Санкт-Петербургский политехнический университет Петра Великого

Институт компьютерных наук и технологий

Высшая школа программной инженерии

Курсовая работа

Тема: «Получение списка белков, содержащихся в геноме SARS-CoV-2»

по дисциплине «Программирование биоинформатических приложений на суперкомпьютере»

Выполнил студент гр.

3540904/00202 Алейников П.И.

Преподаватель Маслаков А. П.

Задание

1. Выбрать любую сборку генома коронавируса [1].
2. Получить все возможные белки, содержащиеся в геноме [2], [3], [4].
3. Сравнить найденные белки с уже определенными для этого генома.
4. Подсчитать для каждого из белков молекулярную массу [5].
5. Попробовать получить модель 3-мерной структуры одного из найденных белков [6].

Выполнение

1. Исходные данные

В качестве исходной сборки генома коронавируса была выбрана сборка: *Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/TWN/CGMH-CGU-01 /2020, complete genome* [7].

Для сравнения потенциальных белков с уже определёнными для этого генома, информация об определенных белках [7] была записана в FASTA формате с указанием *protein_id* и *location*.

2. Реализация

Программа реализована на Java 14. Для каждой подзадачи создан отдельный класс и Unit тесты. Программа считывает геном и определенные для него белки из ресурсных FASTA файлов. Далее при помощи метода Open Reading Frames [4] вычисляются потенциальные белки и формируется файл ***finded_proteins.fasta*** со всеми найденными потенциальными белками и с указанием location (позиция в исходной последовательности ДНК) и mass (масса белка).

3. Результат сравнения найденных белков с уже определёнными

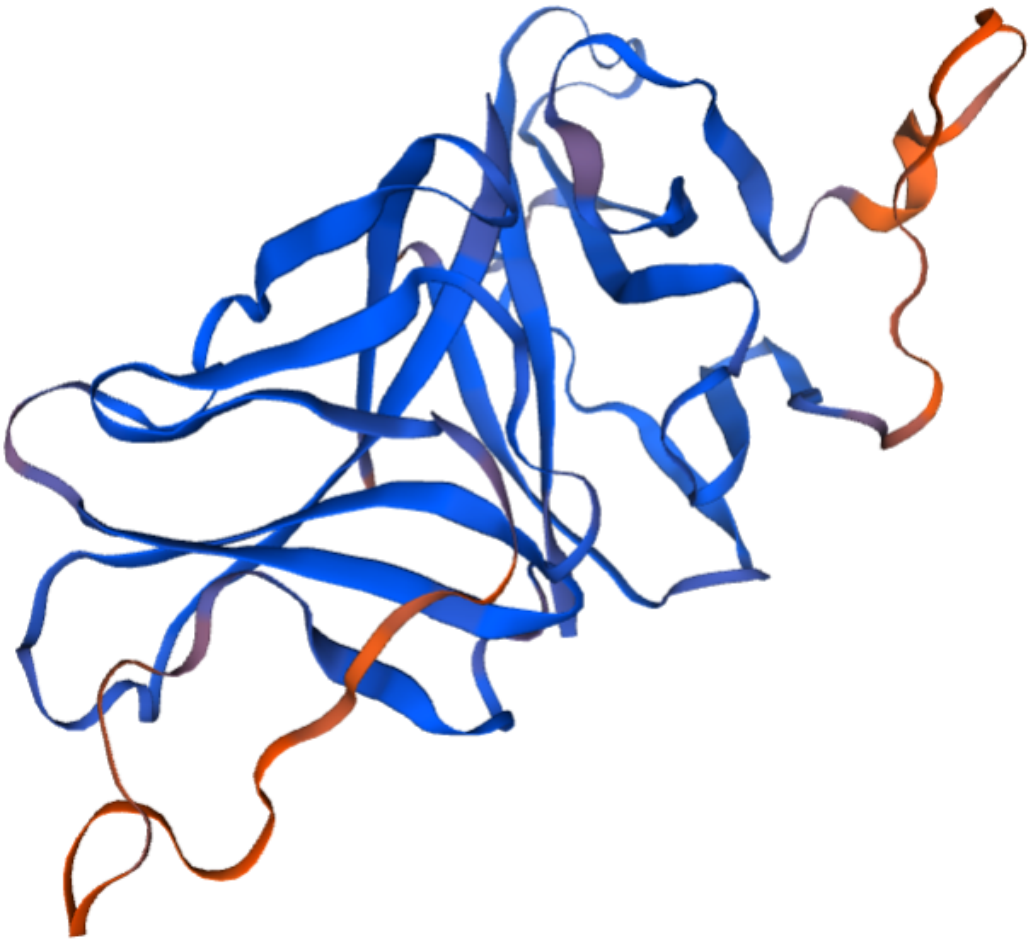
При сравнении потенциальных белков с уже определенными не удалось найти белок *QIK50416.1*. Программа нашла один потенциальный белок со стартовой позицией совпадающей с *QIK50416.1,* но он заканчивается позицией 13456, а не 21528 как у *QIK50416.1*.

```
[main] INFO Main - Read DNA with length: 29862
[main] INFO Main - Find 724 proteins in original DNA
[main] INFO Main - Read 10 defined proteins
[main] INFO task.ComparingProtein - Can't find protein QIK50416.1 with location 238..21528
```

4. Построение модели

Для построение модели 3-мерной структуры был выбран белок:

MKFLVFLGIITTVAAFHQECSLQSCTQHQPYVVDDPCPIHFYSK
WYIRVGARKSAPLIELCVDEAGSKSPIQYIDIGNYTVSCLPFTIN
CQEPKLGSLVVRCSFYEDFLEYHDVRVVLDFI



Список источников

1. Сборки генома SARS-CoV-2. URL: https://www.ncbi.nlm.nih.gov/genome/browse/#!/viruses/86693/
2. Задача перевода ДНК в РНК. URL: http://rosalind.info/problems/rna/
3. Задача транслирования РНК в белок. URL: http://rosalind.info/problems/prot/
4. Метод Open Reading Frames URL: http://rosalind.info/problems/orf/
5. Задача вычисления массы белка. URL: http://rosalind.info/problems/prtm/
6. Сервер моделирования структуры белков. URL: https://swissmodel.expasy.org
7. Выбранная сборка генома SARS-CoV-2. URL: https://www.ncbi.nlm.nih.gov/nuccore/MT192759.1