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

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

Описание:

Программа использует четыре библиотеки: *pandas*, *biopython*, *collections, pickle*, а также модуль *os.path*.

1) Вначале создается словарь *res*, который в будущем будет содержать часть данных из записей GenBank.

2) Затем проверяется наличие файла *dict\_data.txt*, содержащего данные словаря res. Файл создается после первого запуска программы и необходим, чтобы избегать повторного скачивания GenBank-записей при каждом новом запуске программы – удобно, если нужно, к примеру, поменять что-либо в формате вывода данных. Для повторного скачивания данных необходимо перед запуском скрипта удалить *dict\_data.txt*.

3) Далее при отсутствии *dict\_data.txt* выполняется функция *get\_gb\_records(query)*, где *query* – запрос, отправляемый на сервер NCBI (к базе данных NCBI Nucleotide). Функция выполняет три задачи: скачивание данных из GenBank, запись данных некоторых квалификаторов в словарь *res* и сохранение всех скачанных записей GenBank в файл *records*.*gb*.

Если *dict\_data.txt* ужесуществует, то из него подгружаются данные GenBank записей, сохраненные ранее в словарь res, и функция *get\_gb\_records* игнорируется.

4) Затем функцией *file\_writer()* данные из словаря *res* сохраняются в датафрейм *pandas*, оттуда – в файл *metadata*.*csv*

5) Часть данных, сохраненная в словаре *res*, выводится в файл *main\_ output.txt*

Переменной query было передано следующее значение:

**"Mamastrovirus 1"[porgn] AND "6000"[SLEN] : "8000"[SLEN]**

Содержимое *main\_output.txt* приведено ниже (результат скачивания данных по запросу):

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

*Загружено 125 записей*

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

*China: 19*

*France: 15*

*Brazil: 15*

*Japan:Tokyo: 13*

*Kenya: 8*

*USA: 8*

*NaN: 7*

*Russia: Novosibirsk: 6*

*China: Xinjiang: 5*

*China: Fuzhou: 4*

*USA: CA: 4*

*Russia: 4*

*Japan: Gunma, Kiryu: 3*

*South Africa: 2*

*India:Kolkata: 2*

*South Korea: 2*

*China:Fuzhou: 1*

*Guinea: 1*

*China: Pingtan: 1*

*USA: NV: 1*

*Ireland: 1*

*Venezuela: 1*

*Hungary: 1*

*China: Jinzhou: 1*

*Виды-хозяева вирусов:*

*Homo sapiens: 84*

*NaN: 25*

*environment: 11*

*Homo sapiens; infant (<5 years): 2*

*Homo sapiens; sex: M; age: 9 months: 1*

*Homo sapiens; sex: F; age: 3 months: 1*

*Homo sapiens; female: 1*

*Даты взятия образцов:*

*NaN: 17*

*2015: 10*

*2014: 5*

*2019: 4*

*2021-12: 3*

*Nov-2007: 3*

*2021: 2*

*23-Jul-2007: 2*

*05-Feb-2014: 2*

*28-Feb-2006: 2*

*26-May-2021: 2*

*2018-08: 2*

*2020-03: 2*

*2011: 2*

*01-Jun-2019: 2*

*May-2014: 2*

*22-Mar-2018: 2*

*02-Sep-2004: 2*

*2022-01-04: 1*

*2016-05-24: 1*

*2016-05-07: 1*

*2016-04-24: 1*

*24-Jan-2006: 1*

*19-Mar-2014: 1*

*06-Jun-2016: 1*

*23-Jul-2006: 1*

*02-Jun-2016: 1*

*13-Jan-2021: 1*

*10-Mar-2022: 1*

*26-Mar-2022: 1*

*15-Feb-2022: 1*

*13-Feb-2022: 1*

*16-Apr-2022: 1*

*28-Mar-2022: 1*

*31-Mar-2021: 1*

*12-Aug-2018: 1*

*01-Aug-2018: 1*

*17-Jul-2018: 1*

*30-Mar-2018: 1*

*01-Nov-2018: 1*

*2020-02: 1*

*2021-10: 1*

*2021-11: 1*

*2018-10: 1*

*2022-01: 1*

*2019-07: 1*

*15-Apr-2019: 1*

*19-Jun-2019: 1*

*18-Jul-2019: 1*

*26-Apr-2019: 1*

*23-Jul-2019: 1*

*13-Apr-2019: 1*

*May-2013: 1*

*Jan-2013: 1*

*May-2017: 1*

*06-Nov-2010: 1*

*15-Jun-2011: 1*

*22-Feb-2018: 1*

*06-Feb-2018: 1*

*28-Dec-2018: 1*

*2013: 1*

*21-Feb-2004: 1*

*01-Mar-2009: 1*

*21-Apr-2008: 1*

*10-Nov-2003: 1*

*Jun-2010: 1*

*Oct-2006: 1*

*Apr-2005: 1*

*18-Nov-2010: 1*

*20-Mar-2014: 1*

*07-May-2010: 1*

*20-Oct-2005: 1*

*17-Apr-2006: 1*

*10-Aug-2005: 1*

*20-Aug-2012: 1*

*Feb-2006: 1*

*2010: 1*

*Штаммы:*

*NaN: 75*

*Oxford: 8*

*Dresden: 2*

*HAstV-4\_Fuzhou\_2022-10: 1*

*Human/JPN/2016/JP14989: 1*

*Human/JPN/2016/JP14665: 1*

*Human/JPN/2016/JP14641: 1*

*MAstV-1/Hu/BRA/TO-161/2015: 1*

*MAstV-1/Hu/BRA/TO-208/2015: 1*

*MAstV-1/Hu/BRA/TO-171/2015: 1*

*MAstV-1/Hu/BRA/TO-162/2011: 1*

*MAstV-1/Hu/BRA/TO-160/2015: 1*

*MAstV-1/Hu/BRA/TO-98/2015: 1*

*Hu/BRA/TO-32/2014: 1*

*Hu/BRA/TO-57/2014: 1*

*Hu/BRA/TO-40/2014: 1*

*Hu/BRA/TO-65/2015: 1*

*Hu/BRA/TO-26/2014: 1*

*Hu/BRA/TO-71/2015: 1*

*Hu/BRA/TO-207/2014: 1*

*SA6118: 1*

*SA10098: 1*

*KLF/HAstV5/003/2019: 1*

*KLF/HAstV5/002/2019: 1*

*KLF/HAstV3/007/2019: 1*

*KLF/HAstV2/005/2019: 1*

*KLF/HAstV1/010/2019: 1*

*KLF/HAstV1/008/2019: 1*

*KLF/HAstV1/006/2019: 1*

*KLF/HAstV1/001/2019: 1*

*Hu/US/2014/CA-RGDS-1073: 1*

*Hu/US/2014/CA-RGDS-1072: 1*

*Hu/US/2013/CA-RGDS-1071: 1*

*Hu/US/2013/CA-RGDS-1070: 1*

*Hu/US/2017/CA-RGDS-1074: 1*

*Hu/HAstV3/2016/East/Ireland: 1*

*Yuc-8: 1*

*Hu/Nyergesujfalu/HUN4520/2010/HUN: 1*

*Beijing/293/2007/CHN: 1*

*Beijing/291/2007/CHN: 1*

*Beijing/176/2006/CHN: 1*

*Beijing/128/2005/CHN: 1*

*lhar/2011/kor: 1*

*Известная информация о серотипах/пациентах:*

*NaN: 35*

*Child with acute gastroenteritis: 3*

*V1347: 2*

*SY2106: 1*

*G22-011: 1*

*G22-010: 1*

*G22-009: 1*

*G22-008: 1*

*G22-007: 1*

*G22-006: 1*

*G22-005: 1*

*G22-004: 1*

*G22-003: 1*

*G22-002: 1*

*G22-001: 1*

*SEN\_21\_286\_S1: 1*

*2021/Fuzhou/37: 1*

*2019/Pingtan/42: 1*

*2019/Fuzhou/306: 1*

*2019/Fuzhou/289: 1*

*2019/Fuzhou/26: 1*

*HAstV-4/QD-L-22019: 1*

*HAstV-4/QD-S-22022: 1*

*HAstV-4/RC-21045: 1*

*HAstV-4/QD-S-22015: 1*

*HAstV-4/QD-L-22001: 1*

*HAstV-4/RC-21047: 1*

*HAstV-4/QD-F-22056: 1*

*HAstV-4/QD-L-22040: 1*

*HAstV-4/RC-21027: 1*

*AstV/CHN/XJ/env/2018/ht8-ctg1: 1*

*AstV/CHN/XJ/env/2018/ks8-ctg1: 1*

*AstV/CHN/XJ/env/2018/ks7-ctg2: 1*

*AstV/CHN/XJ/env/2018/ht3-ctg1: 1*

*AstV/CHN/XJ/env/2018/ks11-ctg1: 1*

*HAstV1/181289/Tokyo/2018: 1*

*HAstV1/181290/Tokyo/2018: 1*

*HAstV1/192642/Tokyo/2020: 1*

*HAstV1/193055/Tokyo/2020: 1*

*HAstV1/193061/Tokyo/2020: 1*

*HAstV1/210628/Tokyo/2021: 1*

*HAstV1/210764/Tokyo/2021: 1*

*HAstV1/181937/Tokyo/2018: 1*

*HAstV1/210956/Tokyo/2021: 1*

*HAstV1/210963/Tokyo/2021: 1*

*HAstV1/210984/Tokyo/2021: 1*

*HAstV1/211185/Tokyo/2022: 1*

*HAstV3/190610/Tokyo/2019: 1*

*HAstV-3\_RVAB: 1*

*TO-161: 1*

*TO-208: 1*

*TO-171: 1*

*TO-162: 1*

*TO-160: 1*

*TO-98: 1*

*BRA/TO-32: 1*

*BRA/TO-57: 1*

*BRA/TO-40: 1*

*BRA/TO-65: 1*

*BRA/TO-26: 1*

*BRA/TO-71: 1*

*BRA/TO-207: 1*

*HAstV-4/RUS/Novosibirsk/Nsc10-N1190/2010: 1*

*HAstV-2/RUS/Novosibirsk/Nsc11-N2411/2011: 1*

*G19054: 1*

*G19053: 1*

*G19052: 1*

*G19051: 1*

*17W1028: 1*

*2013/Fuzhou/85: 1*

*Rus-Nsc04-H355: 1*

*Rus-Nsc09-B4: 1*

*Rus-Nsc08-3364: 1*

*Rus-Nsc03-H191: 1*

*Guangzhou: 1*

*Goiania/GO/12/95/Brazil: 1*

*Goiania/GO/12/94/Brazil: 1*

*Katano: 1*

*Oxford: 1*

*192-BJ07-CHN: 1*

*JZ: 1*

*kor85: 1*

*Rus-Nsc10-N358: 1*

*Rus-Nsc05-623: 1*

*Rus-Nsc06-1029: 1*

*Rus-Nsc05-430: 1*

*km1: 1*

*DL030: 1*

*Источники выделения:*

*NaN: 47*

*stool: 20*

*sewage: 20*

*feces: 19*

*cell culture: 8*

*stool sample: 4*

*Fecal Specimen: 2*

*faecal specimen of a child suffering from acute watery diarrhoea: 2*

*Stool Sample of non-polio AFP: 1*

*fecal sample: 1*

*human stool: 1*-----------------------------------------------------------------------------------------