ДЗ №3 Кондратюк Александр Группа 3

**Часть 1.**

**Задание 1.**

Я выбрал следующие белки человека:

1. GBRP\_HUMAN
2. CIROP\_HUMAN
3. SOCS2\_HUMAN
4. CLRN2\_HUMAN
5. PPM1D\_HUMAN
6. GLYG2\_HUMAN
7. FLRT2\_HUMAN
8. CHSTA\_HUMAN
9. HAIR\_HUMAN
10. ANGL7\_HUMAN

Также были найдены ортологи у шимпанзе(pan troglodytes)

Далее построим парные выравнивания, используя <https://www.ebi.ac.uk/Tools/psa/emboss_needle/>

1)GBRP\_HUMAN

Ортолог - A0A2J8LXV6

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

2) CIROP\_HUMAN

Ортолог - A0A2I3SAV6

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

3) SOCS2\_HUMAN

Ортолог - A0A6D2VVR1

Изображение выглядит как текст, снимок экрана, Шрифт, алгебра

Автоматически созданное описание

4) CLRN2\_HUMAN

Ортолог - A0A2I3RH01

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

5) PPM1D\_HUMAN

Ортолог - A0A2J8JD59

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

6) GLYG2\_HUMAN

Ортолог - A0A2I3S1P4

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

7) FLRT2\_HUMAN

Ортолог - A0A6D2WRA8

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

8) CHSTA\_HUMAN

Ортолог - A0A6D2XWF8

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

9) HAIR\_HUMAN

Ортолог - A0A6D2W5E7

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

10) ANGL7\_HUMAN

Ортолог - A0A6D2WJB5

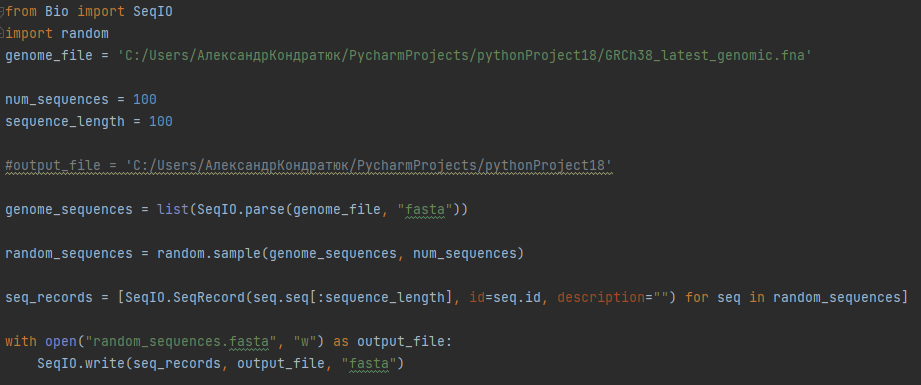
Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

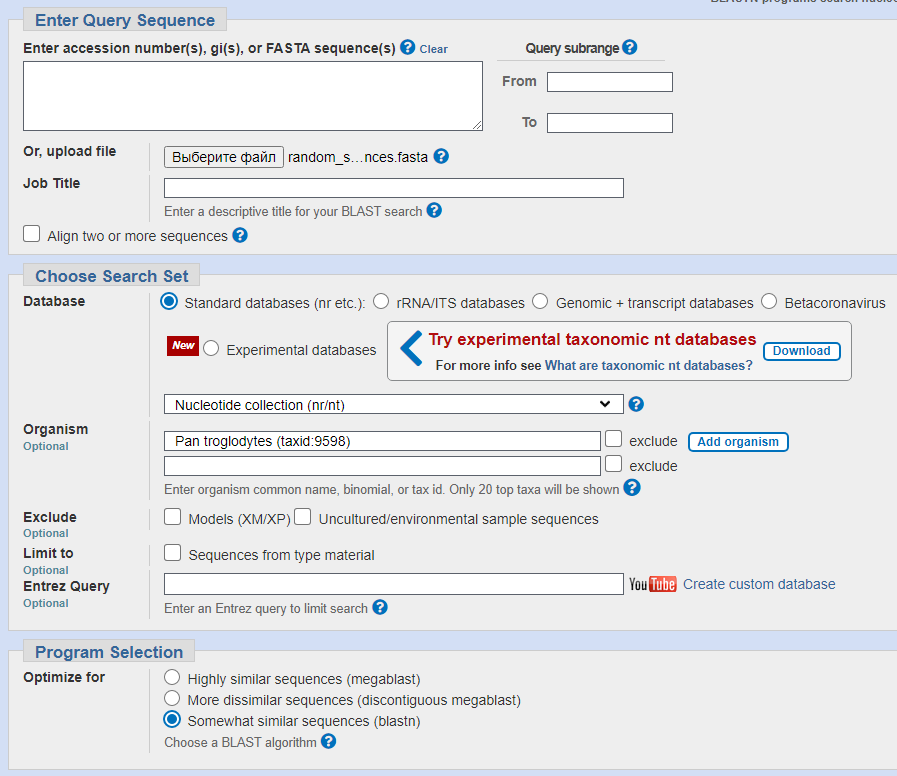
**Задание 2.**

Оценим сходство геномов найдя среднее сходство для 100 нуклеотидных  последовательностей длины 100 выбранных из генома случайно.

Возьмем с NCBI человеческий геном, и с помощью скрипта извлечем 100 случайных последовательностей и запишем их в файл.

****

Полученный файл загружаем в BLAST (<https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&BLAST_SPEC=&LINK_LOC=blasttab&LAST_PAGE=blastx>)



Полученный файл скачиваем в формате CSV и с помощью Excel находим Identity

Изображение выглядит как текст, снимок экрана, число, Параллельный

Автоматически созданное описание

Посчитаем среднее Identity:

Изображение выглядит как текст, число, Шрифт, программное обеспечение

Автоматически созданное описание

Получаем Identity = 81,4373.

**Часть 2.**

**Задание 1.**

**Вопрос 0.** На ПЦР отправили загрязненный образец состоящий из 2х молекул ДНК исследуемого организма и 3х молекул ДНК загрязнения. Считая, что после каждого цикла число молекул удваивается. Определите, сколько % молекул будет принадлежать исследуемому организму после

**а) (0,25 балла)** десяти циклов ПЦР

**б) (0,25 балла)** сорока циклов ПЦ

а) 2\*2^10 / 2\*2^10 + 3 = 0,9985 = 99,85%

б) 2 \* 2^40 / 2 \* 40 + 3 = 0,9999 = 99,99%

Получаем, что в независимости от количества циклов, доля будет равна 40%.

**Вопрос 1.**

Я скачал файл под номером 35.

Делаем BLAST и скачиваем результат.

Произведем поиск для определения есть ли среди них кошка (*Felis catus*) или собака (*Canis lupus familiaris* или *Canis familiaris*).

Сделаем поиск кошки:

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

Видим, что для кошки найдено некоторое количество совпадений.

Теперь произведем поиск собаки

Изображение выглядит как текст, снимок экрана, программное обеспечение, дисплей

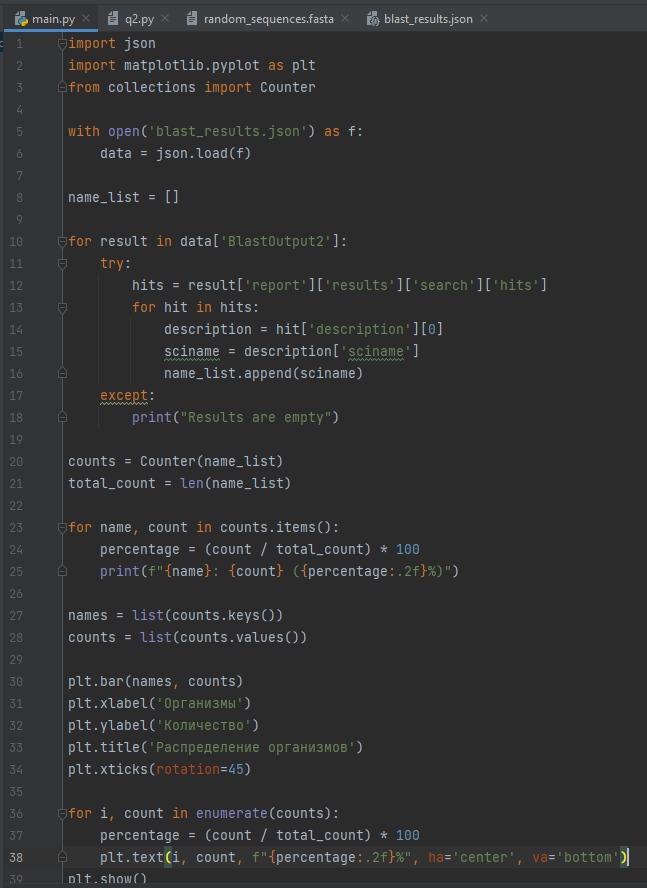
Автоматически созданное описание

Для собаки тоже были найдены совпадения, но их количество гораздо больше, чем у кошки, также есть записи с identity 100%. Могу сделать вывол, что скорее всего загрязнение появилось по вине Иванова.

**Вопрос 2.**

Определим, к геному какого организма относится каждое чтение и найдем долю для каждого представленного вида.

Напишем небольшой код:



Мы проходимся по json файлу и извлекаем все организмы из каждого результата. Затем подсчитываем количество элементов и их долю. (Файл со всеми прикреплю отдельно в формате .txt)

Вот часть результатов работы кода: Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

**На 9-10.**

Теперь чуть изменим код и сделаем топ 10 организмов

Чуть изменим код:

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Получаем результат:

Изображение выглядит как текст, Шрифт, снимок экрана

Автоматически созданное описание

Также получим таблицу: Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание(названия немного поехали)

**Задание 2. Фрагментация ДНК**

Сначала напишем скрипт, который сделает нам нужный файл, и возьмем AGTATAGTTCAGTTGTTTTCCTGTGTGAAGTCTCTGTAGCATTGACTGAATGTATAAGGGGACGAAGAGACAGAAGCTTCCTAGCGTAAGAAACATACCA

Изображение выглядит как текст, программное обеспечение, Мультимедийное программное обеспечение, Значок на компьютере

Автоматически созданное описание

Получаем файл fragments.fasta (прикреплю)

Изображение выглядит как текст, снимок экрана, дисплей

Автоматически созданное описание

а) Делаем BLAST

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, программное обеспечение, число

Автоматически созданное описание

Видим, что после 76 последовательности Blast перестает что-либо находить, а ее E-value = 0.038. Получается что при длине последовательности меньше 25, E-value становится больше 0.05.

б) Теперь ограничим поиск человеком и посмотрим что изменится

Изображение выглядит как текст, снимок экрана, программное обеспечение, веб-страница

Автоматически созданное описание

Если запускать BLAST таким образом, то значение изменится и теперь E-value становится больше 0.05 только при последовательности меньше 21.

Изображение выглядит как текст, снимок экрана, программное обеспечение, веб-страница

Автоматически созданное описание

Так как мы ограничиваем поиск человеком, уменьшается и количество возможный выравниваний, следовательно те последовательности который раньше имели e-value больше 0.05, теперь имеют меньше.

**На 9-10:**

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

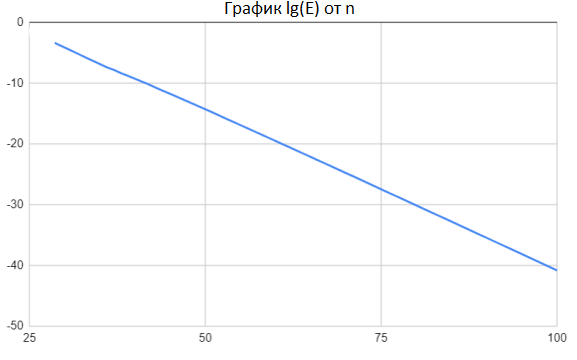
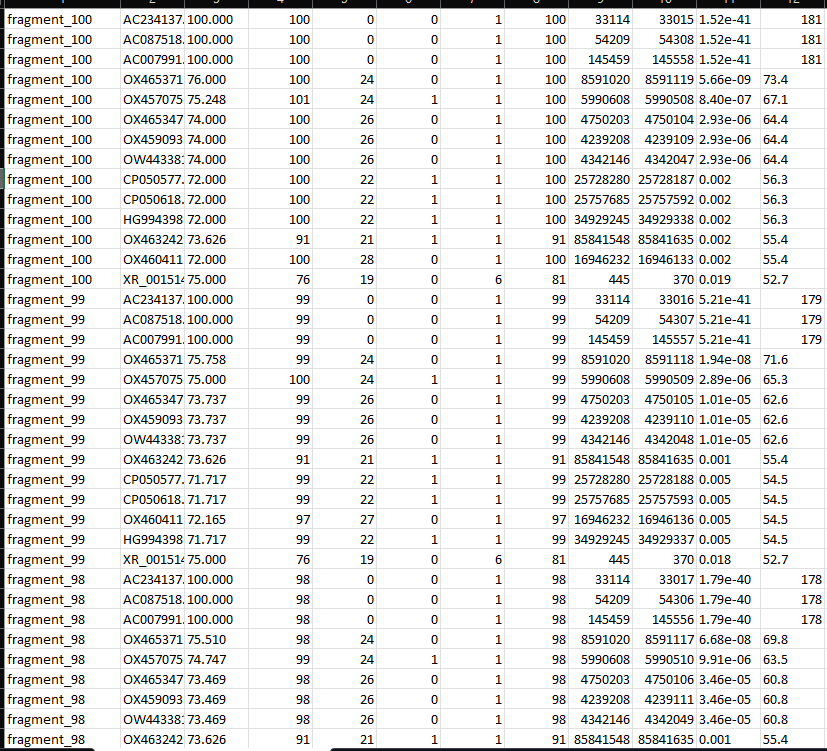
****

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

Вручную посмотрел в таблице BLAST, везде max\_N для каждой длины фрагмента будет 3(вот доказательство, таблицу также приложу blast76 называется)



Значит график будет выглядеть так:

Изображение выглядит как текст, линия, снимок экрана, График

Автоматически созданное описание