**Программа для получения отдельных данных записей GenBank из NCBI**

Ссылка на GitHub: <https://github.com/xfluffymuffin/astrovirus.git>

Описание:

Программа использует три библиотеки: pandas, biopython, collections.

1) Первым делом программа проверяет, имеется ли уже файл *output.txt*с записями GenBank, с которым можно начать работу. Если файл не обнаруживается, то происходит скачивание данных по запросу, сохраняемому как единственный параметр *query*функции *get\_gb\_records(query)***.** Если текстовый файл существует, то данные сразу анализируются функцией *parse()*.

2) Функция *get\_gb\_records(query)* загружает ID из базы данных NCBI Nucleotide, затем по этим ID скачивает соответствующие записи формата GenBank. Таким образом, при вызове функции: **get\_gb\_records('"Mamastrovirus 1"[porgn] AND "6000"[SLEN] : "8000"[SLEN]')**

будет скачано 125 записей, которые этой же функцией помещаются сначала в словарь *res*, а затем из словаря – в файл *output.txt* по следующему образцу:

…  
*GBID: OR371570*

*country: China:Fuzhou*

*host: Homo sapiens*

*collection\_date: 2022-01-04*

*strain: HAstV-4\_Fuzhou\_2022-10*

*isolate: N/A*

*isolation\_source: feces*

*…*

Из приведенного примера видно, что при отсутствии данных для какого-либо поля (в данном случае *isolate*) функция возвращает *N/A*.

3) Функция *parse()* с помощью библиотеки *pandas* создает *DataFrame,* заполняет его данными из *output.txt* и сохраняет его в переменной *data*.

Первые несколько строк датафрейма выглядят следующим образом:

*0 1*

*0 GBID OR371570*

*1 country China:Fuzhou*

*2 host Homo sapiens*

*3 collection\_date 2022-01-04*

*4 strain HAstV-4\_Fuzhou\_2022-10*

*.. ... ...*

В датафрейме содержится вся та же информация, что и в файле *output.txt*.

Далее функция *parse()* обращается к указанному датафрейму и собирает данные из строк с названиями *GBID*, *collection\_date, country, host, strain, isolate, isolation\_source*. Очевидно, что количество строк с каждым из этих названий равно количеству скачанных записей (125 при текущем запросе).

Полученные данные выводятся на экран в следующем формате:

----------------------------------------------------------------------------------------------------  
*Загружено 125 записей*

*Образцы выделены в следующих точках мира (чаще всего China, 19 раз(а)):*

*China:Fuzhou*

*China*

*Japan:Gunma, Kiryu*

*…*

*Виды-хозяева вирусов:*

*Homo sapiens*

*environment*

*Homo sapiens; infant (<5 years)*

*…*

*Даты взятия образцов (известны у 108 из 125)):*

*2022-01-04*

*2021*

*2016-05-24*

*…*

*Штаммы:*

*HAstV-4\_Fuzhou\_2022-10*

*Human/JPN/2016/JP14989*

*Human/JPN/2016/JP14665*

*…*

*Известная информация о серотипах/пациентах:*

*SY2106*

*Child with acute gastroenteritis*

*G22-011*

*…*

*Источники выделения:*

*feces*

*stool*

*sewage*

*…*

*Более полное описание - в файле output.txt*

----------------------------------------------------------------------------------------------------