam:PF00069]	
Diguanylate cyclases	GGDEF domains: [Pfam:PF00990]	
Adenylate cyclases	AC1 domains: [Pfam:PF01295], AC2 domains: [Pfam:PF01928], or AC3 domains: [Pfam:PF00211]	
Predicted c-di-GMP phosphodiesterases	EAL domains: [Pfam:PF00563], HD-GYP domain: [Pfam:PF01966]	

1.1. Выберите бактерию

Микроорганизмы, населяющие стабильные экологические ниши, кодируют относительно примитивные сигнальные системы, тогда как микроорганизмы окружающей среды обычно имеют сложные системы восприятия окружающей среды и передачи сигналов [1]. Было любопытно оценить IQ бактерии, обитающей в почвах.

Была выбрана бактерия вида *Agrobacterium tumefaciens*. Это вид грамотрицательных, облигатно аэробных палочковидных почвенных бактерий рода *Rhizobium*. Выбранная бактерия способна трансформировать клетки растений при помощи специальной плазмиды и приводить к образованию у растений опухолей (корончатых галлов)[2].

1.2. Скачайте последовательности всех белков своей бактерии

Рассматривался штамм *Agrobacterium tumefaciens* K84. Сборка – ASM1626v1.
Скачаем последовательности всех белков выбранной бактерии из командной строки с помощью команды `wget`. Скачать последовательность можно по ссылке:

```
https://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/Agrobacterium_tumefaciens/latest_assembly_versions/GCF_000016265.1_ASM1626v1/GCF_000016265.1_ASM1626v1_protein.faa.gz
```

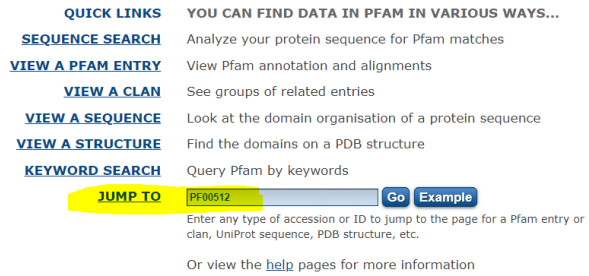
После скачивания распакуем архив с помощью команды `gunzip`.

```
oslik08@LAPTOP-DR23BJOV:~$ mkdir pb_hw5
oslik08@LAPTOP-DR23BJOV:~$ cd pb_hw5
oslik08@LAPTOP-DR23BJOV:~/pb_hw5$ wget https://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/Agrobacterium_tumefac
iens/latest_assembly_versions/GCF_000016265.1_ASM1626v1/GCF_000016265.1_ASM1626v1_protein.faa.gz
--2022-04-06 23:36:51--  https://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/Agrobacterium_tumefaciens/latest_as
sembly_versions/GCF_000016265.1_ASM1626v1/GCF_000016265.1_ASM1626v1_protein.faa.gz
Resolving ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)... 130.14.250.7, 165.112.9.229, 2607:f220:41f:250::229, ...

Connecting to ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.7|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1392020 (1.3M) [application/x-gzip]
Saving to: ‘GCF_000016265.1_ASM1626v1_protein.faa.gz’

GCF_000016265.1_ASM1626v1_pro 100%[=================================================>]   1.33M  1.43MB/s    in 0.
9s

2022-04-06 23:36:53 (1.43 MB/s) - ‘GCF_000016265.1_ASM1626v1_protein.faa.gz’ saved [1392020/1392020]
```

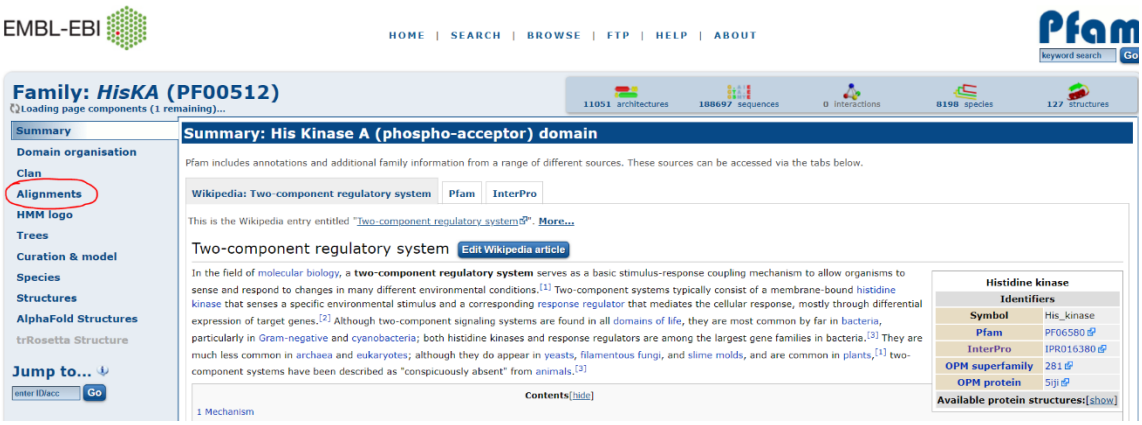
1.3. Скачайте выравнивания-затравки для всех нужных доменов

Скачаем выравнивания-затравки для каждого из 13-ти интересующих нас доменов (Таблица 1). Для этого воспользуемся сайтом PFAM.

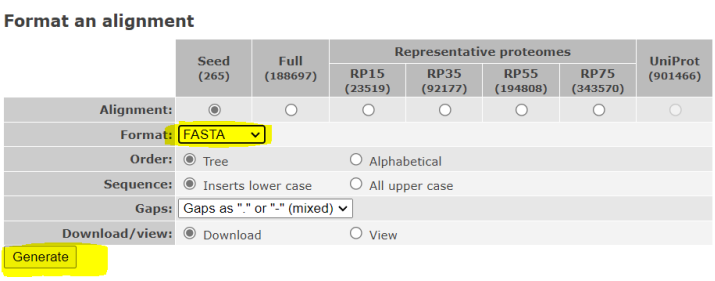
- Введем идентификатор домена в Pfam (указан в Таблице 1).



- Перейдем во вкладку Alignments (обведена красным):



- Выбираем формат FASTA и нажимаем «Generate»



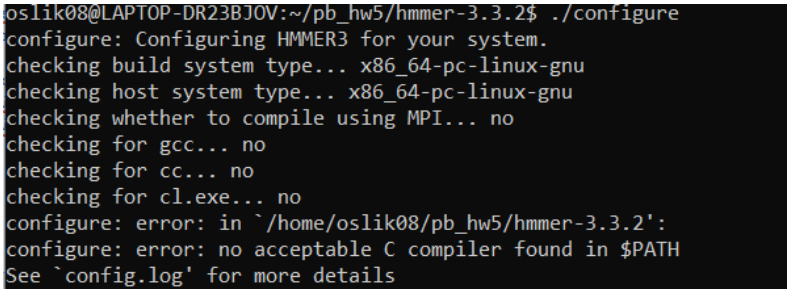
В результате чего файл с seed автоматически скачивается.

1.4. Установите HMMER

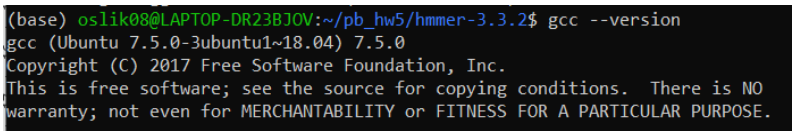
Для установки программы были выполнены следующие команды в терминале:

```
sudo apt-get update
sudo apt install build-essential

wget eddylab.org/software/hmmer/hmmer.tar.gz
tar xf hmmer.tar.gz
cd hmmer-3.3.2
./configure
make
make check
```

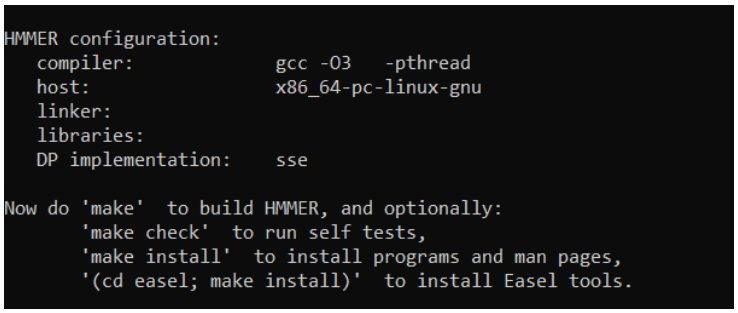
Первые две команды были выполнены, так как не были установлены компиляторы и при конфигурации, соответственно, возникала ошибка:



После выполнения первых двух команд была проверена версия gcc:



Выполнение конфигурации:



1.5. Напишите скрипт, который запустит hmmer на всех выравниваниях

Теперь нам необходимо для каждого из 13 доменов построить скрытую Марковскую модель по выравниваниям-затравкам из PFAM, которые были скачаны в пункте 1.3. Для этого нужно воспользоваться командой **hmmbuild**. Затем с помощью команды **hmmsearch** мы «пройдемся» последовательно каждым из построенных профилей по белкам бактерии (файл GCF_000016265.1_ASM1626v1_protein.faa). Запускаемые файлы программ hmmbuiled и hmmsearch лежат в папке src (путь из текущей директории проекта: ./hmmer-3.3.2/src).

Все файлы с выравниваниями были помещены в директорию seeds, файлы с профилями сохранялись в папку profile, а файлы с результатами поиска – в директорию results, которые были предварительно созданы командой **mkdir**.

Для построения профиля и поиска им по последовательностям белка, был написан и затем запущен следующий bash-скрипт (файл в архиве bash_script_hw5.sh).

```
#!/bin/bash
for file in ./seeds/*
do
b=$(basename $file)
./hmmer-3.3.2/src/hmmbuild ./profiles/profile_$b.hmm $file
./hmmer-3.3.2/src/hmmsearch --notextw  ./profiles/profile_$b.hmm
./GCF_000016265.1_ASM1626v1_protein.faa > ./results/rez_file_$b.txt
done
```

Здесь добавлена опция `--notextw`, которая ответственна за выведение полного описания белка в поле Description в результирующем файле **hmmsearch**.

Для того чтобы скрипт заработал, нужно указать для него флаг исполняемости:

```
chmod ugo+x bash_script_hw5.sh
```

Запустим скрипт:

```
./bash_script_hw5.sh
```

Ниже приведен результат выполнения скрипта, а именно команд для последнего файла с затравкой:

```
# 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:             ./seeds/PF07730_seed.fasta
# output HMM file:                  ./profiles/profile_PF07730_seed.fasta.hmm
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

# idx name                  nseq  alen  mlen eff_nseq re/pos description
#---- -------------------- ----- ----- ----- -------- ------ -----------
1     PF07730_seed           163    86    68     8.49  0.826
```

1.6. Проанализируйте результаты, посчитайте число сигнальных белков

Хорошими находками в каждом из файлов с результатами поиска мы считаем те, которые находятся выше ------ inclusion threshold -----, например, здесь «хорошей» находкой будет лишь один белок:

```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.3.2 (Nov 2020); http://hmmer.org/
# Copyright (C) 2020 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query HMM file:                  ./profiles/profile_PF00069_seed.fasta.hmm
# target sequence database:        ./GCF_000016265.1_ASM1626v1_protein.faa
# max ASCII text line length:      unlimited
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       PF00069_seed  [M=263]
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence       Description
    ------- ------ -----    ------- ------ -----   ---- --  --------       -----------
     0.0091   14.0   0.0      0.019   13.0   0.0    1.4  1  WP_007702710.1  MULTISPECIES: 2-polyprenylphenol 6-hydroxylase [Agrobacterium]
  ------ inclusion threshold ------
       0.02   12.9   0.0      0.032   12.2   0.0    1.2  1  WP_015917718.1  MULTISPECIES: phosphotransferase [Agrobacterium]
      0.026   12.5   0.3       0.19    9.7   0.4    2.1  2  WP_012653153.1  aminoglycoside phosphotransferase family protein [Agrobacterium rhizogenes]
```

Порог ------ inclusion threshold ----- может отсутствовать файле, если не было найдено ни одной хорошей находки или если все найденные последовательности оказались выше ------ inclusion threshold -----, пример такого случая приведен ниже:

```
Query:       PF07536_seed  [M=83]
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence       Description
    ------- ------ -----    ------- ------ -----   ---- --  --------       -----------
    4.5e-22   77.2   0.0    9.5e-22   76.2   0.0    1.5  1  WP_174013793.1  MULTISPECIES: sensor histidine kinase [Agrobacterium]
    1.2e-18   66.2   0.8    1.2e-18   66.2   0.8    2.1  2  WP_034517514.1  MULTISPECIES: sensor histidine kinase [Agrobacterium]
    9.4e-15   53.7   0.6    2.2e-14   52.5   0.6    1.7  1  WP_012649235.1  MULTISPECIES: PAS domain S-box protein [Agrobacterium]


Domain annotation for each sequence (and alignments):
```

Хорошей находкой в случае гистидин киназы считается та, которая содержит домены обоих типов: АТФазный и любой из фосфоакцепторов. Поэтому нам необходимо проверить каждый из обнаруженных белков, содержащий домен PF02518, на наличие у него любого из доменов PF00512, PF07568, PF07730, PF07536.

Анализ проводился в Jupyter Notebook. Далее я приведу скриншоты исполняемых на том или ином шаге ячеек, поскольку так код более читаемый, но на всякий случай к отчету будет прикреплен ноутбук.

1) Сначала пройдемся по всем 13 файлам и создадим словарь, в котором ключом будет ID домена в PFAM, а значением – список последовательностей, располагающихся выше ------ inclusion threshold ----- в результате поиска профилем домена по последовательностям белков бактерии. (В таблице поле Sequence – это 9 колонка).

```python
files = [f for f in os.listdir("./results")]  # список имен файлов
domains = {} # словарь, в котором ключи - id доменов, а значения - набор белков, в которых был найден профиль домена

for f in files:

    file = open('./results/' + f) # открываем текущий файл
    domain = f[9:16] # извлекаем из имени файла название домена
    domains[domain] = []

    # прочитаем заголовок (первые 15 строк в результирующем файле)
    for _i in range(15):
        file.readline()

    # просматриваем находки и записываем хорошие в словарь
    for line in file:
        line = line.rstrip()
        if line: # если строка не пустая
            if '------ inclusion threshold ------' in line: # если порог не достигнут
                break
            else:
                columns = line.split()
                protein_id = columns[8]
                domains[domain].append(protein_id)
        else:
            break
```

2) Далее посчитаем число «хороших» находок для каждого из доменов в отдельности:

```python
# посчитаем количество находок для каждого домена

num_of_good_founds = {}
for key, value in domains.items():
    num_of_good_founds[key] = len(value)

df = pd.DataFrame.from_dict(num_of_good_founds, orient='index').rename(columns={0:'Num_of_proteins'})
df.to_excel('num_of_proteins.xlsx')
df
```

	Num_of_proteins
PF00015	22
PF00069	1
PF00211	14
PF00512	41
PF00563	16
PF00990	23
PF01295	0
PF01928	2
PF01966	5
PF02518	51
PF07536	3
PF07568	1
PF07730	4

И сведем это в единый DataFrame, который сохраним в отдельный файл.

3) Теперь найдем количество белков, которые можно считать гистидин киназами. Для этого мы найдем пересечения множеств хороших находок для домена PF02518 с идентификаторами хороших находок каждого из четырех доменов. Полученный список преобразуем в множество и найдем его мощность.

```
# найдем и посчитаем гистидин киназы
signal_proteins = []
histidine_kinasases = []

histidine_kinasase_domain1 = ['PF00512', 'PF07568',
                              'PF07730', 'PF07536']
ATPase_domain = 'PF02518'

for hkd1 in histidine_kinasase_domain1:
    hk = set(domains[hkd1]).intersection(set(domains[ATPase_domain]))
    histidine_kinasases.extend(list(hk))
    signal_proteins.extend(hk)

n_of_hd = len(set(histidine_kinasases))
print(f'Количество белков, содержащих оба домена гистидин киназы: {n_of_hd}')
```

```
Количество белков, содержащих оба домена гистидин киназы: 46
```

Получим 46 белков, которые могут быть признаны гистидин киназами. Как видно из полученной ранее таблицы, домен PF02518 был распознан в 51 последовательностях, то есть большая доля хороших находок содержит и второй домен.

4) Посчитаем количества уникальных хороших находок для каждого из остальных 5 типов ферментов.

```
macp = 'PF00015'
print(f'Количество хороших находок для Methyl-accepting chemotaxis proteins: {len(set(domains[macp]))}')

sty_kinases = 'PF00069'
print(f'Количество хороших находок для Ser/Thr/Tyr kinases: {len(set(domains[sty_kinases]))}')

diguanylate_cyclases = 'PF00990'
print(f'Количество хороших находок для Diguanylate cyclases: {len(set(domains[diguanylate_cyclases]))}')

adenylate_cyclases_domains = ['PF01295', 'PF01928', 'PF00211']
adenylate_cyclases = []
for d in adenylate_cyclases_domains:
    adenylate_cyclases.extend(domains[d])
print(f'Количество хороших находок для Adenylate cyclases: {len(set(adenylate_cyclases))}')

phosphodiesterases_domains = ['PF00563', 'PF01966']
phosphodiesterases = []
for d in phosphodiesterases_domains:
    phosphodiesterases.extend(domains[d])
print(f'Количество хороших находок для Predicted c-di-GMP phosphodiesterases: {len(set(phosphodiesterases))}')
```

```
Количество хороших находок для Methyl-accepting chemotaxis proteins: 22
Количество хороших находок для Ser/Thr/Tyr kinases: 1
Количество хороших находок для Diguanylate cyclases: 23
Количество хороших находок для Adenylate cyclases: 16
Количество хороших находок для Predicted c-di-GMP phosphodiesterases: 21
```

Видим, что хотя бы по одному белку для каждого из типов ферментов нашлось.

5) Теперь найдем общее число обнаруженных сигнальных белков. Белок не учитывается дважды, если вдруг в нем встречались домены альтернативных видов.

```
# посчитаем сигнальные белки

for key, value in domains.items():
    if (key != ATPase_domain) and (key not in histidine_kinasase_domain1):
        signal_proteins.extend(value)

num_of_signal_proteins = len(set(signal_proteins))
print(f'Общее число найденных сигнальных белков: {num_of_signal_proteins}')
```

```
Общее число найденных сигнальных белков: 114
```

Получим общее число сигнальных белков, равное 114.

1.7. Посчитайте IQ

Чтобы посчитать IQ бактерии, нам необходимо разделить число сигнальных белков на общее число белков.

```
# посчитаем количество белков

proteins = open('GCF_000016265.1_ASM1626v1_protein.faa')
num_of_proteins = 0

for l in proteins:
    if l[0] == '>':
        num_of_proteins += 1

print(f'Количество белков у выбранной бактерии: {num_of_proteins}')
print(f'IQ бактерии: {round(num_of_signal_proteins/num_of_proteins, 3)}')
```

```
Количество белков у выбранной бактерии: 6701
IQ бактерии: 0.017
```

Таким образом, IQ бактерии вида *Agrobacterium tumefaciens* оказался равным 0.017.

Также количество белков можно было найти в конце каждого из результирующих файлов:

```
Internal pipeline statistics summary:
-------------------------------------
Query model(s):                            1  (235 nodes)
Target sequences:                       6701  (2116533 residues searched)
Passed MSV filter:                       268  (0.039994); expected 134.0 (0.02)
Passed bias filter:                      234  (0.0349202); expected 134.0 (0.02)
Passed Vit filter:                        37  (0.00552156); expected 6.7 (0.001)
Passed Fwd filter:                        16  (0.0023877); expected 0.1 (1e-05)
Initial search space (Z):               6701  [actual number of targets]
Domain search space  (domZ):              16  [number of targets reported over threshold]
# CPU time: 0.03u 0.01s 00:00:00.04 Elapsed: 00:00:00.02
# Mc/sec: 18620.43
//
[ok]
```

Выводы

В данной работе мы оценили IQ бактерии вида *Agrobacterium tumefaciens* как долю сигнальных белков к общему числу белков бактерии, он оказался равным 0.017. В статье для расчета IQ была приведена следующая формула для расчета IQ [1]:

$$IQ = 5 \cdot 10^4 \cdot (n-5)^{\frac{1}{2}} \cdot L^{-1}$$

Где n – общее число сигнальных белков, L – размер всего генома (включая плазмиды) в kb.

Полная длина генома рассматриваемого штамма K84 равна 7273,3 kb [3]. Количество сигнальных белков – 114. Посчитаем для интереса IQ бактерии по этой формуле:

$$IQ = 5 \cdot 10^4 \cdot (114-5)^{\frac{1}{2}} \cdot 7273{,}3^{\ -1} = 71.8$$

В статье приведены значения IQ грамотрицательных бактерий, имеющих высокий уровень «интеллекта», у всех он превышает 130 баллов (Table 1: Bacteria with the highest adaptability index ("highest IQ"), [1]). Как можно заметить, наша бактерия им существенно проигрывает.

Стоит отметить, что у бактерии был найден хотя бы 1 фермент каждого из типов. И 15 находок совпали для доменов PF00990 и PF00563.

```
len(set(domains['PF00990']).intersection(set(domains['PF00563'])))
```

```
15
```

Ниже приведен код, который подсчитывал число пересекающихся находок и выводил идентификаторы соответствующих белков.

```python
# найдем ID совпадающих белков

signal_proteins = []

for hkd1 in histidine_kinasase_domain1:
    hk = set(domains[hkd1]).intersection(set(domains[ATPase_domain]))
    signal_proteins.extend(hk)

k = 0
for key, value in domains.items():
    if (key != ATPase_domain) and (key not in histidine_kinasase_domain1):
        for v in value:
            if v in signal_proteins:
                k += 1
                print(key, v)
        signal_proteins.extend(value)

num_of_signal_proteins = len(set(signal_proteins))
print(f'Число пересекающихся находок: {k}')
```

```
PF00990 WP_012650734.1
PF00990 WP_012650023.1
PF00990 WP_131596066.1
PF00990 WP_012649406.1
PF00990 WP_161990913.1
PF00990 WP_012649519.1
PF00990 WP_034519250.1
PF00990 WP_012651936.1
PF00990 WP_007693876.1
PF00990 WP_174082022.1
PF00990 WP_041722718.1
PF00990 WP_007699742.1
PF00990 WP_012652379.1
PF00990 WP_012649429.1
PF00990 WP_041722445.1
Число пересекающихся находок: 15
```

Получается, что в 15 белках были найдены домены как Diguanylate cyclases, так и Predicted c-di-GMP phosphodiesterases. Если, к примеру, поискать в базе NCBI первый белок из списка по ACCESSION, то мы обнаружим, что:

MULTISPECIES: EAL domain-containing protein [Agrobacterium]

Download Datasets

NCBI Reference Sequence: WP_012650734.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       WP_012650734            778 aa            linear   BCT 22-JUL-2021
DEFINITION  MULTISPECIES: EAL domain-containing protein [Agrobacterium].
ACCESSION   WP_012650734
VERSION     WP_012650734.1
KEYWORDS    RefSeq.
SOURCE      Agrobacterium
  ORGANISM  Agrobacterium
            Bacteria; Proteobacteria; Alphaproteobacteria; Hyphomicrobiales;
            Rhizobiaceae; Rhizobium/Agrobacterium group.
REFERENCE   1  (residues 1 to 778)
  AUTHORS   Galperin,M.Y., Nikolskaya,A.N. and Koonin,E.V.
  TITLE     Novel domains of the prokaryotic two-component signal transduction
            systems
  JOURNAL   FEMS Microbiol Lett 203 (1), 11-21 (2001)
```

В целом, было бы достаточно просто посмотреть в результирующий файл, поле Description:

```
Query:       PF00990_seed  [M=161]
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence       Description
    -------  ------ -----    ------- ------ -----   ---- --  --------       -----------
    2.2e-52  175.7   0.0    3.5e-52  175.0   0.0    1.4  1  WP_041722476.1  PleD family two-component system response regulator [Agrobacterium tumefaciens]
    2.6e-48  162.4   0.0    4.4e-48  161.7   0.0    1.4  1  WP_015918147.1  sensor domain-containing diguanylate cyclase [Agrobacterium sp. B131/95]
      1e-46  157.2   0.0      2e-46  156.3   0.0    1.5  1  WP_012650734.1  MULTISPECIES: EAL domain-containing protein [Agrobacterium]
    1.1e-46  157.1   0.0    1.9e-46  156.4   0.0    1.4  1  WP_012650023.1  EAL domain-containing protein [Agrobacterium tumefaciens]
      1e-45  154.0   0.0    2.5e-45  152.8   0.0    1.7  1  WP_131596066.1  EAL domain-containing protein [Agrobacterium sp. B131/95]
    1.3e-44  150.4   0.0    3.4e-44  149.0   0.0    1.7  1  WP_012651641.1  MULTISPECIES: GGDEF domain-containing protein [Agrobacterium]
```

Было бы интересно подробнее посмотреть на эти два домена и поисследовать каждую из последовательностей, а также проанализировать бактерию на предмет того, экстраверт она или интроверт.

Список источников

1 – Michael Y Galperin. A census of membrane-bound and intracellular signal transduction proteins in bacteria: Bacterial IQ, extroverts and introverts. BMC Microbiology 2005

2 – Agrobacterium tumefaciens - Wikipedia

3 – ASM1626v1 - Genome - Assembly - NCBI (nih.gov)