Санкт-Петербургский политехнический университет Петра Великого
Институт компьютерных наук и технологий
Высшая школа программной инженерии

Курсовая работа

Тема: «Поиск и анализ белков SARC-COV-2»

по дисциплине «Программирование биоинформатических приложений на суперкомпьютере»

Выполнил студент
Гр. 3540904/00201 Прохоров М. А.

Руководитель Маслаков А. П.

«21» мая 2021 г.

Санкт-Петербург
2021

Содержание

Введение

В качестве задания студенту было предложено:

- Выбрать один из имеющихся геномов коронавируса и получить все возможные белки, содержащиеся в геноме.
- Сравнить полученные последовательности с уже определенными для этого генома.
- Подсчитать для каждого из белков молекулярную массу
- Получить трехмерную модель структуры одного из найденных белков

1. Выбор генома

Среди имеющихся в источнике геномов был выбран один из самых последних от 9 марта 2020 года:

Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/CDC-CruiseA-6/2020, complete genome

GenBank: MT159722.2

FASTA Graphics

Go to:

```
LOCUS       MT159722               29882 bp    RNA     linear   VRL 29-JUL-2020
DEFINITION  Severe acute respiratory syndrome coronavirus 2 isolate
            SARS-CoV-2/human/USA/CDC-CruiseA-6/2020, complete genome.
ACCESSION   MT159722
VERSION     MT159722.2
KEYWORDS    .
SOURCE      Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
  ORGANISM  Severe acute respiratory syndrome coronavirus 2
            Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes;
            Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae;
            Betacoronavirus; Sarbecovirus.
REFERENCE   1  (bases 1 to 29882)
  AUTHORS   Uehara,A., Tao,Y., Paden,C.R., Queen,K., Zhang,J., Li,Y.,
            Keckler,M.S., Laufer Halpin,A.S., Wang,H., Padilla,J., Lee,J.,
            Elkins,C.A., Gerber,S.I. and Tong,S.
  TITLE     Cruise A sequences
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 29882)
  AUTHORS   Uehara,A., Tao,Y., Paden,C.R., Queen,K., Zhang,J., Li,Y.,
            Keckler,M.S., Laufer Halpin,A.S., Wang,H., Padilla,J., Lee,J.,
            Elkins,C.A., Gerber,S.I. and Tong,S.
  TITLE     Direct Submission
  JOURNAL   Submitted (07-MAR-2020) Division of Viral Diseases, Centers for
            Disease Control and Prevention, 1600 Clifton Rd NE, Atlanta, GA
            30033, USA
COMMENT     On Jun 3, 2020 this sequence version replaced MT159722.1.
```

Рис. 1. Страница генома

Геном в формате FASTA был скопирован в TXT-файл и помещен в папку datasets.

2. Поиск возможных белков в геноме

Кодирующие последовательности могут располагаться на любой из двух цепей ДНК (если ДНК не является одноцепочечной, как у многих вирусов), в любой из трех возможных фаз: по три, начиная со старт-кодона AUG, по три с +1 нуклеотида от AUG и так далее. При анализе просматриваются все шесть вариантов, поскольку при отсутствии дополнительной информации они являются равнозначными.

Для поиска белков в геноме была разработана следующая функция:

```
public static Set<String> openReadingFrames(String dataset) {
    String rnaString = transcribingDNAIntoRNA(dataset);
    Set<String> result = new HashSet<>();

    List<Integer> startIndexes = new ArrayList<>();
    int lastIndex = 0;
    while (lastIndex != -1) {
        lastIndex = rnaString.indexOf(START_CODON, lastIndex);
        if (lastIndex != -1) {
            startIndexes.add(lastIndex);
            lastIndex += 1;
        }
    }

    List<Integer> complementStartIndexes = new ArrayList<>();
    String reverseComplementString = complementingAStrandOfDNA(dataset);
    String reverseComplementRNAString =
transcribingDNAIntoRNA(reverseComplementString);
    lastIndex = 0;
    while (lastIndex != -1) {
        lastIndex = reverseComplementRNAString.indexOf(START_CODON, lastIndex);
        if (lastIndex != -1) {
            complementStartIndexes.add(lastIndex);
            lastIndex += 1;
        }
    }

    for (int index : startIndexes) {
        StringBuilder proteinString = new StringBuilder();
        for (int i = index; i < rnaString.length() - CODON_LENGTH; i +=
CODON_LENGTH) {
            String key = rnaString.substring(i, i + CODON_LENGTH);
            String str = Tasks.codonTable.get(key);
            if (str.equals("Stop")) {
                if (proteinString.length() >= MIN_PROTEIN_LENGTH) {
                    result.add(proteinString.toString());
                }
                break;
            }
            proteinString.append(str);
        }
    }

    for (int index : complementStartIndexes) {
        StringBuilder proteinString = new StringBuilder();
        for (int i = index; i < reverseComplementRNAString.length() - CODON_LENGTH;
i += CODON_LENGTH) {
```

```
            String key = reverseComplementRNAString.substring(i, i + CODON_LENGTH);
            String str = Tasks.codonTable.get(key);
            if (str.equals("Stop")) {
                if (proteinString.length() >= MIN_PROTEIN_LENGTH) {
                    result.add(proteinString.toString());
                }
                break;
            }
            proteinString.append(str);
        }
    }
    return result;
}
```

Функция анализирует геном, находит все последовательности белков и сохраняет их в текстовый файл.

3. Вычисление молекулярной массы белка

Подсчет молекулярной массы белка производился на основе таблицы моноизотопных масс:

```
private static Map<Character, Double> massTable = new HashMap<Character, Double>()
{{
    put('A', 71.03711);
    put('C', 103.00919);
    put('D', 115.02694);
    put('E', 129.04259);
    put('F', 147.06841);
    put('G', 57.02146);
    put('H', 137.05891);
    put('I', 113.08406);
    put('K', 128.09496);
    put('L', 113.08406);
    put('M', 131.04049);
    put('N', 114.04293);
    put('P', 97.05276);
    put('Q', 128.05858);
    put('R', 156.10111);
    put('S', 87.03203);
    put('T', 101.04768);
    put('V', 99.06841);
    put('W', 186.07931);
    put('Y', 163.06333);
}};

static String calculatingProteinMass(String dataset) {
    double result = 0.0;
    for (char ch : dataset.toCharArray()) {
        result += massTable.get(ch);
    }
    return String.format(Locale.US,"%.3f", result);
}
```

4. Результат работы функций

Функции были опробованы на имеющемся геноме. В результате запуска функции поиска белков получено около 1000 белков, что сильно превосходит количество представленных белков в источнике. Это связано с тем, что там представлены уникальные последовательности, больше некоторой длины. В моей реализации возможно перекрытия последовательностей, что также корректно, исходя из информации на сайте Rosalind.info. Также последовательности не ограничены по минимальной длине.

```
MWSTKMHKL
MRFYFSNISSIQISWTKLLSICH
MSLSEQLRKQIRSAAKKNNLPFKLTCATTRQVVNVVTTKIALKGGKIVNNWLKQLIKVTLVFLFVAAIFYLITPVHVMSKHTDFSSEIIGYKAIDGGVTRDIASTDTCFANKHADFDTWFSQRG
MFLARGIVFMCVEYCPIFFITGNTLQCIMLVYCFLGYFCTCYFGLFCLLNRYFRLTLGVYDYLVSTQEFRYMNSQGLLPPKNSIDAFKLNIKLLGVGGKPCIKVATVQSKMSDVKCTSVVLLSV
MELQQLH
MVQVTCGTTTLNGLWLDDVVYCPRHVICTSEDMLNPNYEDLLIRKSNHNFLVQAGNVQLRVIGHSMQNCVLKLKVDTANPKTPKYKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNFTIKGSF
MNLHLLMLLDNAQVLLSKVQ
MNVLTLVYKVYYGNALDQAISMWALIISVTSNYSGVVTTVMFLARGIVFMCVEYCPIFFITGNTLQCIMLVYCFLGYFCTCYFGLFCLLNRYFRLTLGVYDYLVSTQEFRYMNSQGLLPPKNSI
MSRGVMLLEKLLVPIYLYS
MVLGSLAATVRLQAGNATEVPANSTVLSFCAFAVDAAKAYKDYLASGGQPITNCVKMLCTHTGTGQAITVTPEANMDQESFGGASCCLYCRCHIDHPNPKGFCDLKGKYVQIPTTCANDPVGFT
MADGRFC
MHANYIFWRNTNPIQLSSYSLFDMSKFPLKLRGTAVMSLKEGQINDMILSLLSKGRLIIRENNRVVISSDVLVNN
MRKTHSQSLCQHLYFYRDTHKVCVVNCGHTYRQFCYHQ
MEPLQML
MVILLLLLKHLKNILLKPSHLLVPIKIGPILDNLHN
MHKL
MFSTKKTVTQQP
```

Рис. 2. Фрагмент результата поиска белков в геноме

Все имеющиеся в источнике белки были также получены в результате запуска функции.

Для этих белков были вычислены молекулярные массы и помещены в другой файл:

```
MWSTKMHKL
1142.573
MRFYFSNISSIQISWTKLLSICH
2755.398
MSLSEQLRKQIRSAAKKNNLPFKLTCATTRQVVNVVTTKIALKGGKIVNNWLKQLIKVTLVFLFVAAIFYLITPVHVMSKHTDFSSEIIGYKAIDGGVTRDIASTDTCFANKHADFDTWFSQRGGSYTNDK
187322.127
MFLARGIVFMCVEYCPIFFITGNTLQCIMLVYCFLGYFCTCYFGLFCLLNRYFRLTLGVYDYLVSTQEFRYMNSQGLLPPKNSIDAFKLNIKLLGVGGKPCIKVATVQSKMSDVKCTSVVLLSVLQQLRVE
71854.840
MELQQLH
879.427
MVQVTCGTTTLNGLWLDDVVYCPRHVICTSEDMLNPNYEDLLIRKSNHNFLVQAGNVQLRVIGHSMQNCVLKLKVDTANPKTPKYKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNFTIKGSFLNGSCGS
124588.239
MNLHLLMLLDNAQVLLSKVQ
2274.259
MNVLTLVYKVYYGNALDQAISMWALIISVTSNYSGVVTTVMFLARGIVFMCVEYCPIFFITGNTLQCIMLVYCFLGYFCTCYFGLFCLLNRYFRLTLGVYDYLVSTQEFRYMNSQGLLPPKNSIDAFKLNI
76233.103
MSRGVMLLEKLLVPIYLYS
2206.226
MVLGSLAATVRLQAGNATEVPANSTVLSFCAFAVDAAKAYKDYLASGGQPITNCVKMLCTHTGTGQAITVTPEANMDQESFGGASCCLYCRCHIDHPNPKGFCDLKGKYVQIPTTCANDPVGFTLKNTVCT
17409.168
MADGRFC
780.305
MHANYIFWRNTNPIQLSSYSLFDMSKFPLKLRGTAVMSLKEGQINDMILSLLSKGRLIIRENNRVVISSDVLVNN
8589.543
MRKTHSQSLCQHLYFYRDTHKVCVVNCGHTYRQFCYHQ
4698.161
MEPLQML
842.403
MVILLLLLKHLKNILLKPSHLLVPIKIGPILDNLHN
```

Рис. 3. Фрагмент результата подсчета массы белков

5. Построение трехмерной модели белка

Для построения модели использовался сервис swissmodel.expasy.org.

Был выбран один из полученных белков:

MLFTMLRKLDNDALNNIINNARDGCVPLNIIPLTTAAKLMVVIPDYNTYKNTCDGTTFTYASA
LWEIQQVVDADSKIVQLSEISMDNSPNLAWPLIVTALRANSAVKLQNNELSPVALRQMSCAA
GTTQTACTDDNALAYYNTTKGGRFVLALLSDLQDLKWARFPKSDGTGTIYTELEPPCRFVTD
TPKGPKVKYLYFIKGLNNLNRGMVLGSLAATVRLQAGNATEVPANSTVLSFCAFAVDAAKA
YKDYLASGGQPITNCVKMLCTHTGTGQAITVTPEANMDQESFGGASCCLYCRCHIDHPNPKG
FCDLKGKYVQIPTTCANDPVGFTLKNTVCTVCGMWKGYGCSCDQLREPMLQSADAQSFLNG
FAV

После применения функции моделирования было получено несколько моделей:

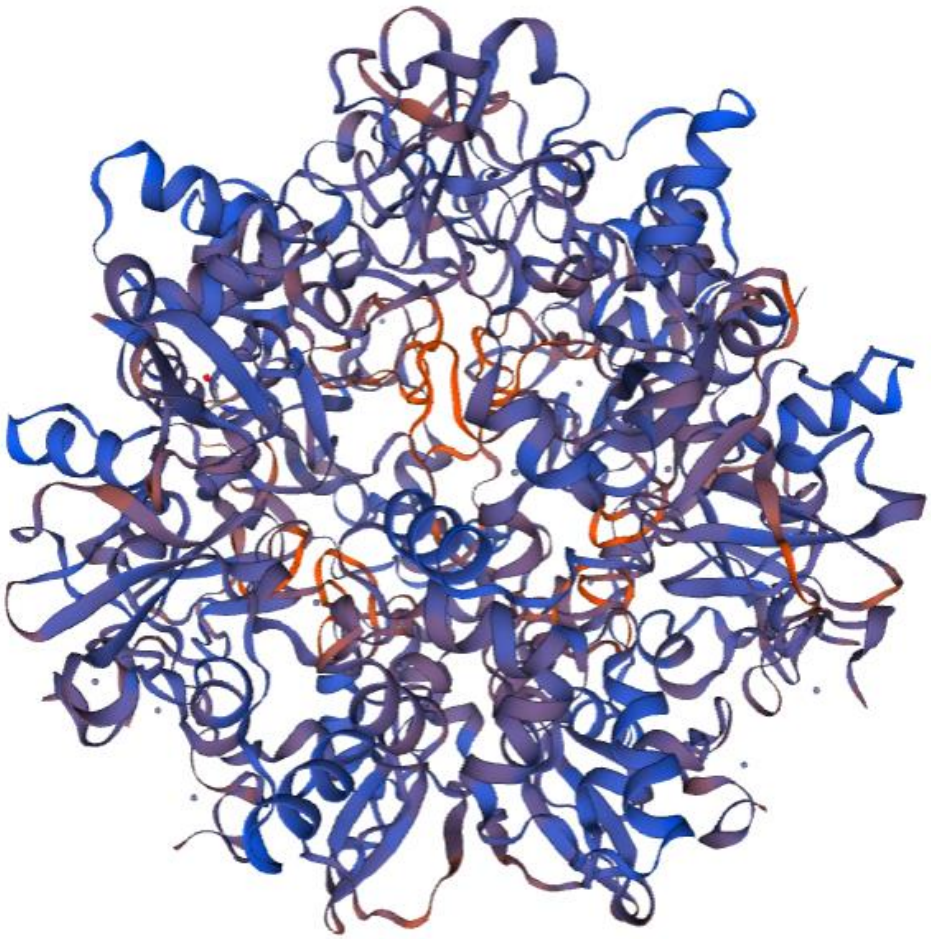

Рис. 4. Модель структуры белка

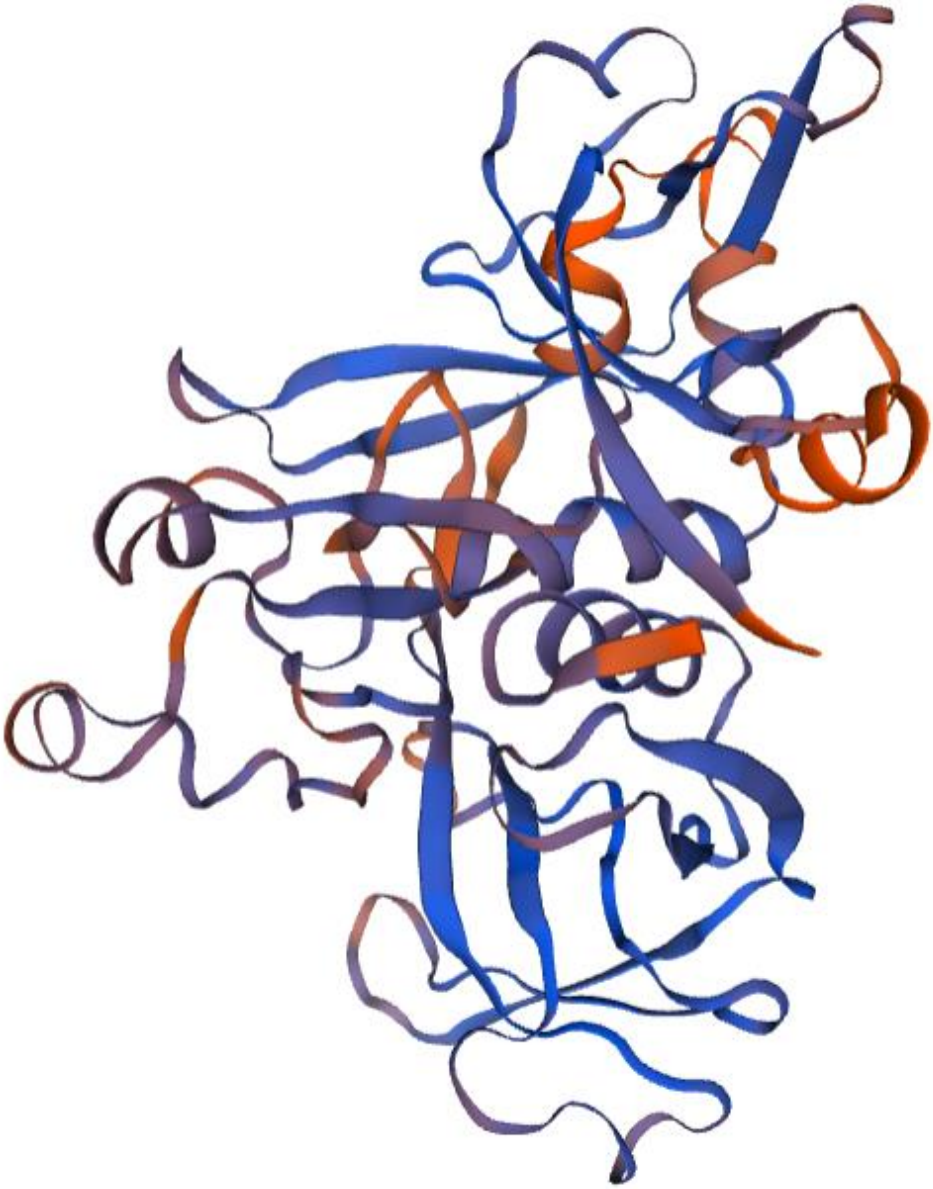

Рис. 5. Модель структуры белка

Заключение

В данной работе был следован геном SARS-COV-2, получены все возможные белки и подсчитаны их молекулярные массы.

Для одного из полученных белков была построена трехмерная модель.