

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ  
ВЫСШЕГО ОБРАЗОВАНИЯ  
«САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО»  
ВШ программной инженерии



**ПОЛИТЕХ**

Санкт-Петербургский  
политехнический университет  
Петра Великого

**Отчет**  
**по дисциплине «Программирование биоинформатических**  
**приложений на суперкомпьютере»**

Выполнил  
студент гр. № 3540202/00201

Е.В. Ковальчук

Руководитель

А.П. Маслаков

Санкт-Петербург  
2021 г.

## Оглавление

Постановка задачи .....	3
Выбор сборки генома коронавируса.....	4
Получение белков генома и расчет молекулярной массы.....	4
Построение модели белка .....	5
Заключение.....	7
Список использованных источников.....	8

## Постановка задачи

В рамках работы были поставлены следующие задачи:

1. Выбрать любую сборку генома коронавируса;
2. Получить все возможные белки, содержащиеся в геноме;
3. Подсчитать для каждого белка молекулярную массу;
4. Получить трехмерную модель одного из найденных белков.

Для решения основных задач необходимо ознакомиться и решить задачи в Rosalind [1]:

1. Расшифровка DNA в RNA;
2. Перевод RNA в белок;
3. ORF и другие.

## Выбор сборки генома коронавируса

На сайте [2] представлена таблица, была выбрана первая запись данной таблицы. После выбора сборки генома (рис. 1) необходимо перейти во вкладку fasta для получения самого генома.

The screenshot displays the NCBI COVID-19 Information page. At the top, there is a navigation bar with links for Public health information (CDC), Research information (NIH), SARS-CoV-2 data (NCBI), Prevention and treatment information (HHS), and Español. The main content area is titled "Severe acute respiratory syndrome coronavirus 2" and provides download options for FASTA format (genome, protein), GFF, GenBank, or tabular format. It also lists all 92 reference or representative genomes for the species. A "Summary" section shows the assembly level as "Complete Genome" and provides statistics: total length (Mb): 0.029903, protein count: 12, GC%: 38. A "Replicon Info" table is shown below the summary. The "Genome Region" section includes a "Go to nucleotide" link and a "FASTA" link. The bottom part of the image shows a genomic map with tracks for genes and proteins.

Type	Name	RefSeq	INSDC	Size (Kb)	GC%	Protein	Gene
Chr	-	NC_045512.2	MN908947.3	29.9	38.0	12	11

Рисунок 1.

## Получение белков генома и расчет молекулярной массы

На данном этапе необходимо найти белки, которые содержатся в геноме, и их молекулярную массу.

Каждому символу белка соответствует некоторое значение. Таким образом для получения молекулярной массы белка необходимо сложить значения символов, входящих в этот белок.

Большинство геномов, в том числе геном человека и геномы всех остальных клеточных форм жизни, построены из ДНК.

Был разработан алгоритм, который находит все возможные белки и для каждого из них находит молекулярную массу. Часть результата алгоритма представлена на рисунке 2, где первая строчка - это найденный белок, вторая - молекулярная масса.

1	MLSALIQYN	
2	1021.4902699999999	
3	MESLVP6FNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQHLKD6TCGLVEVEKGVLPQLEQPYVFIKRS DARTAPHGHVMVELVAELEGIQYGRSGETLGVLPVHVGEIPVAYRKVLLRKNGNKGAGGHSYGADL	
4	489654.0014000103	
5	MALVA	
6	485.26717999999994	
7	MCSSNVRLMELHLMVMLWLSW	
8	2560.22852	
9	MLELHLMVMLWLSW	
10	1782.9023300000001	
11	MVMLWLSW	
12	1046.50816	
13	MLWLSW	
14	816.3992599999999	
15	MVELVAELEGIQYGRSGETLGVLPVHVGEIPVAYRKVLLRKNGNKGAGGHSYGADLKSFDLGDDELGTDPYEDFQENWNTHKSSGVTRELMRELNGGAYTRYVDNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFIDT	
16	480360.1352900104	
17	MWAKYQWLTARFFVRTVIKELVAIVTAPI	
18	3578.9991600000008	
19	MKIFKKTGTLNIAVLPVNSCVSLTEGHTLMSITTSVALMATLLSALKTF	
20	5345.9242100000002	
21	MRELNGGAYTRYVDNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFIDTKRGVYCCREHEHEIAWYTERSEKSYELQTPFEIKLAKKFDTFNGECPNFVPLNSIIKTIQPRVEKKLDGFMGRISVYPVASPNECN	
22	470666.26992001	
23	MSITTSVALMATLLSALKTF	
24	2080.1311999999994	
25	MATLLSALKTF	
26	1176.65765	
27	MHFVRTGLY	
28	1205.60154	
29	MSMKLLGTRNVLKRAMNCRHLLKLNWQRNLTPSMGNVQILYFP	
30	5066.70057	
31	MKLLGTRNVLKRAMNCRHLLKLNWQRNLTPSMGNVQILYFP	
32	4848.62805	
33	MNCRHLLKLNWQRNLTPSMGNVQILYFP	
34	3367.72565	
35	MGNVQILYFP	
36	1162.5844900000002	
37	MSKFCISLKFHNQDYSTKG	
38	2215.05542	
39	HALWVEFDLSIQLRHQMNA TKCAFQLS	
40	7444.56104	

Рисунок 2.

## Построение модели белка

Модель белка можно построить с помощью сервиса [3].

Для того, чтобы построить модель, необходимо выбрать белок и вставить последовательность белка в окно ввода, затем построить модель. Модель будет доступна в правой части окна.

Для построения трехмерной модели был выбран следующий белок: MTTFLKYSKKRKSTSILLVTLNLMKRSPLFWHLFLLPQVLLWKL

Модель выбранного белка представлена на рисунке 3. С полученной моделью можно взаимодействовать.


**SWISS-MODEL**

Modelling
Repository
Tools
Documentation
Log in
Create Account

All Projects

**Untitled Project**
Created: today at 16:35

Summary
Templates 1
Models 2
Project Data

Model Results
Order by: GMQE



Model 01
Structure Assessment


<b>Oligo-State</b> Monomer	<b>Ligands</b> None	<b>GMQE</b> 0.23	<b>QMEAN</b> -0.91
-------------------------------	------------------------	---------------------	-----------------------

**Global Quality Estimate**

QMEAN	-0.91
C $\beta$	-0.97
All Atom	-0.68
solvation	-1.09
torsion	-0.29

**Local Quality Estimate**


**Comparison**


<b>Template</b> 4bhh.1.E	<b>Seq Identity</b> 22.58%	<b>Coverage</b> 	<b>Description</b> NUCLEOPROTEIN Crystal structure of tetramer of La Crosse virus nucleoprotein in complex with ssRNA
-----------------------------	-------------------------------	---	---

**Model-Template Alignment**

```

Model_01  MNKSLG  NVLKKANNCKRLKLNKQNNLTFSKNN  LFFP  43
4bhh.1.E  ----- RLELTATNATATLT  PEP  P  G  E  V  -----  71

```



Model 02
Structure Assessment

<b>Oligo-State</b> Monomer	<b>Ligands</b> None	<b>GMQE</b> 0.19	<b>QMEAN</b> -2.67
-------------------------------	------------------------	---------------------	-----------------------

**Global Quality Estimate**

**Local Quality Estimate**

**Comparison**

<b>Template</b> 1o6e.1.B	<b>Seq Identity</b> 18.75%	<b>Coverage</b> 	<b>Description</b> CAPSID PROTEIN P40 Epstein-Barr virus protease
-----------------------------	-------------------------------	---	---

**Model-Template Alignment**



Cartoon

Рисунок 3.

6

## **Заключение**

В ходе данной работы из генома коронавируса были получены все возможные белки и рассчитаны их молекулярные массы. Для одного из найденных белков была построена трехмерная модель.

Также были решены соответствующие задачи в Rosalind.

### **Список использованных источников**

1. Rosalind [Электронный ресурс]. - URL: <http://rosalind.info/problems/list-view/> (Дата обращения: 15.05.2021).
2. Severe acute respiratory syndrome coronavirus 2 [Электронный ресурс]. - URL: <https://www.ncbi.nlm.nih.gov/genome/browse/#!/viruses/86693/> (Дата обращения: 17.05.2021).
3. Swiss-model [Электронный ресурс]. - URL: <https://swissmodel.expasy.org/> (Дата обращения: 19.05.2021).