Биоинформатика ДЗ3
Русанов Андрей

Часть 1

Задание 1

Uniprot Entry Name (Human)	Uniprot Entry Name (Chimpanzee)	Identity, %	Similarity, %
LAT_HUMAN	A0A6D2W5U3_PANTR	77.8	78.1
FADD_HUMAN	H2Q4B6_PANTR	98.1	99.5
ACE2_HUMAN	A0A2J8KU96_PANTR	99.0	99.4
IFIH1_HUMAN	H2QIW3_PANTR	99.5	99.7
HELZ_HUMAN	H2QDQ4_PANTR	99.6	99.7
RIPK3_HUMAN	K7CE96_PANTR	98.8	99.4
RHEX_HUMAN	A0A2J8QVX4_PANTR	99.4	99.4
NMI_HUMAN	H2QIT8_PANTR	99.3	99.3
IRF3_HUMAN	K7D3V0_PANTR	69.5	71.1
PML_HUMAN	H2Q9S3_PANTR	94.2	94.2

Среднее значение Identity - 93.52%
Среднее значение Similarity - 93.98%

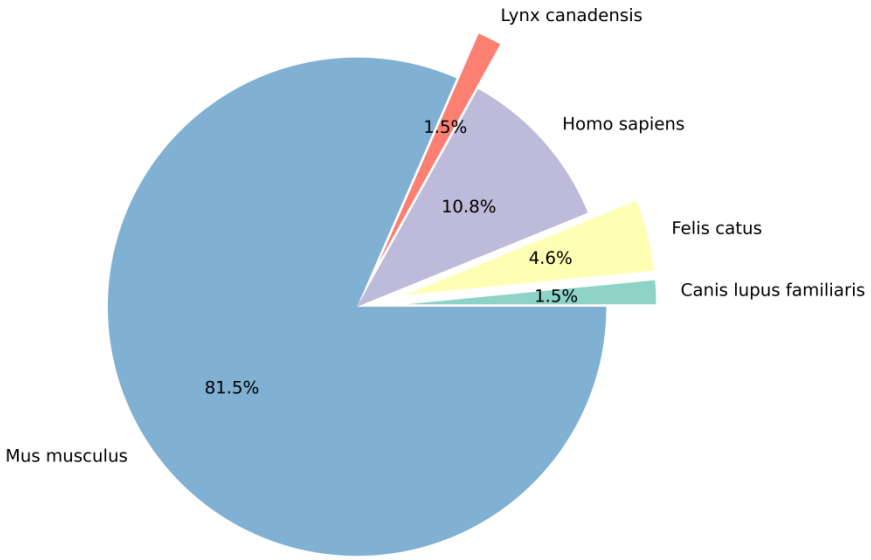
Задание 2

Скрипт get_sequences, использующий пакет Bio.Entrez выбирает 100 последовательностей длиной 100.
Далее подаём наши последовательности на вход BLAST на сайте NCBI
Полученный json-файл подаём на вход скрипту parse_alignment.

Полученное среднее значение для identity - 92.9565%

Часть 2
Задание 1

Подаём данный нам файл на вход BLAST. Получаем json-файл и с помощью скрипта анализируем его и строим график со всеми представленнми видам.



Представленные виды с alignment score ≥ 160

Вопрос 0

Если у нас есть специфичные праймеры для нужной нам последовательности, то только они будут удваиваться. Тогда после n итераций соотношение:

$$\nu = \frac{2^{n+1}}{2^{n+1}+3} \cdot 100\%$$

$$n = 10, \quad \nu \sim 99.85\%$$
$$n = 40, \quad \nu \sim 100\%$$

Вопрос 1

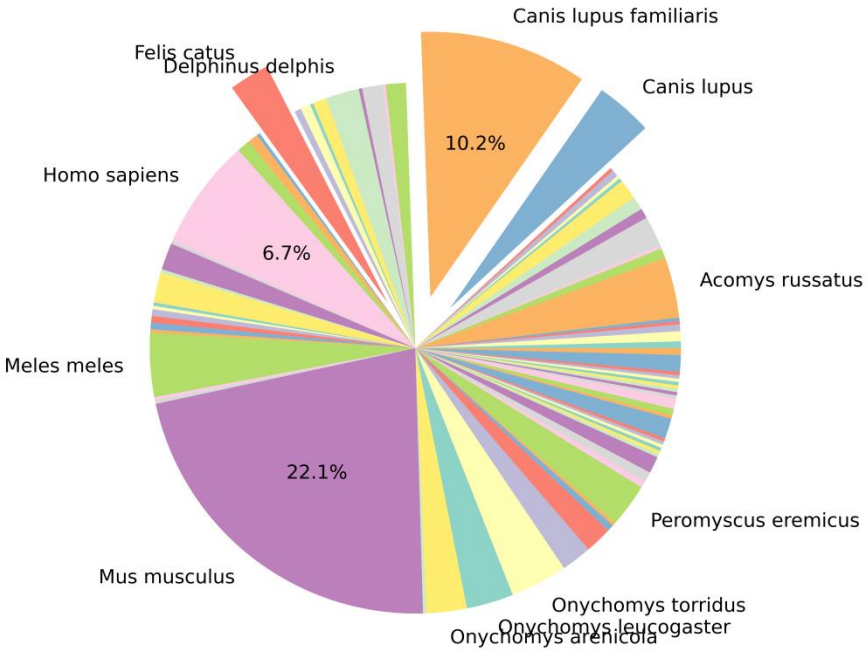
По графику видно, что в образце присутствует и ДНК кошки (Felis catus), и ДНК собаки (Canis lupus familiaris), значит и кошка, и собака являются виновниками загрязнения.

Вопрос 2

Среди загрязнителей наблюдаем еще и ДНК человека (Homo sapiens). Также было найдено ДНК канадской рыси (Lynx canadensis) - его я отнес к кошачьему источнику загрязнения.

На 9-10

С помощью скрипта пройдемся по топу 10 хитов (или меньше если 10 не нашлось) для каждого рида.



Соотношение представленности видов в топ 10 выравниваний

Задание 2

С помощью скрипта создаем нужные фрагменты, после чего подаём на вход BLAST.

Вопрос 1

При фрагменте длиной 25 E-value становиться больше 0,05.

Вопрос 2

Если при запуске BLAST ограничить поиск человеком, то критическая длина уменьшается до 21. E-value в BLAST оценивает вероятность случайного соответствия между запросом и базой данных, поэтому, уменьшая объем данных, мы уменьшаем E-value для всех находок.

На 9-10

С помощью скрипта построим графики:

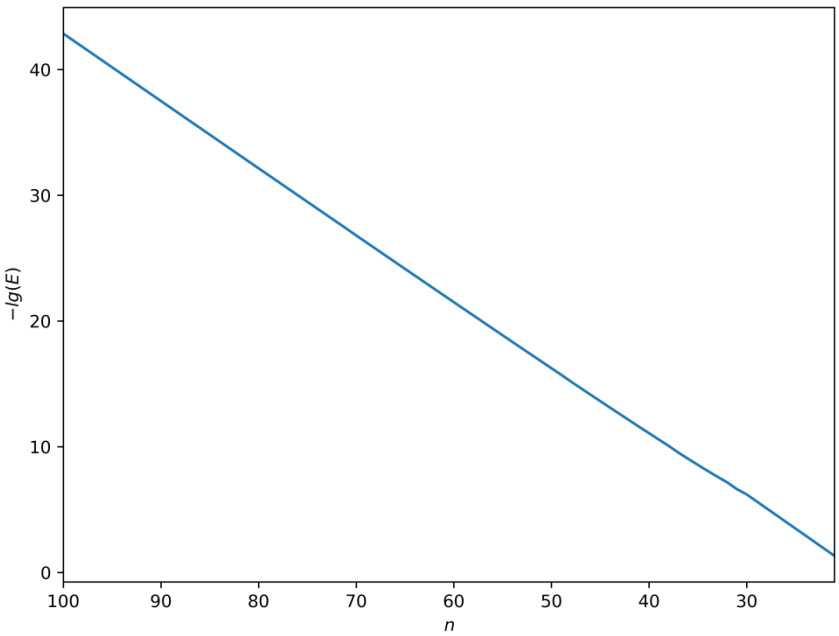


График lg(E) от n, где n длина фрагмента, а E-evalue лучшей находки.

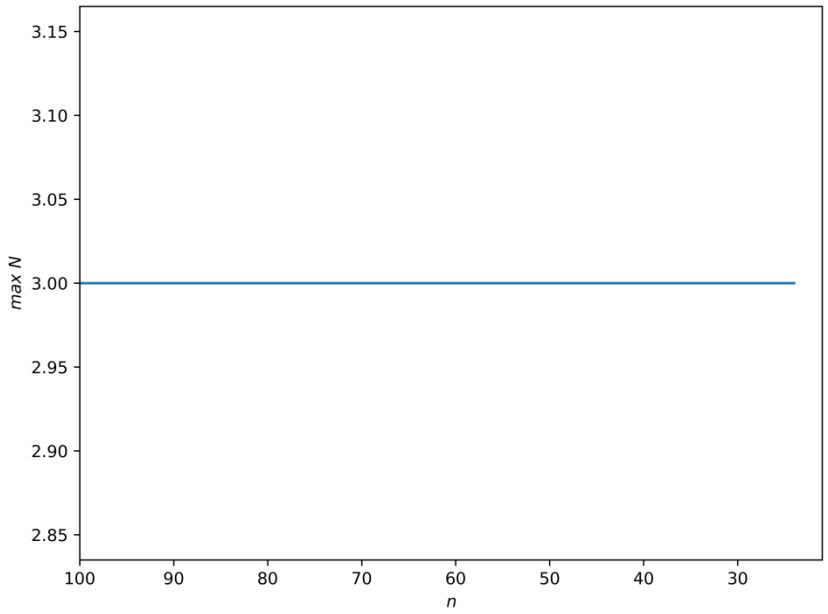


График max_N от n, где n длина фрагмента, а max_N число результатов с identity равным identity лучшего результата.