13.03.24

**Продолжение работы над скриптом для получения данных об астровирусах из GenBank и парсинга этих данных**

1) Добавлена функция *file\_writer()*, упрощающая визуальное восприятие части кода, ответственной за запись данных в *metadata.csv* *и main\_output.txt*.

2) Изменено название выходного файла с GenBank-записями: вместо *records.gb* он теперь называется *Mamastrovirus\_1\_complete\_genome\_records.gb* – так легче сразу понять, что файл содержит.

3) Выходные файлы добавлены в репозиторий на GitHub, чтобы избежать необходимости запускать скрипт лишний раз при просмотре данных с другого компьютера.

4) Выходная информация о частотах встречаемости значений квалификаторов теперь отсортирована от больших значений к меньшим (в функцию *count\_occurrences()* добавлен метод *most\_common()* для вывода ключей и значений словаря res)

5) Значение квалификатора теперь может быть равным *NaN*, если необходимые данные отсутствуют в GenBank (понадобится в дальнейшем). В связи с этим из вывода *main\_output.txt* убрано лишнее упоминание о том, у какого количества образцов известны даты выделения: это и так понятно по количеству значений *NaN*.

6) Согласно описанным изменениям исправлен документ с описанием работы программы.

14.04.24

**Создание скрипта для последующей работы со значениями квалификаторов /gene и /product – получение csv-файла**

1) Запустил скрипт *genecds\_field\_distribution.py* на входном файле *Mamastrovirus\_1\_complete\_genome\_records*. Команда:

python alignment\_of\_orfs/genecds\_field\_distribution.py -input Mamastrovirus\_1\_complete\_genome\_records.gb -odir alignment\_of\_orfs\_output -oname Mamastrovirus\_1\_gene\_product\_output.txt

Вывод в командной строке:

*Number of records in file: 125*

*Records with annotated CDS: 121*

*CDS with product fields 353*

*CDS with gene fields 111*

Основным файлом вывода при этом послужил *Mamastrovirus\_1\_gene\_product\_output.txt*

2) Написал программу *genecds\_create\_csv.py*, которая сначала запускает скрипт *genecds\_field\_distribution.py* и по его выходным данным составляет csv-таблицу формата:

|  |
| --- |
| 3C-like serine protease motif,1A |
| NS,1A |

…

Таблица понадобится в дальнейшем как входной файл скрипта *get\_orfs\_coord.py*, достающего координаты ORF из GenBank-файла *Mamastrovirus\_1\_complete\_genome\_records.gb.*

15.03.24

**Минорные изменения во вчерашнем скрипте и попытка получения координат ORF**

1) Внес небольшие изменения в *genecds\_create\_csv.py* после более внимательного изучение GenBank файла. Теперь значение “NS” означает рамку считывания 1A, “Non-structural protein” – 1AB, а "nonstructural protein” – сразу три возможных варианта (1A, 1B, 1AB).

2) Для получения координат ORF была попытка запустить *get\_orfs\_coord.py* с помощью следующей команды:

python alignment\_of\_orfs/get\_orfs\_coord.py -input Mamastrovirus\_1\_complete\_genome\_records.gb -orf\_map Mamastrovirus\_1\_CDS\_gene\_product.csv

File with coordinates of ORFs: Mamastrovirus\_1\_complete\_genome\_records\_orf.txt

Получена следующая ошибка:

*Traceback (most recent call last):*

*File "D:\PyCharm\BioPython\alignment\_of\_orfs\get\_orfs\_coord.py", line 173, in <module>*

*orf\_coord(args.input\_file, args.orf\_map\_file, args.remove\_exceptions)*

*File "D:\PyCharm\BioPython\alignment\_of\_orfs\get\_orfs\_coord.py", line 85, in orf\_coord*

*dict\_coord[rec.name][product] = [int(feature.location.\_start) + cod\_start, int(feature.location.\_end)]*

*^^^^^^^^^^^^^^^^^^^^^^^*

AttributeError: 'CompoundLocation' object has no attribute '\_start'. Did you mean: 'start'?

Попытка исправить ее приводит к поломке вывода:

*File with coordinates of ORFs: Mamastrovirus\_1\_complete\_genome\_records\_orf.txt*

*Encountered joined locations in Z25771 entry*

*[SimpleLocation(ExactPosition(85), ExactPosition(2839), strand=1), SimpleLocation(ExactPosition(2823), ExactPosition(4380), strand=1)]*

*[85, 2839]*

19.03.24

**Получение координат ORF. Перевод GenBank-записей в fasta-формат. Разбивка геномов по координатам. Фильтрация записей для дальнейшей работы.**

1) Проблема решалась небольшим изменением формата записи строки с nonstructural protein в csv-файле. Координаты ORF получены и теперь лежат в выходном файле *Mamastrovirus\_1\_complete\_genome\_records\_orf.txt*.

2) Проверка выходного файла показывает, что полностью координаты отсутствуют только у четырех записей:

*MN510442,NA-NA,NA-NA,NA-NA*

*MN510441,NA-NA,NA-NA,NA-NA*

*MN510440,NA-NA,NA-NA,NA-NA*

*MN510439,NA-NA,NA-NA,NA-NA*

Источник всех четырех образцов – сточные воды

Частично (для одной рамки) координаты отсутствуют у 14 записей.

3) Использован скрипт parser\_gb.py, возвращающий записи в формате fasta после конвертации их из GenBank-файла. Записи помещены в файл *Mamastrovirus\_1\_complete\_genome\_records.fasta*. В результаты работы программы выводилось сообщение “ Couldn\t find key for environment”, которое, однако, не повлияло на общее число записей в итоговом файле (все 125 были сохранены)

Команда: *python alignment\_of\_orfs/parser\_gb.py -input Mamastrovirus\_1\_complete\_genome\_records.gb -min 1 -max 10000 -f 'host,country,collection\_date,strain,isolate,isolation\_source'.*

4) Использован скрипт *split\_genome.py* для разбивки геномов из fasta-файла по координатам ORF. Если в какой-либо из записей координаты для рамки отсутствовали (предположительно), то для такой рамки выводилось сообщение вида: *ON807348 not in coord file*.

Команда:

5) В итоге были получены три файла, содержащих соответственно рамки 1А, 1B и 2:

*Mamastrovirus\_1\_complete\_genome\_records\_1A.fasta*

*Mamastrovirus\_1\_complete\_genome\_records\_1B.fasta*

*Mamastrovirus\_1\_complete\_genome\_records\_2.fasta*

Каждый файл содержит 107 записей, что имеет смысл, т.к.

125 (исходных записей) – 4 (нет координат всех рамок) – 14 (нет координат у одной рамки) = 107 оставшихся записей.

14 записей в формате: GBID (источник выделения):

ON807348 (sewage), ON807318 (sewage), OQ630461 (Stool Sample of non-polio AFP), NC\_030922 (faecal specimen of a child suffering from acute watery diarrhoea), MG571777 (stool), HQ398856 (fecal sample), ATVPOLY6A (Оксфорд), FJ755405 (stool), FJ755404 (stool), FJ755403 (stool), FJ755402 (stool), KF211475 (isolate = JZ), KC342249 (stool sample), AB308374 (faecal specimen of a child suffering from acute watery diarrhoea).

6) Создано описание для скрипта *genecds\_create\_csv.py*

23.03.24

**Ручная обработка данных по координатам для проблемных записей.**

1) «Сломанные» записи с источником выделения “sewage” (4 с полностью отсутствующими координатами и 2 с частично отсутствующими) были проигнорированы как малозначимые. 12 записей подверглись ручной обработке:

Данные о рамках на самом деле были в GB-файле:

* OQ630461
* MG571777
* HQ398856
* ATVPOLY6A
* FJ755405
* FJ755404
* FJ755403
* FJ755402
* KF211475

Были взяты значения сдвига 61 нт для рамок 1A\1B и 8 нт для 1B/2:

* NC\_030922
* KC342249
* AB308374

Исправления понадобились потому, что скрипт *genecds\_field\_distribution.py* учитывал только поля “gene” и “product” в записях, но игнорировал поля “note”, в которых тоже содержалась информация о названиях рамок. Вследствие этого скрипт *get\_orfs\_coord.py* затем не сработал полностью корректно.

Измененный файл с координатами был назван *FIXED\_Mamastrovirus\_1\_complete\_genome\_records\_orf.txt*

2) *split\_genome.py* был запущен повторно для получения исправленных фаста-файлов.

Обнаружены записи, которые имеются в общем фаста-файле с геномами (и GB-файле), но отсутствуют в файле с координатами рамок:

*ON807348 not in coord file*

*ON807318 not in coord file*

*MN510442 not in coord file*

*MN510441 not in coord file*

*MN510440 not in coord file*

*MN510439 not in coord file*

*NC-030922 not in coord file*

*L23513 not in coord file*

Предпоследний файл все же существовал (с “\_” вместо “-“). Источник получения остальных, кроме последнего, образцов – сточные воды (записи проигнорированы). Для последней записи координаты внесены в файл *FIXED\_Mamastrovirus\_1\_complete\_genome\_records\_orf.txt* вручную.

Новый запуск *split\_genome.py* показал, что фаста-файлы с последовательностями рамок теперь содержат **118 записей** – они и будут использоваться для дальнейшей работы. Записи содержатся в файлах с теми же названиями:

*Mamastrovirus\_1\_complete\_genome\_records\_1A.fasta*

*Mamastrovirus\_1\_complete\_genome\_records\_1B.fasta*

*Mamastrovirus\_1\_complete\_genome\_records\_2.fasta*