Фаза 2: База податоци
16.05.2021

ВОВЕД ВО БИОИНФОРМАТИКА

База на податоци и претпроцесирање

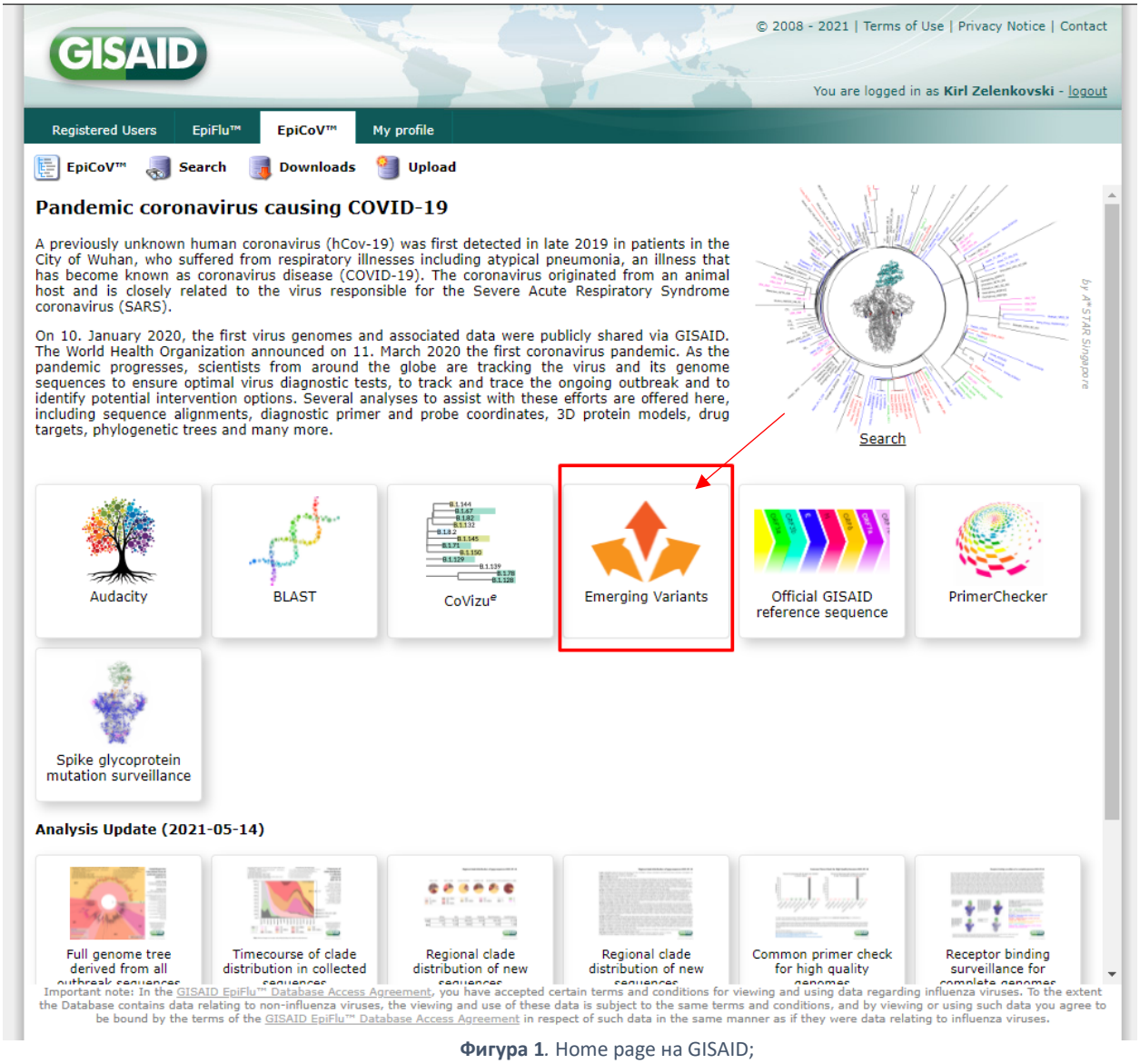
Предвидување на варијатни на глико-протеинот на SARS-Cov2 вирусот

Професор*:* д-р Невена Ацковска
Студент*:* Кирил Зеленковски, 161141

База податоци

Базата на податоци како што образложив во претходната фаза ќе ја создавам користејќи податоци од GISAID [1]. Се одлучив за оваа страна како извор на податоците врз кои ќе учат моделите бидејќи повеќето dashboards и трудови што кружат околку интернетот, дали од престижни универзитети или фирми - повеќето влечат податоци од GISAID (не ги цитирав бидејќи навистина има премногу). GISAID е овозможено од грантови од Германија, Америка и Сингапур.

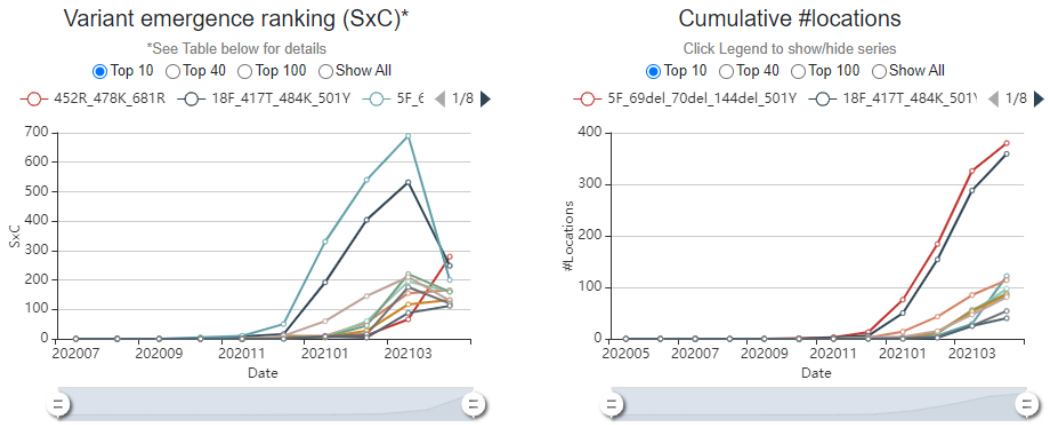
На главната веб-страна на GISAID има неколку одели (sections) и сите можат да бидат корисни на некој начин за базата. Јас главно се фокусирав на оделот *“Emerging Variants”.*



Фигура 1. Home page на GISAID;
Одел Emerging variants

Во овој одел од страница помага да се следат новите варијанти на hCoV-19, кои би можеле да станат релевантни заради знаци на зголемено ширење (проценето со промена на бројот на локации) во комбинација со потенцијални ефекти врз врзувањето на рецепторот или антителата, коментирани во CoVsurver. Во моментов, 124 промени на аминокиселините и бришења во глико-протеинот (Spike) кои се јавуваат во најмалку 10 различни географски локации и беа идентификувани во студиите за да предизвикаат бегство на антитела, зголемување на врзувањето за ACE2 или зголемување на изразот и стабилноста на протеинот Spike се сметаат како дел од формирање на комбинации на потенцијалните варијанти што треба да се следат.

Промените со наставка „X“ претставуваат недефинирани основи на соодветните страници, кои исто така можат да вклучуваат страници за бришење Spike. Конечно, варијантите за секој месец (според датумот на собирање) се рангирани според SxC, што е производ на промената на бројот на локации (во споредба со претходните месеци; слично на ширењето S) и бројот на релевантни промени на аминокиселини со потенцијал ефект што придонесува кон комбинации (C).



Фигура 2. Ранкинг (лево), Бр. локации (десно)

На самата страница има мета-податоци за секој месец почнувајќи од **Јули 2020 – Април 2021**. Податоците за секој месец се организирани во табели за секој месец посебно каде има поголем број на колони (кои се карактеристики подаоцна за учење, features) и една таква табела пример за месец Април 2021 изгледа вака:

	A	B	C	D	E	F	G	H	I	J
1	Variant	#Genomes	#Top Location	#Top Clade	#Top Lineage	Co-occurring Changes	#Co-occurring Changes	Δ#Loc(S)	#aachanges(C)	(SxC)
2	452R_478K_6	1889	779 England	1887 G	1800 B.1.617.2	Spike_T19R, Spike_E156G, Sp	16	93	3	279
3	18F_417T_48	10075	1205 Sao Paulo	9088 GR	10034 P.1	Spike_D138Y, Spike_R190S, S	20	63	4	252
4	5F_69del_70c	10073	2573 England	10011 GRY	10067 B.1.1.7	Spike_A570D, Spike_Y145del	20	44	5	220
5	69del_70del_	603	180 Germany	599 GRY	603 B.1.1.7	Spike_A570D, Spike_Y145del	21	33	5	165
6	69del_70del_	702	213 Tyrol	696 GRY	699 B.1.1.7	Spike_A570D, Spike_Y145del	20	33	5	165
7	69del_70del_	655	265 England	639 GRY	655 B.1.1.7	Spike_A570D, Spike_F490S, S	22	32	5	160
8	69del_70del_	143	24 Utenos apskri	143 G	143 B.1.620	Spike_H245Y, Spike_P681H,	15	17	8	136
9	452R_484Q_6	1756	757 Maharashtra	1752 G	1649 B.1.617.1	Spike_Q1071H, Spike_D614(	11	44	3	132
10	18F_69del_7(	3161	2143 England	3146 GRY	3160 B.1.1.7	Spike_A570D, Spike_Y145del	21	26	5	130
11	69del_70del_	220	58 England	219 GR	210 C.36	Spike_S12F, Spike_Q677H, S	23	28	4	112
12	18F_242del_2	176	40 Luxembourg	176 GH	176 B.1.351	Spike_D215G, Spike_D80A, S	14	14	8	112

Фигура 3. Пример од првите записи за месец Април 21’

Како што можеме да видиме од Фигура 3, имаме вкупно 10 колони. Подолу е табелата за колку се вкупно записи и од кој месец колку варијанти има.

Секоја една колоните претставува потенцијално информација за учење на моделите, и нивното значење е следно:

- **Variant (**стринг**):** Име по кое се води варијантата (името се задава врз основа на каде настанала мутација на глико-протеинот но референтно во главниот SARS-Cov2 [2] геном)
- **Genomes (**бројка**):** Вкупен број на геноми што се секвенционирани со овие мутации
- **Top Location (**бројка + стринг**):** Овој податок се состои од два дела:
    - Најголема бројка на вакви секвенционирани геноми
    - Држава од каде потекнуваат
- **Top Clade (**бројка + стринг**):** Clade претставува група сродни организми кои потекнуваат од заеднички предок. Повторно овој податок се состои од два дела:
    - Најголема бројка на секвенционирани геноми од овој Clade
    - Име на Clade
- **Top Lineage (**бројка + стринг**):** Lineage доаѓа како лоза (потекло) на секоја нова варијанта. Овој податок се состои од два дела:
    - Најголема бројка на секвенционирани геноми од ова лоза
    - Име на лоза
- **Co-occurring changes (**стринг**):** Овој податок е најкорисен од сите, претставува како листа од мутации на самата варијанта. Секоја мутација е преставена со конкатениран стрин:
    - Spike = Во кој ген настанала промената во геномот на вирусот (овој е ист за сите)
    - Протеин во главна секвенца = Како на примерот горе, првиот ред има T
    - Позиција во глико-протеинска секвенца = Како на примерот горе, првиот ред има 19, тоа е 19-та позиција во првобитниот глико-протеин
    - Мутиран протеин во новата секвенца = Како на примерот горе, првиот ред има R, тоа значи дека ако има T19R тоа значи дека протеинот T на позиција 19 е заменет со R во оваа варијанта
- **#Co-occurring changes (**бројка**):** Број на промени во оваа варијанта. Ова бројче е корисно за да се даде overall score на самата мутација за да се одреди нејзината важност.
- **#Loc(S) (**бројка**):** Промена во бројот на локации споредено во период од 3 месеци на оваа варијанта
- **#aachanges(C) (**бројка**):** Ова аа е кратенка за Amino-Acid Substitution и овие промени се доста важни за добивање на самиот score на оваа варијанта
- **(SxC) (**бројка**):** Производ на промената на бројот на локации (Loc или на кратко S) и бројот на релевантни промени на аминокиселини со потенцијал ефект (aachanges или на кратко C)

Вакви табели има за секој месец посебно од Јули 20’ – Април 21’ има 10 месеци и за секој месец има различен вкупен број на варијанти. Секако првите месеци со целото избувнување на вирусот бројот на евидентирани [i] е помал, но веќе април месец бројот на вакви мутации изнесува 765 варијанти.

Број на варијанти евидентирани по месец:

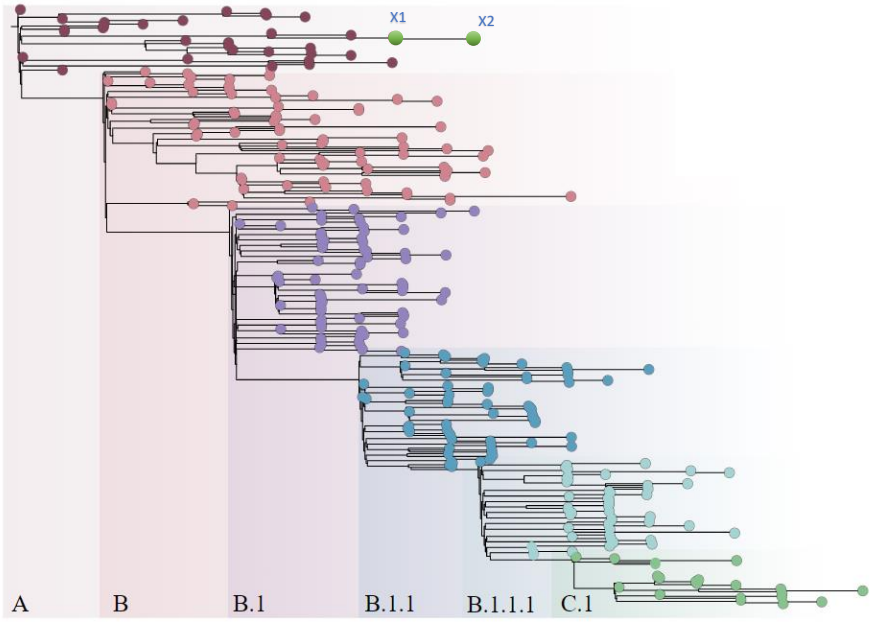
Јул.	Авг.	Септ.	Окт.	Ноем.	Дек.	Јан.	Феб.	Мар.	Апр.
50	75	82	117	245	249	425	592	767*	764

Овие се симнуваат како посебни датотеки и после со помош на Python сркипти ги конкатенирав во една голема база од варијанти каде ја додадов само колоната на кој месец припапаѓаат и вкупниот број на варијанти резултираше со:

50 + 75 + 82 + 117 + 245 + 249 + 425 + 592 + 766* (едно се повторуваше) + 764 = 3 365 варијанти

Податоците беа доста средени бидејќи е доста официјална страната. Единствено нешто што е сум во фаза да го завршувам е поврзување врз основа на lineage. Бидејќи повеќето варијанти се од сличини лози но месец од месец не се поврзани, додека го правев ова почнав да додавам и еден вид на similarity score меѓу некои од различни месеци што се премногу слични. Ова го правев со помош на порамнување. Но, бидејќи премногу записи се и не сакам да правам грешки ова сеуште го правам.

Ги поврзувам главно врз основа на вакви филогенетски дрва кои доаѓаат од Outbreak.info [3]:



Фигура 4. Хиерархиски дрва (PANGO)

Откако ќе завршам евидентирање на овие информации идејата е да се најдат претходниците на секоја нова мутација бидејќи не се добро евидентирани.

Пример точките X1 и X2 ќе бидат дел од еден ред каде X1 е вид на влез (претходник на варијанта) и после излезот е после сите мутации X2 (следбеник на варијанта).

Референци

[1] GISAID официјална веб-страна, проследна на: https://www.gisaid.org/

[2] NCBI Reference Sequence: NC_045512.2, проследно на:
https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2/

[3] Outbreak.info: информации за важност на варијанти, проследно на:
https://outbreak.info/situation-reports/caveats

[i] Под евидентирани ја посочувам можноста за поголема стапка на мутација на вирусот на почетокот, поради целата ситуација оваа слика можеби не е најреална.