Выполнил Куваев Максим
2 курс СПб ВШЭ ПМИ, 2 группа

Отчет.

Часть 1

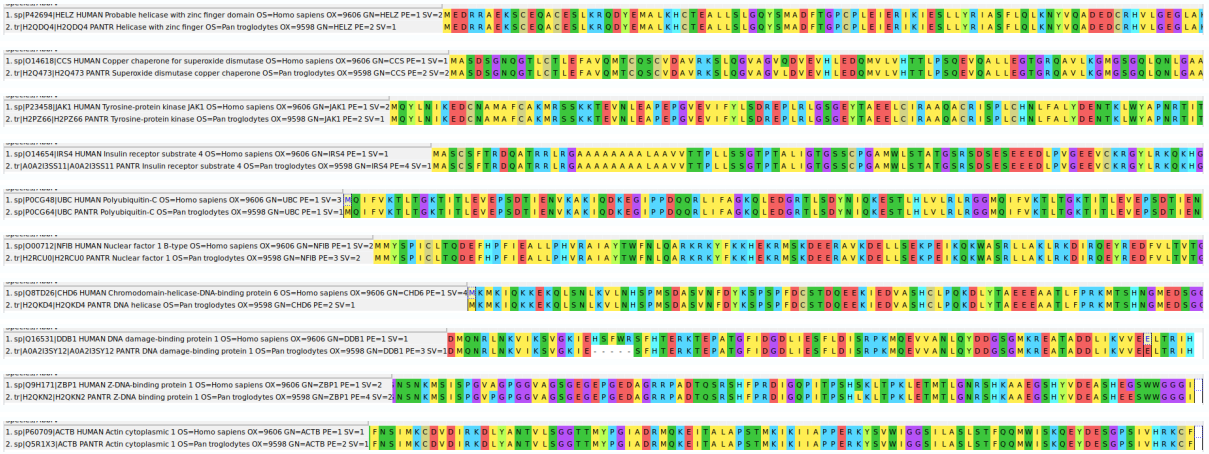
1) Для анализа были выбраны следующие протеины человека:

HELZ_HUMAN	CCS_HUMAN	JAK1_HUMAN	IRS4_HUMAN
UBC_HUMAN	NFIB_HUMAN	CHD6_HUMAN	DDB1_HUMAN
ZBP1_HUMAN	ACTB_HUMAN		

Далее при помощи продвинутого поиска в uniprot ищем ортологи для этих протеинов (протеины специально выбирались так, чтобы у каждого были ортологи).

H2QDQ4_PANTR	H2Q473_PANTR	H2PZ66_PANTR	
A0A2I3SS11_PANTR	UBC_PANTR	H2RCU0_PANTR	H2QKD4_PANTR
A0A2I3SY12_PANTR	H2QKN2_PANTR	ACTB_PANTR	

Далее выравниваем при помощи меги попарно все эти протеины с их ортологами.



И при помощи BLAST ищем среднее identity и similarity:

NAME	IDENTITIES	SIMILARITIES
H2QDQ4_PANTR	99%	99%
H2Q473_PANTR	99%	98%
H2PZ66_PANTR	99%	99%
A0A2I3SS11_PANTR	97%	98%
UBC_PANTR	100%	100%
H2RCU0_PANTR	93%	92%
H2QKD4_PANTR	98%	98%
A0A2I3SY12_PANTR	98%	98%
H2QKN2_PANTR	99%	98%
ACTB_PANTR	99%	99%

Таким образом, среднее identity равно 98.1%, среднее similarity 97.9

Часть 2

Задание 1

0) Просто считаем по формуле. После 10 циклов ПЦР:
$(2 * 2^{10})/(2 * 2^{10} + 3 * 2^{10}) = 0.4$, т.е. 40%
После 40 циклов ПЦР:
$(2 * 2^{40})/(2 * 2^{40} + 3 * 2^{40}) = 0.4$, т.е. те же самые 40%

1) Загружаем файл 80.fasta, подгружаем его в blast. Смотрим на все риды, выбираем в каждом лучшую находку. Ставим ограничение, чтобы identity было не менее 99%. Результаты сохраняем в файл reads.txt (сохраняем только названия видов у которых лучшая находка для каждого read). Если в read нет ничего, оставляем пустую строчку. Результаты сохранены в файле read.txt. Как мы можем наблюдать, хотя несколько чтений (а именно 2) и принадлежат Felis catus, все же чтений, которые принадлежат Canis lupus familiaris больше, а именно 6. Кроме того есть 3 чтения, которые принадлежат кошачьим (Acinonyx jubatus, Lynx canadensis, Panthera pardus) и 8 которые принадлежат Canis lupus (волку). Опираясь на этот результат, можно сделать вывод, что скорее всего виновник загрязнения это владелец собаки, т.е. Иванов (т.к. чтений, которые принадлежат Canis lupus familiaris больше, чем чтений, которые принадлежат Felis catus. Если считать все чтения, которые принадлежат кошачьим и собачьим, то тоже складывается впечатление, что больше загрязнения пришло от владельца собаки.
2) Теперь давайте воспользуемся файлом, который мы получили в пункте 1. Вот сводка по организмам:

Вид	Проценты
Не определен	20%
Felis catus	2%
Canis lupus familiaris	6%
Canis lupus	8%
Acinonyx jubatus	1%
Lynx canadensis	1%
Panthera pardus	1%
Sciurus carolinensis	1%
Homo sapiens	11%

Mus musculus	48%
Mastomys coucha	1%

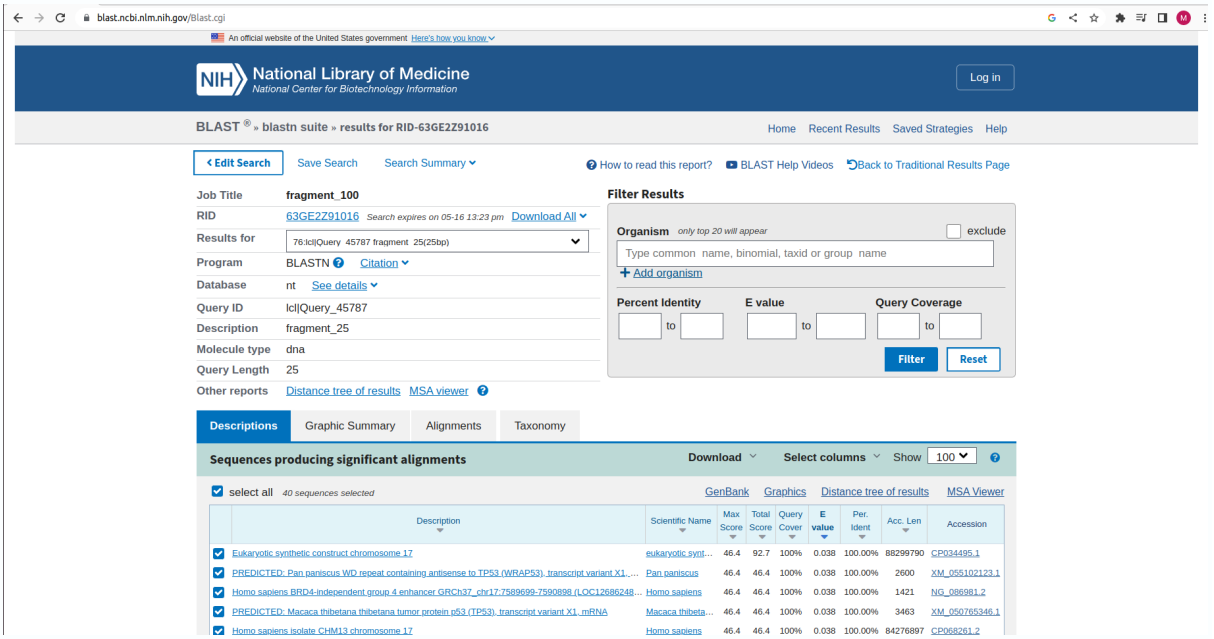
Итак, мы видим, что среди источников загрязнения есть 3 основных группы. Первая – это человек (возможно лаборант занес свое днк или что-нибудь в этом роде). Также есть кошачьи и собачьи, причем процент собачьих значительно больше. Основные выводы по этому поводу сделаны в пункте 1, тут можно лишь добавить, что возможно не один Иванов виноват в загрязнении, вина Петрова тоже есть, т.к. кошачьи откуда-то взялись. Все остальное – это 48% Mus musculus днк которой исследовали и по 1% на 2 других грызунов.

Задание 2

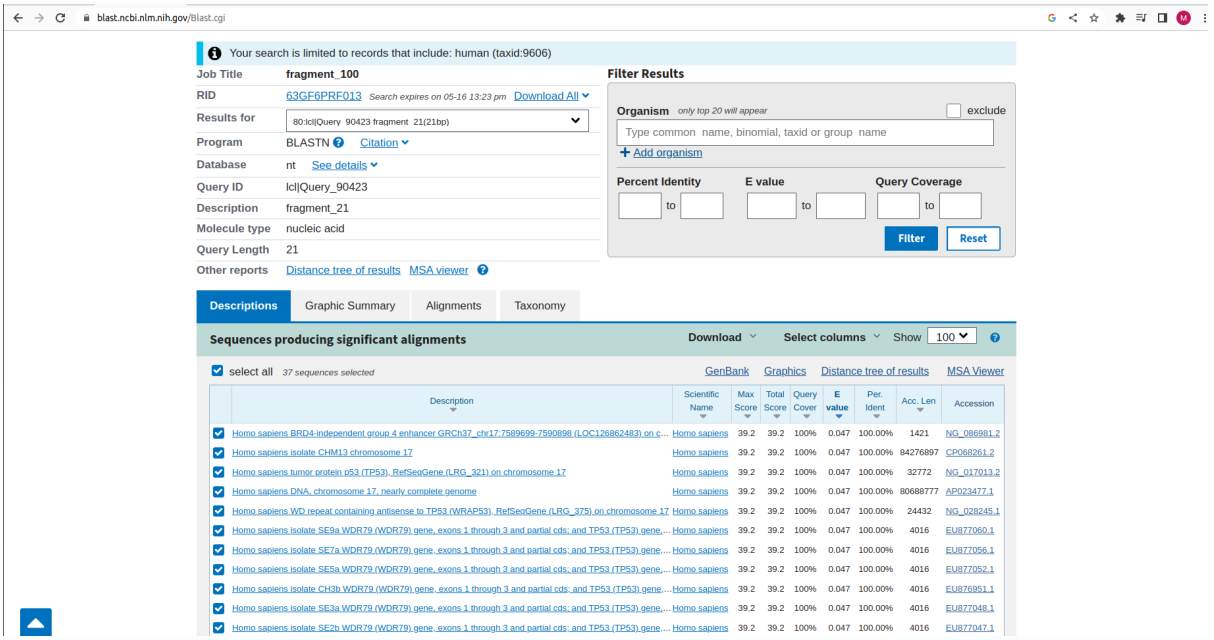
a) Воспользуемся следующим скриптом, чтобы сгенерировать требуемый fasta файл:

```
with open('seq.fasta', 'w') as file:
    s =
"CCGGCAGCCCTGTTATTGTTTGGCTCCACATTTACATTTCTGCCTCTTGCAGCAGCATT
TCCGGTTTCTTTTTGCCGGAGCAGCTCACTATTCACCCGAT"
    print(len(s))
    for i in range(100):
        file.write('>fragment_')
        file.write(str(100-i))
        file.write('\n')
        file.write(s[:(100-i)] + '\n')
```

Далее загружаем этот файл в blast, пока что не ограничиваем поиск только человеком. Получаем, что последняя длина последовательности, когда E-value ещё нормальное (меньше 0.05) это 25



b) А теперь добавим в организмы человек. Получаем, что последняя длина последовательности, когда E-value ещё нормальное (меньше 0.05) это 21, т.е. она уменьшилась.



Это так, потому что E-Value показывает вероятность случайного совпадения. Соответственно, если мы ограничим поиск только одним лишь человеком она будет меньше и нам нужен будет меньший кусок последовательности.