

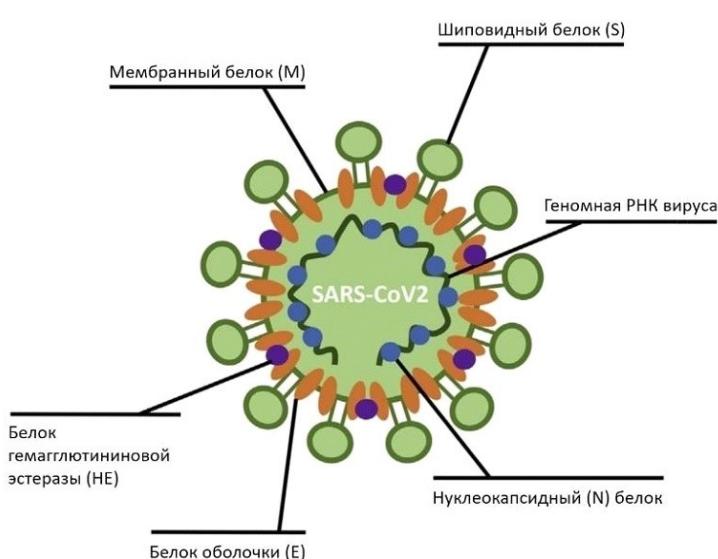
Решение задания хакатона от команды «Профессора 20»

На первом этапе работы мы ознакомились с объектом дальнейшего исследования (вирусом SARS-CoV-2). Мы выяснили, какие макромолекулы SARS-CoV-2 могут выступать в роли доноров эпитопов для создания вакцины. Среди этих молекул, а именно белков, оказались N-, S-, E- и M-белки. Каждую из этих структур мы рассмотрели в отдельности.

- Изучив строение N-белка, мы выяснили, что он не может выступать в роли донора эпитопа, так как является нуклеопротеином (служит для компактизации РНК вируса) и расположен непосредственно внутри вируса (см. рис.)
- Мембранный белок (M-белок), играет ключевую роль в сборке вируса, поскольку привлекает все остальные элементы к месту сборки.
- Белок оболочки Е представлен малым числом копий. Он формирует пентамерные ионные каналы, которые нарушают строение клеточных мембран клетки-хозяина в ходе размножения вируса.
- S-белку свойственно пептидное средство со многими белками человека, в частности и интерлейкинам, интегринам, различным рецепторам и компонентам комплемента - белкам иммунной системы, локализующимися на поверхности клеток или циркулирующими в крови. Присутствие таких гомологичных последовательностей в вакцине против Covid-19 повышает риск развития аутоиммунных заболеваний и гетерологичного иммунитета. Поэтому необходимо выбрать для вакцины

такие последовательности белков SARS-CoV-2, которые минимально гомологичны белкам человека и будут соответственно распознаваться иммунной системой как «чужие» с образованием к ним антител.

Исходя из вышеперечисленных данных основным объектом на роль доноров эпитопов для нас стали E- и M-белки SARS-CoV-2. Последовательности же S-белков будут рассматриваться нами в



последнюю очередь, при этом последовательности, кодирующие S-белки будут подвергаться предварительной проверке на совпадение с усреднённым геномом человека в базе данных NCBI.

В процессе поиска решения проблемы мы опирались на научные исследования и статьи, ссылки и названия приведены ниже:

1. <https://visual-science.com/ru/projects/sars-cov-2/3d-model/> (3D-модель CoV-2)
2. https://www.epidemvac.ru/jour/article/view/977?locale=ru_RU
3. Статья Татьяны Зайчук, Юрия Нечипуренко и др. "Обзор технологий для создания вакцин против бетакоронавирусов и вирус Сендей как возможный вакцинный вектор"

Для того чтобы получить эпитопы антигенов выбранных белков мы использовали базу данных IEDB (The Immune Epitope Database).[1]

The screenshot shows the IEDB homepage with a search results table. The search filters are set to 'Organism: Severe acute respiratory syndrome-related coronavirus (ID:694009)'. The results table has columns for Details, Epitope, Antigen, Organism, # References, and # Assays. The first few rows show epitopes for Spike glycoprotein, Replicase polyprotein 1ab, and other proteins. A pending filters sidebar on the left includes sections for Epitope (Any Epitopes selected), Antigen (Organism: Severe acute respiratory syndrome selected), and various search parameters like 3D structure availability and amino acid modification.

Details	Epitope	Antigen	Organism	# References	# Assays
16156	FIAGLIAIV	Spike glycoprotein	Severe acute respiratory syndrome-related coronavirus (Human coronavirus (strain SARS))	12	29
33572	KSYEHQTPF	Replicase polyprotein 1ab	Severe acute respiratory syndrome-related coronavirus (Human coronavirus (strain SARS))	10	29
2998	ALWEIQQVV	Replicase polyprotein 1ab	Severe acute respiratory syndrome-related coronavirus (Human coronavirus (strain SARS))	8	40
21347	GMSRIGMEV	Other Severe acute respiratory syndrome-related coronavirus	Severe acute respiratory syndrome-related coronavirus	8	19

С неё мы скачали эпитопы всех трёх белков:

Е-белок — 53 эпитопа,

М-белок — 171 эпитоп,

S-белок — 1095 эпитопов.

[Папка №1]

Так как вакцина будет тем лучше, чем больше механизмов адаптивного иммунитета она вызывает, то мы решили, что уместим в ограничение 900 аминокислот сначала все эпитопы Е-белка и М-белка, а оставшееся место выделим для эпитопов S-белка.

Чтобы не навредить организму человека, то сначала мы должны сравнить полученные эпитопы с эпитопами антигенов клеток человека на предмет схожести. Мы должны убрать из нашего реестра все эпитопы, которые имеют схожесть с человеческими, потому что если их оставить, то у

человека можно спровоцировать аутоиммунное заболевание, так как его организм начнёт синтезировать антитела на собственные клетки.

Для этой проверки написали код на python, который удаляет из реестра все такие эпитопы, используя для сравнения базу данных белков человека из [Uniprot](#).[2]

The screenshot shows the UniProt Proteomes interface for Homo sapiens. The main header says "Proteomes - Homo sapiens (Human)". On the left, there's a sidebar with "Overview", "Components", and "Publications" checked under "None". Below that is a "Map to" section with links to UniProtKB (75,777), Swiss-Prot (Reviewed 20,379), and TrEMBL (Unreviewed 55,398). The main content area is titled "Overview" and contains a table with the following data:

Status	R Reference proteome
Proteins ⁱ	75,777
Gene count ⁱ	20,609 - Download one protein sequence per gene (FASTA)
Proteome ID ⁱ	UP000005640
Taxonomy	9606 - Homo sapiens
Last modified	September 3, 2020
Genome assembly and annotation ⁱ	GCA_000001405.27 from Ensembl full
Busco ⁱ	C:99.6%[S:39.3%,D:60.3%],F:0%,M:0.4%,n:13780 primates_odb10
Completeness ⁱ	Outlier

Below the table, a note states: "Homo sapiens (Homo sapiens sapiens) or modern humans are the only living species of the evolutionary branch of great apes known as hominids. Divergence of early humans from chimpanzees and gorillas is estimated to have occurred between 4 and 8 million years ago. The genus Homo (Homo habilis) appeared in Africa around 2.2 million years ago and shows the first signs of stone tool usage. The exact lineage of Homo species is: H".

В этой базе данных хранятся пептидные последовательности белков в их вариации, в которой их чаще всего можно встретить, это позволяет нам добиться того, что вакцина будет подходить большему количеству людей. Таким образом мы избавились от двух эпитопов S-белка (VSATK и SATKL)

[Папка №2]

Для того чтобы скомпоновать все эпитопы в одну последовательность мы написали код на python, который с помощью выравнивания находит одинаковые части разных эпитопов и накладывает их друг на друга до тех пор, пока не получит одну последовательность. Программа первой версии (compprof1) справилась с задачей, соединив все эпитопы Е-белка.

Позже выяснилось, что такой код подойдёт только для работы с небольшим количеством небольших эпитопов, так программа на смогла справиться с эпитопами М-белка. Была написана вторая версия (compprof2), которая справилась с первой задачей в 6 раз быстрее, и могла, хоть и через продолжительно время, справиться с большим количеством эпитопов. В отличии от первой версии, которая выравнивала большие последовательности друга на друга, вторая версия сравнивала на схожесть только их концы, что позволяло ускорить процесс. Из-за того что эпитопы М-белка (как и S-белка) частично различались между собой наша

программа не смогла должным образом их обработать. В связи с этим эпитоны М-белка были наполовину соединены программой и наполовину вручную, а соединенный набор эпитопов S-белка, гомологичный нашему, мы нашли с помощью инструмента BLAST (Отличия были в несколько аминокислот) [3]

The screenshot shows the NCBI Protein page for entry Q0J87214.1. The page header includes links for NCBI Resources, How To, and a search bar. A prominent red banner at the top provides COVID-19 information from CDC and NIH, along with a link to NCBI's SARS-CoV-2 literature. Below the banner, the protein name is listed as "surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]". The "GenPept" tab is selected. On the right, there are buttons for "Send to:" (with "Change region shown"), "Customize view", and a sidebar with options like "Analyze this sequence", "Run BLAST", and "NCBI Virus". The main content area displays detailed sequence information, including Locus (Q0J87214), Definition (surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]), Accession (Q0J87214), Version (Q0J87214.1), DBLINK (BioProject: PRJNA639066, BioSample: SAMN16352193, Sequence Read Archive: SRR12765287), DBSOURCE (accession MW093487.1), KEYWORDS (.), and SOURCE (Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)).

В финальную пептидную последовательность мы добавили не все эпитопы S-белка, так как мы решили, что сначала добавим эпитопы всех остальных белков. В итоге была взята последовательность длиной 588 аминокислот и, составляющая в себе 490 эпитопов S-белка и не содержащая в себе тех двух исключенных эпитопа. Все эпитопы, которые мы в итоге использовали находятся в трёх файлах в папке «Использованные эпитопы» в Папке №3 - S-proteins epitope (490 эпитопа) ; E-proteins epitope (53 эпитопа) ; M-proteins epitope (171 эпитоп).

[Папка №3]

Итоговый пептид длиной 895 аминокислот состоит из 53 эпитопов Е-белка, 171 эпитопа М-белка и 490 эпитопов S-белка.

- [1] - http://www.iedb.org/result_v3.php?cookie_id=055064&active_tab=Epitopes
- [2] - <https://www.uniprot.org/uniprot/?query=&fil=organism%3A%22Homo+sapiens+%28Human%29+%5B9606%5D%22>
- [3] - [https://www.ncbi.nlm.nih.gov/protein/Q0J87214.1?report=genbank&log\\$=protalign&blast_rank=1&RID=WU7TFNVX013](https://www.ncbi.nlm.nih.gov/protein/Q0J87214.1?report=genbank&log$=protalign&blast_rank=1&RID=WU7TFNVX013)