Analiza mutacija u genomu SARS-CoV-2

Andrijana Ivković
Broj indeksa: 115/2019

Ivana Nestorović
Broj indeksa: 130/2019

Januar 2024

Profesor: Nenad Mitić
Projekat na kursu Istraživanje Podataka 2

Sadržaj

1 Uvod

1.1 Uopšteno

SARS-CoV-2 je glavni uzročnik bolesti COVID-19. Od kad je otkriven postao je glavna pretnja javnom zdravlju i odgovoran je za pandemiju koja je zahvatila čitav svet. Ovaj virus pripada porodici koronavirusa i ima veoma složenu genetiku i biologiju.

1.2 Genom i Struktura

Genom sadrži više od 30.000 nukleotida (kodiraju različite proteine neophodne za replikaciju i infekciju domaćina). Njegovu strukturu čine ORF1ab, spike (S) protein, membranski (M) protein i nukleokapsidni (N) protein. To je struktura koja omogućava virusu da prepozna ćelije domaćina, veže se i inficira ih. [8][6]

1.3 Patogeneza i Klinička Slika

Virus u organizam ulazi preko respiratornog trakta, posebno kapljičnim putem. Spike (S) protein virusa se vezuje za ACE2 receptore kod domaćina i na taj način omogućava virusu ulazak u ćeliju. Nakon vezivanja, virus oslobađa genetski materijal u ćelije domaćina. Ovaj virus ima sposobnost brze replikacije, pogotovu u respiratornom traktu. Infekcija aktivira imunski sistem. Negativni odgovor imunskog sistema može izazvati hiperinflamatorne reakcije, što može dovesti do ozbiljnih zdravstvenih problema. Virus može dovesti do sistemske upale, putujući kroz krvotok domaćina. Na taj način se povećava rizik od komplikacija, kao što su npr tromboza.. Klinička slika može imati više oblika:
1. Blaga infekcija, kod većine pacijenata, simptomi su: groznica, kašalj, umor, gubitak ukusa i mirisa.
2. Teška infekcija, simptomi se pogoršavaju i dolazi do poteškoća u disanju, upale pluća
3. Komplikacije: plućna embolija, rizik od srčanih komplikacija, oštećenja srca, bubrega..
4. Sindrom hiperinflamacije ili poznatiji kao čytokine storm"
5. Dugotrajni simptomi izazvani COVID-om mogu biti umor, glavobolja, gubitak daha, mentalne smetnje itd.
[2][4][1][5]

1.4 Epidemiologija

SARS-CoV-2 se prenosi kapljičnim putem, uglavnom kada osoba kašlje, kija ili govori. Moguća je i infekcija dodirivanjem zaraženih površina i osoba.
Inkubacioni period, od izlaganja virusu do pojave prvih simptoma, može biti od nekoliko dana do 2 nedelje. Većina zaraženih ima blage simptome. Neki ljudi mogu biti samo asimptomatski nosioci, što otežava prepoznavanje virusa i njegovu dalju kontrolu širenja.
Starije osobe i osobe sa postojećim zdravstvenim problemima (srce, dijabetes, hronične bolesti disajnih puteva, itd..) imaju veći rizik od zaraze i težih komplikacija.

Da bi kontrolisali širenje, stručnjaci su dali preporuke za određene mere:

Fizičko distanciranje
Nošenje maski
Higijena ruku
Testiranje i obaveštavanje osoba sa kojima ste bili u kontaktu Imunizacija
Informisanje javnosti

[7][3]

1.5 Cilj istraživanja

Radi boljeg razumevanja evolucije SARS-CoV-2 na lokalnom nivou, sprovedeno je istraživanje genoma virusa iz uzoraka prikupljenih na teritoriji Srbije u periodu od 2020. do 2023. godine. Cilj ovog istraživanja je bilo analiziranje genetičkih varijacija, identifikacija mutacija i utvrđivanje učestalosti promena na nivou nukleotida i aminokiselina.

U prvom delu istraživanja, izvršeno je poravnanje nukleotidnih sekvenci prema referentnom izolatu SARS-CoV-2 (NCBI identifikacija NC 045512.2). Zatim su određene moguće granice proteina. To nam je omogućilo detaljan uvid u genetsku raznolikost virusa prisutnog u lokalnoj populaciji.
U drugom delu, fokusirali smo se na identifikaciju najčešće i najređe mutiranih mesta u genomu SARS-CoV-2. Kako bismo analizirali učestalost promena, primenjivali smo različite strategije, uključujući opcije koje uključuju i isključuju neidentifikovane nukleotide (N). Ova analiza nam pruža ključne informacije o ucestalosti mutacija u lokalnim sojevima virusa.
U daljim koracima, identifikovali smo regione od 5 nukleotida sa najmanjom i najvećom stopom mutiranosti. Dublja analiza nam je omogućila uvid u određene delove genoma. To može da nam pomogne da prepoznamo očuvane nukleotidne sekvence. Nukleotidne sekvence su ključne za funkcionalnost virusa.
Konačno, za proteine čije su pozicije odredjene nakon poravnanja nukleotidnih sekvenci, izvršena je translacija u aminokiselinsku sekvencu. Procenat mutacija na aminokiselinskom nivou analiziran je i uporedjen sa rezultatima mutacija na nukleotidnom nivou, pružajući dodatnu perspektivu na evoluciju virusa. Cilj istraživanja je između ostalog, razumevanje genetičkih karakteristika lokalnih sojeva SARS-CoV-2.

2 Podaci

2.1 Izvor podataka

Podaci koje koristimo su dobijeni iz baze podataka **GISAID** (Global Initiative on Sharing All Influenza Data). Ovi podaci obuhvataju niz nukleotidnih sekvenci genoma SARS-CoV-2 virusa, prikupljenih iz uzoraka na teritoriji Srbije tokom perioda od 2020. do 2023. godine. Referentna sekvenca SARS-CoV-2 virusa (NCBI identifikacija NC 045512.2) takođe je uključena u analizu.

2.2 Struktura podataka

Podaci su organizovani u obliku datoteke u Fasta formatu koja sadrži poravnate nukleotidne sekvence. Ova poravnanja su generisana pomoću alata **MAFFT**. Svaka sekvenca u datoteci odgovara jednom uzorku i uključuje informacije o genomu SARS-CoV-2.
Pre poravnanja, izvršeno je spajanje referentnog izolata sa sekvencama uzoraka iz Srbije[1]. Referentni izolat je učitan iz datoteke NC_045512.2.fasta, dok su sekvence iz Srbije učitane iz datoteke 1699819380462.sequences.fasta. Nakon toga, sve sekvence su spojene i sačuvane u novoj datoteci pod nazivom spojene_sekvence.fasta.

3 Istraživanje

3.1 Poravnanje nukleotidnih sekvenci

Nakon prikupljanja nukleotidnih sekvenci i spajanja, sledeći korak u analizi genoma SARSCoV-2 virusa bio je njihovo poravnanje radi precizne analize evolutivnih odnosa i identifikacije mutacija. Za ovu svrhu, koristili smo MAFFT (Multiple Alignment using Fast Fourier Transform) alat, koji omogućava efikasno i tačno poravnanje nukleotidnih sekvenci.
[2] MAFFT koristi heuristički pristup za brzo poravnanje sekvenci, čineći ga efikasnim izborom za analizu genomske varijabilnosti velikog broja uzoraka. Ovaj alat takođe može efikasno rukovati sa sekvencama različitih dužina i detektovati konzervirane regione, što je posebno važno u analizi virusnih genoma koji mogu imati različite dužine i strukture.

Nakon poravnanja, imali smo precizan okvir koji nam je omogućio dalju analizu genomske varijabilnosti, identifikaciju mutacija, i određivanje granica proteina na nukleotidnom nivou.
Koristili smo JalView alat za vizuelizaciju rezultata. JalView nam je omogućio precizno praćenje konzerviranih regiona, identifikaciju mutacija i vizuelno poredjenje više sekvenci radi boljeg razumevanja genomske varijabilnosti.

[1] Program za spajanje: Spajanje.ipynb
[2] Poravnanje sekvenci iz terminala. 4.3.1

3.2 Identifikacija granica proteina

Nakon poravnanja nukleotidnih sekvenci, sledeći korak bio je identifikacija granica proteina u genomu SARS-CoV-2 virusa [3]. Za ovo smo koristili unapred definisane pozicije proteina, a informacije o granicama proteina nalaze se u promenljivoj protein_info. Ova promenljiva sadrži tuple-ove sa početnim i završnim pozicijama za svaki od proteina od interesa.
Na osnovu ovih pozicija u neporavnatom genomu, izdvojili smo sekvence za svaki protein i sačuvali ih u odvojene fajlove. Na primer, za protein "ORF1a polyprotein" sa pozicijama (266, 13483), izdvojili smo region od 266. do 13483. nukleotida iz svake sekvence i sačuvali u poseban fajl koji nosi naziv proteina.
Ovaj kod definiše funkciju find_aligned_positions, koja se koristi za pronalaženje početne i krajnje pozicije poravnate sekvence koja odgovara odredjenom delu referentnog proteina.
Funkcija find_aligned_positions prima tri argumenta: start_position i end_position, koji predstavljaju početnu i krajnju poziciju referentnog proteina, i aligned_sequence, koja je poravnata sekvenca referentnog proteina. U prvoj petlji, funkcija prolazi kroz poravnatu sekvencu dok ne nađe poziciju koja odgovara start_position. Ovo se postiže brojanjem ne '-' karaktera u poravnatoj sekvenci i upoređivanjem sa vrednošću start_position. Nakon što se pronađe početna pozicija, druga petlja se koristi za pronalaženje krajnje pozicije tako što ponovo prolazi kroz poravnatu sekvencu i broji karaktere do kraja proteina. Konačno, funkcija vraća početnu i krajnju poziciju u poravnatoj sekvenci koja odgovara delu referentnog proteina.
U kontekstu poravnanja sekvenci, simbol '-' se koristi za označavanje praznih pozicija ili umetanja u odnosu na referentnu sekvencu. Kada se analizira više sekvenci, svaka sekvencija se poravnava s referentnom sekvencijom kako bi se identifikovale zajedničke pozicije i razlike. Ako se na određenim pozicijama u referentnoj sekvenci nalaze aminokiseline, a u nekim drugim sekvencama ne, umesto nedostajućih aminokiselina koristi se '-' kako bi se očuvala struktura poravnanja.

"Proteini bez minusa"[4] se odnosi na proteine kod kojih se ne javljaju "minus"ili negativne vrednosti za dužinu proteina prilikom identifikacije granica. Kada se radi identifikacija granica proteina, vodi se računa o pojavama umetanja i brisanja u referentnom genomu, kao i u ostalim sekvencama. Ovi događaji mogu uticati na dužinu proteina, pa se granice proteina mogu prilagoditi u skladu sa prisutnim umetanjima i brisanjima kako bi se tačno odredila funkcionalna regija proteina.
Proces izdvajanja proteina iz poravnatih sekvenci uključuje čišćenje ovih proteina od '-' karaktera kako bi se dobio samo niz aminokiselina koji čini protein, bez praznih mesta ili umetanja. Ovi čisti proteini se zatim koriste za dalju analizu ili istraživanje.
Za učitavanje poravnatih sekvenci koristili smo AlignIO iz biblioteke Bio, o kojoj će kasnije biti više reči. Ovaj pristup omogućava nam rad sa pojedinačnim proteinima, olakšava analizu i pruža jasniji uvid u varijacije na nivou pojedinačnih proteina.
Fokusirali smo se na nekoliko ključnih proteina koji igraju važnu ulogu u strukturi i funkciji virusa. Kroz našu metodologiju, obradjivali smo proteine koji su zajedno sa svojim pozicijama u originalnom referentnom genomu navedeni ispod:

- ORF1a polyprotein (266, 13483)
- ORF1ab polyprotein (266, 13468)
- Surface Glycoprotein (21563, 25384)
- ORF3a Protein (25393, 26220)
- Envelope Protein (26245, 26472)
- Membrane Glycoprotein (26523, 27191)
- ORF6 Protein (27202, 27387)
- ORF7a Protein (27394, 27759)
- ORF7b Protein (27756, 27887)
- ORF8 Protein (27894, 28259)
- Nucleocapsid Phosphoprotein (28274, 29533)
- ORF10 Protein (29558, 29674)

[3] Program za identifikaciju granica: graniceProteina.ipynb
[4] Sacuvani u fajlu proteiniBezMinusa.txt

3.3 Analiza mutacija na nukleotidnom nivou

3.3.1 Uopšteno

Nakon poravnanja, analizirali smo različite tipove mutacija na nukleotidnom nivou. Uzeli smo u obzir substitucije, insercije i delecije, identifikujući specifične promene na svakoj poziciji u odnosu na referentnu sekvencu. Ove informacije omogućavaju nam da pratimo dinamiku genetičkih promena i identifikujemo potencijalno relevantne regione u genomu virusa. Jedna od ključnih varijabli u analizi bila je obrada neidentifikovanih nukleotida (N). Razmatrali smo tri opcije: računanje N kao nukleotidne zamene (N računamo kao mutaciju), računanje N kao referentnog nukleotida (bez mutacije), i isključivanje sekvenci sa N iz analize određene pozicije. Ova razmatranja omogućila su nam bolje razumevanje uticaja neidentifikovanih nukleotida na analizu mutacija.

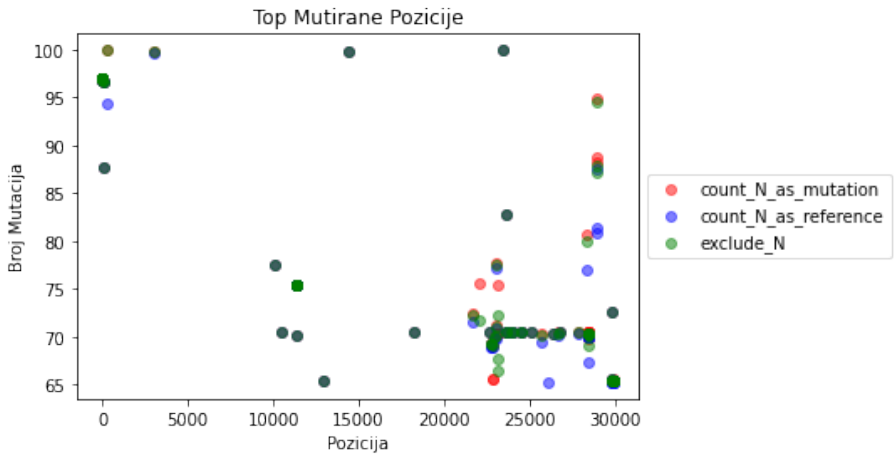
Nakon identifikacije mutacija, analizirali smo frekvencije mutiranih pozicija. Koristeći Counter objekat, mapirali smo procenat mutacije na svakoj poziciji, čime smo dobili uvid u pozicije sa najvećim procentom mutacije, kao i pozicije sa najmanjim procentom mutacije u sekvencama. U zavisnosti od tehnike koju smo koristili za obradu neidentifikovanih nukleotida dobili smo različite rezultate.
Za izdvajanje pozicija sa najvećom i najmanjom mutacijom iz Counter objekta koristili smo funkciju find_top_mutated_positions_within_range, ona nam omogućava pronalaženje top N mutiranih pozicija unutar određenog opsega. Evo kako funkcija radi:
mutation_counter: Ovaj parametar predstavlja Counter objekat koji sadrži informacije o broju mutacija na različitim pozicijama.
top_n: Broj najviše (mutiranih pozicija koje želimo pronaći.
lower_bound i upper_bound: Granice opsega unutar kojeg tražimo mutirane pozicije. Podrazumevane vrednosti su 0 za lower_bound i dužina mutation_counter za upper_bound.
find_max: Logička vrednost koja određuje hoćemo li pronaći najveće ili najmanje mutirane pozicije. Ako je True, funkcija će pronaći najveće mutirane pozicije, inače će pronaći najmanje.
threshold: Prag koji određuje minimalni broj mutacija koji je potreban da bi pozicija bila uzeta u obzir. Ako je postavljen, samo će pozicije sa brojem mutacija većim od ovog praga biti uključene u rezultate. Ako nije postavljen, sve pozicije unutar granica će biti uzete u obzir.

U daljem tekstu izdvojićemo pozicije 5 najređe i 5 najčešće mutiranih pozicija (u fajlove smo izdvojili 100 pozicija, uzimajući u obzir da se dužine sekvenci razlikuju, pa to može uzrokovati pojavu da su najčešće mutirane početne i krajnje pozicije).
Sve prikazane pozicije se odnose na pozije u odnosu poravnate sekvence.

3.3.2 Pozicije sa najvećim procentima mutacija:

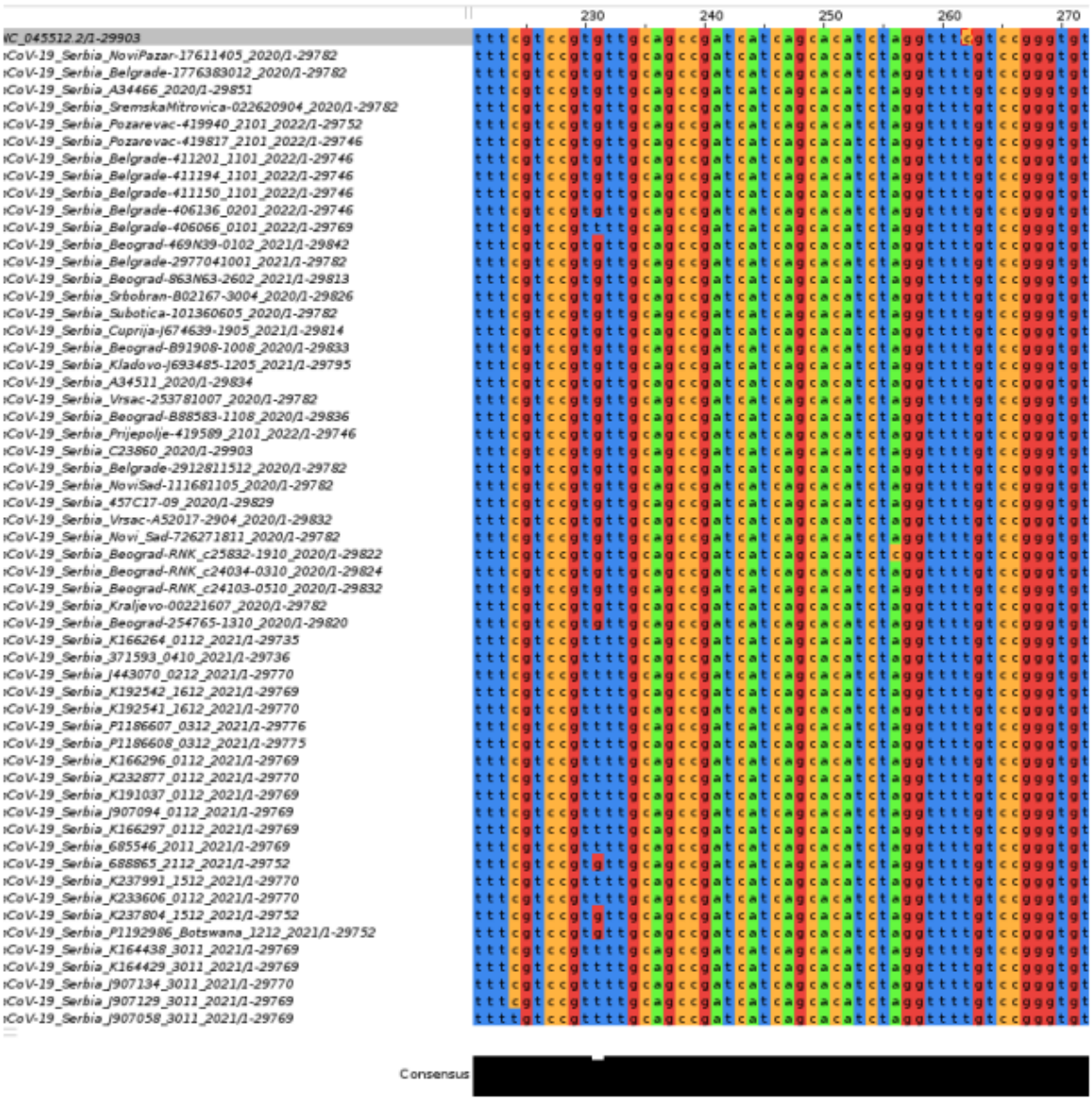
Tabela 1: Top 5 pozicija sa najvećim procentima mutacija

Opcija za N	Pozicija	Procent mutacije
count N as mutation	Pozicija 262	99.96%
	Pozicija 23478	99.96%
	Pozicija 14435	99.87%
	Pozicija 3064	99.79%
	Pozicija 2	97.06%
count N as reference	Pozicija 23478	99.96%
	Pozicija 14435	99.87%
	Pozicija 3064	99.66%
	Pozicija 2	98.06%
	Pozicija 1	98.02%
exclude N	Pozicija 23478	99.96%
	Pozicija 262	99.96%
	Pozicija 14435	99.87%
	Pozicija 3064	99.79%
	Pozicija 2	98.06%



Slika 1: Pozicije sa najvećim procentom mutacije

Na grafiku su predstavljene pozicije sa najvećom mutacijom, za ovo iscrtavanje koristili smo podatke dobijene za sva tri različita razmatranja N nukleotida.
Primećujemo da su mutacije najviše skoncentrisane oko 30 000 nukleotida, ali one nam nisu od značaja, jer svi proteini koje obradjujemo se završavaju pre nje.
Osim tih mutacija, pozicije sa velikim mutacijama nalaze se oko 23 hiljadite pozicije. Pozicije se uglavnom poklapaju za sve tri opcije za N.



Slika 2: Pozicija 262 sa najvećom mutacijom

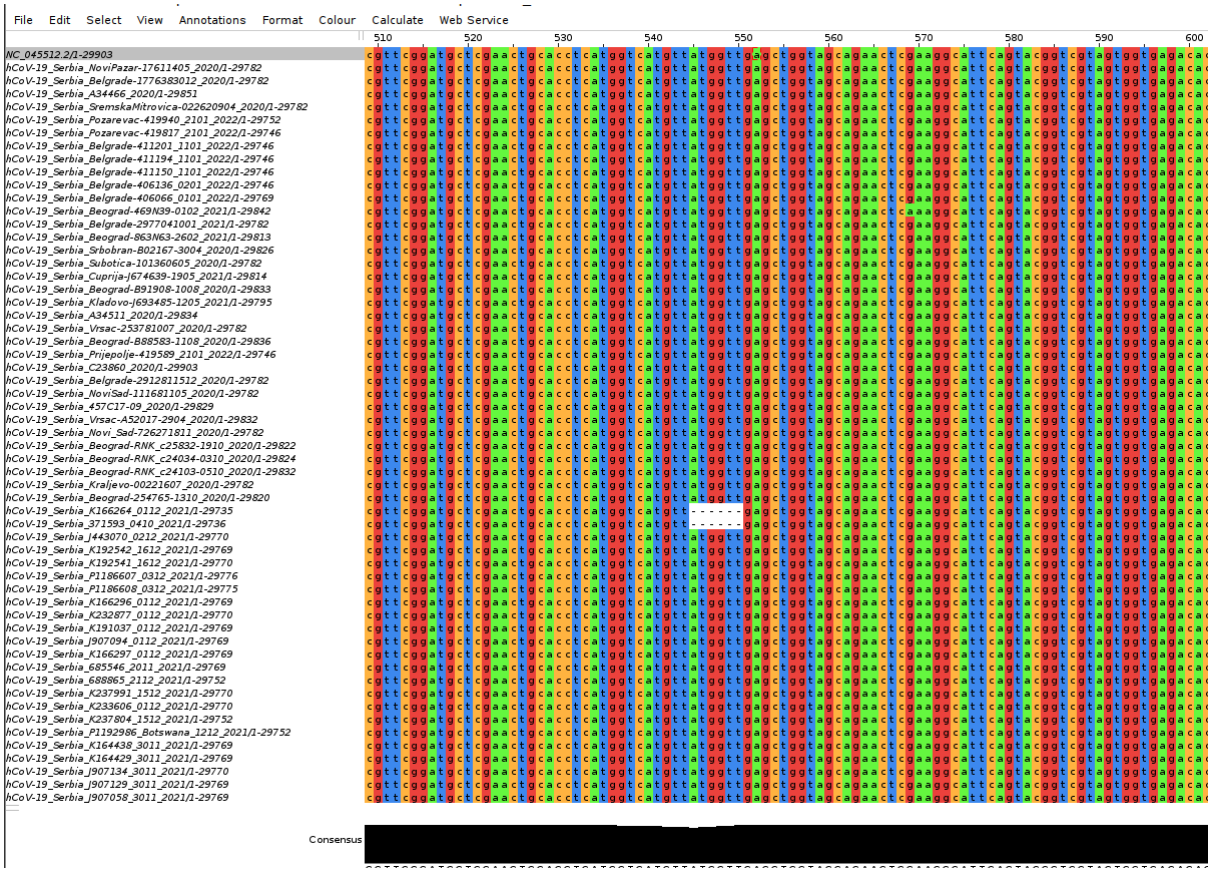
Na slici možemo da uočimo da je na poziciji 262. citozin iz referentne sekvence najčešće zamenjen timinom.
U fajlu mutacije_proteina.txt izdvojene su pozicije sa najvećom frekvencijom mutacije za svaki protein posebno.

3.3.3 Pozicije sa najmanjim procentima mutacija:

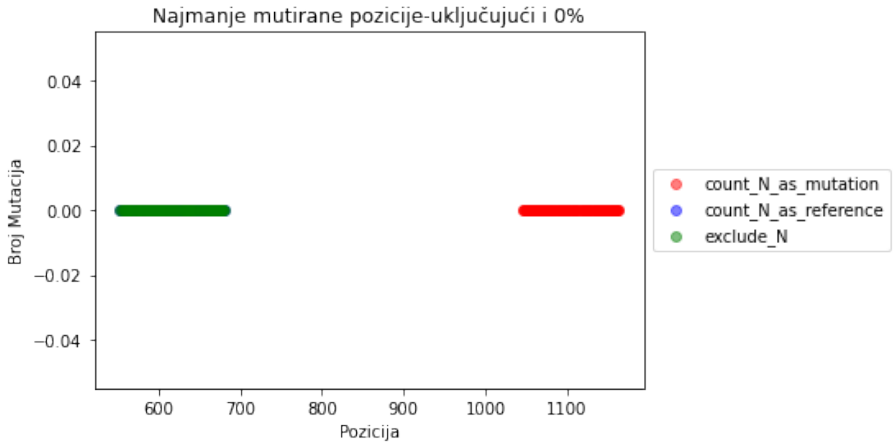
Ukoliko računamo pozicije na kojima nema mutacije, odnosno na kojima je procenat mutacije 0%, dobijamo veliki broj takvih pozicija. Prve pozicije bez mutacije, za svaku od opcija su prikazane u tabeli.

Tabela 2: Top 5 pozicija sa najmanjim procentima mutacija

Opcija za N	Pozicija	Procent mutacije
count N as mutation	Pozicija 1046	0.00%
	Pozicija 1047	0.00%
	Pozicija 1048	0.00%
	Pozicija 1049	0.00%
	Pozicija 1050	0.00%
count N as reference	Pozicija 552	0.00%
	Pozicija 555	0.00%
	Pozicija 557	0.00%
	Pozicija 558	0.00%
	Pozicija 559	0.00%
exclude N	Pozicija 552	0.00%
	Pozicija 555	0.00%
	Pozicija 557	0.00%
	Pozicija 558	0.00%
	Pozicija 559	0.00%

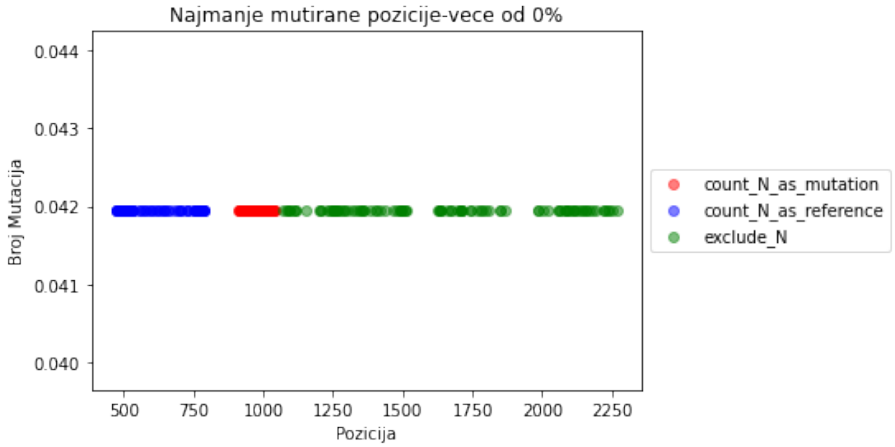


Slika 3: Pozicije sa najmanjom mutacijom



Slika 4: Pozicije sa 0% mutacije

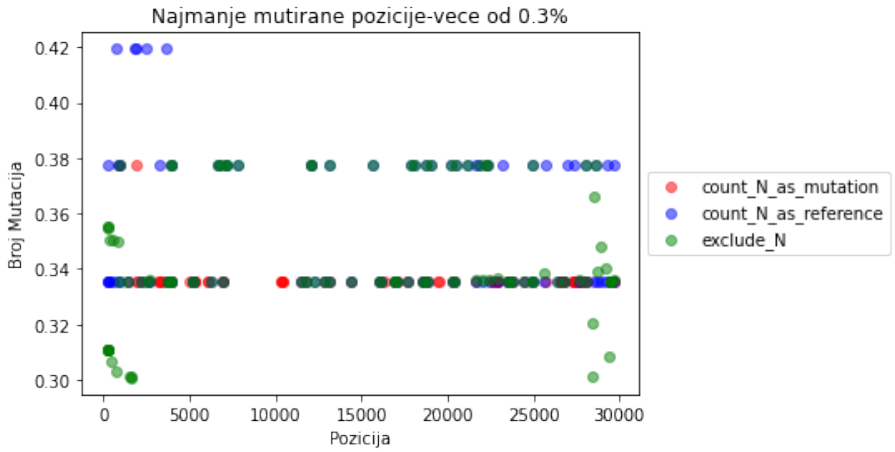
Pozicije sa najmanjim procentom mutacije su prikazane na grafiku, sve pripadaju proteinima ORF1a polyprotein i ORF1ab polyprotein.
Za opciju računanja N kao nukleotid referentne sekvence i opciju iskljucivanja sekvence u kojoj se javlja N dolazi do preklapanja pozicija.



Slika 5: Pozicije sa 0% mutacije

Ako tražimo mutacije veće od 0, dobićemo grafik iznad, koji nam još uvek ne govori mnogo o pozicijama unutar proteina, jer su sve pozicije pre 2250 i mutacije iznose 0.042%.
Ali ukoliko tražimo mutacije veće od 0.3 dobijamo zanimljiviji grafik.[5]. Ponovo primećujemo škupljanje"pozicija oko nekoliko vrednosti mutacija.

[5]Ukoliko želite da promenite najmanju vrednost mutacije, podesite argument threshold



Slika 6: Pozicije sa 0% mutacije

3.3.4 Regioni sa najređim mutacijama

Ako učestalost određujemo kao procenat zamena u odnosu na nukleotid referentnog izolata na konkretnoj poziciji, a mutiranost za region sabiranjem vrednosti za svaku poziciju i za N (neindentifikovani nukleotid) na toj poziciji smatramo da nema mutacije, dobijemo sledeći rezultat.
Top 5 regiona od 5 nukleotida koji su najređe mutirani:

180-184: 0.00%
181-185: 0.00%
182-186: 0.00%
183-187: 0.00%
184-188: 0.00%

Kao i za pojedinačne mutirane pozicije, veliki broj regiona dužine 5 imaće procenat mutacije 0. Da bismo ispitali koji regioni imaju procenat mutacije veći od nekog broja, možemo podesiti parametar threshold u funkciji find_least_mutated_regions (takodje menjanjem parametra region_size možemo promeniti dužinu regiona).
Ako za threshold postavimo 0, odnosno izdvojimo regione dužine 5, sa najmanjom mutacijom većom od 0, dobijamo sledeće rezultate:

478-482: 0.04%
556-560: 0.04%
557-561: 0.04%
565-569: 0.04%
571-575: 0.04%

U tekstualnom fajlu least_mut_regions u folderu Mutacije, izdvojili smo 100 takvih regiona.

3.3.5 Najduži regioni

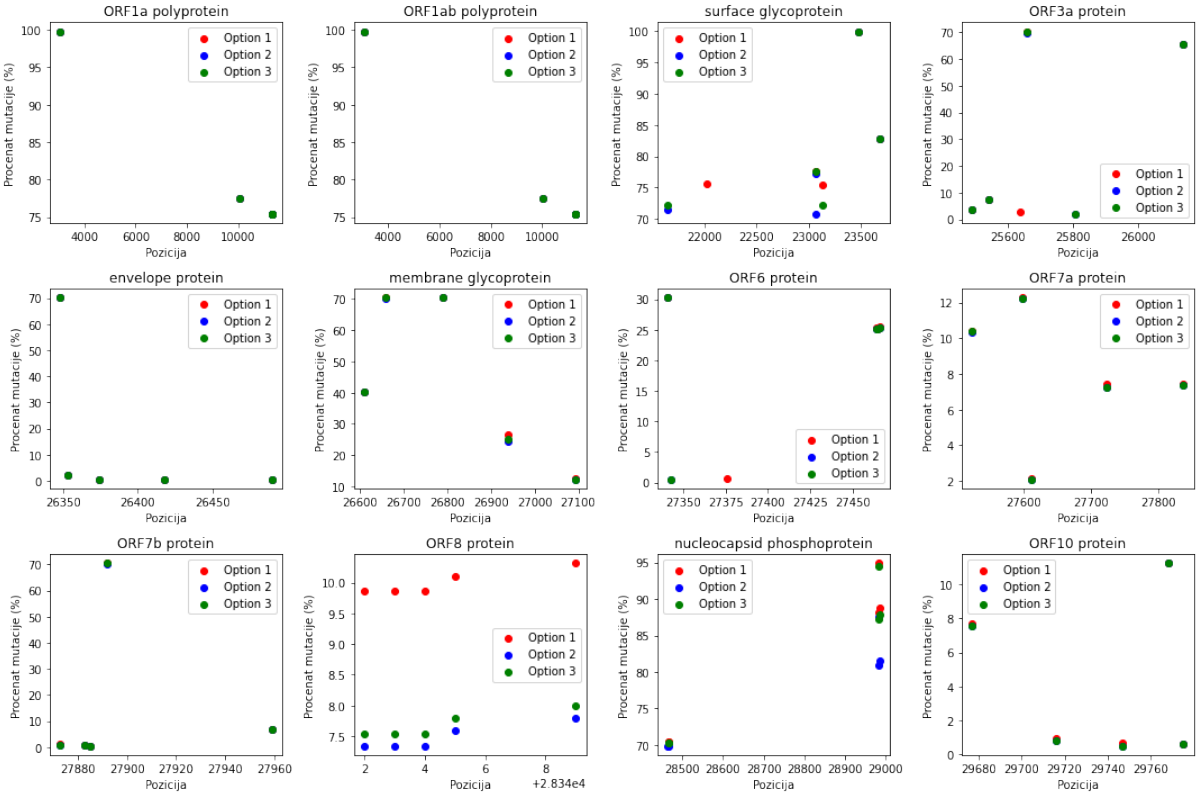
Najdužih 5 regiona (nukleotidnih nizova) u genomu SARS-CoV-2 virusa iz uzoraka sa teritorije Srbije, cija je stopa mutiranosti manja od 15%:

Region 1: 7201-8150,Duzina: 949, Stopa mutacije: 14.97%
Region 2: 7194-8142,Duzina: 948, Stopa mutacije: 14.85%
Region 3: 7201-8149,Duzina: 948, Stopa mutacije: 14.97%
Region 4: 7202-8150,Duzina: 948, Stopa mutacije: 14.93%
Region 5: 7219-8167,Duzina: 948, Stopa mutacije: 14.93%

Kao i pozije sa najmanjim procentom mutacije, tako i najduži regioni sa mutacijom ispod 15% pripaaju proteinima ORF1a polyprotein i ORF1ab polyprotein.

3.3.6 Najveće mutacije po proteinima

Na narednim graficima prikazani su pojedinačni proteini i njihove pozicije sa najvećom mutacijom za svaku od opcija za računanje N nukleotida.



Korišćena je biblioteka matplotlib.pyplot za crtanje grafika, a funkcija scatter se koristi za prikazivanje rasporeda tačaka na grafiku.

3.4 Analiza mutacija na aminokiselinskom nivou

Nakon što smo identifikovali mutacije na nukleotidnom nivou, prešli smo na analizu mutacija na aminokiselinskom nivou. Ova analiza se fokusirala na prevođenje nukleotidnih sekvenci u aminokiselinske sekvence, kako bismo dobili uvid u promene na nivou proteina.

3.4.1 Prevođenje nukleotidnih sekvenci u aminokiselinske

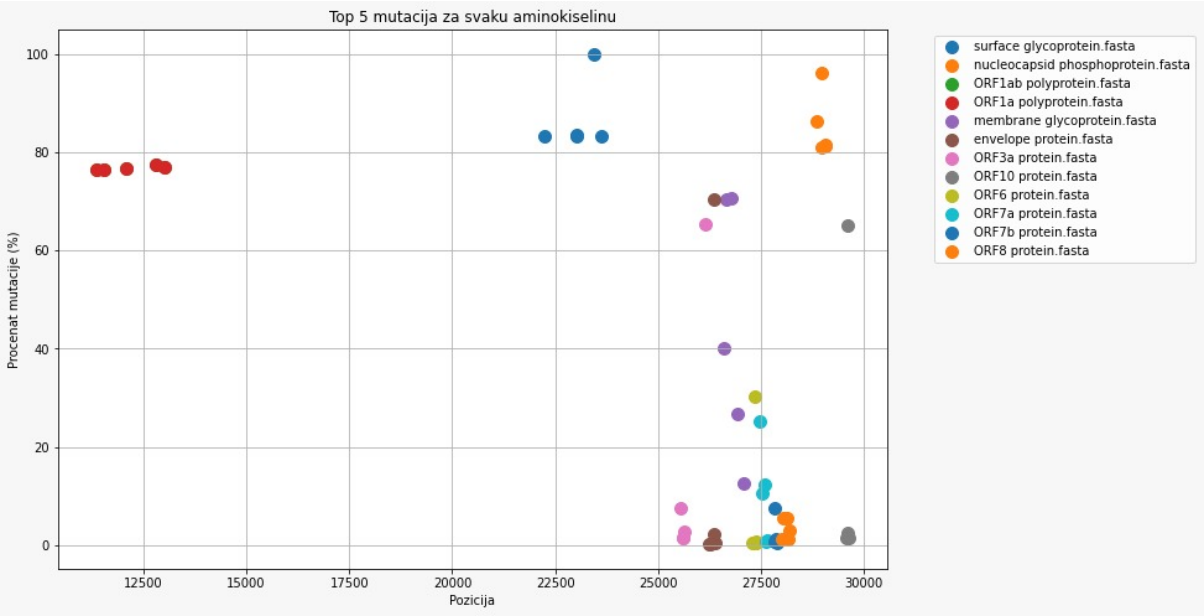
Korišćenjem standardnog genetskog koda (transl table=1), nukleodine sekvence smo preveli u aminokiselinske. Ovaj proces smo ponovili za svaki protein posebno. Koristili smo informacije o granicama proteina koje smo prethodno identifikovali.
Za prevodjenje smo koristili biološki-informacionu biblioteku Bio.Seq.
Prevedene aminokiseline za svaki protein nalaze se u Aminokiseline.

3.4.2 Upoređivanje rezultata

Na nivou aminokiselina, primetili smo da se kod različitih proteina frekvencije mutacija značajno razlikuju. Najmanju stopu mutacija uočili smo kod aminokiseline dobijene iz proteina ORF7b, gde pozicija sa najvećom mutacijom iznosi 7.63%. Pozicije sa najvećim procentom mutacije imaju aminokiseline koje su dobijene iz proteina surface glycoprotein i nucleocaspid phosphoprotein. Na nukleotidnom nivou zapazili smo da ovi proteini obuhvataju neke od pozicija sa najvećom stopom mutacije (npr pozicija 23478) Najčešće mutirane pozicije u aminokiselinama se ne nalaze u okviru najdužih regiona sa procentom mutacije ispod 15% na nukleotidnom nivou. Takođe, primetili smo da za skoro sve aminokiseline postoji jedna pozicija sa mnogo većim procentom mutacije u odnosu na prvu sledeću poziciju.
Da bismo mogli da poredimo pozije na nukleotidnom nivou sa pozicijama na aminokiselinskom nivou, bilo je neophodno da pozicije sa aminokiselinskog nivoa prevedemo na originalne pozicije.

Za konvertovanje aminokiselinske pozicije u broj nukleotida koristili smo formulu originalna_pozicija = start_position + (pozicija - 1) * 3, gde je start_positon početna pozicija proteina. Svaka aminokiselina u sekvenci predstavlja triplet nukleotida (kodon), i svaka pozicija u aminokiselinskoj sekvenci se odnosi na jedan od tih kodona. Oduzima se 1 jer aminokiselinske pozicije počinju od 1, ali indeksiranje u Pythonu počinje od 0. Zatim se množi sa 3 jer svaki kodon ima tri nukleotida.

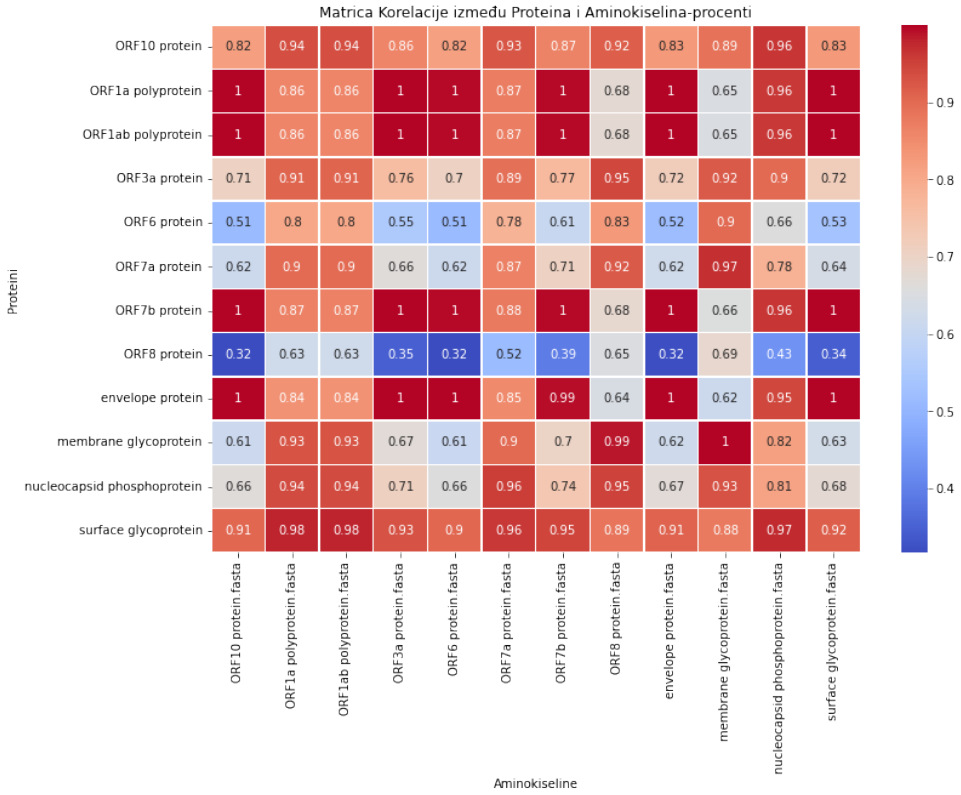
U nastaku vizuelno su prikazane pozicije sa najčešćim mutacijama za aminokiseline za opciju računanja nukleotida kao iz referentne sekvence, pozicije u aminokiselinama su konvertovane na već naveden tekstu.



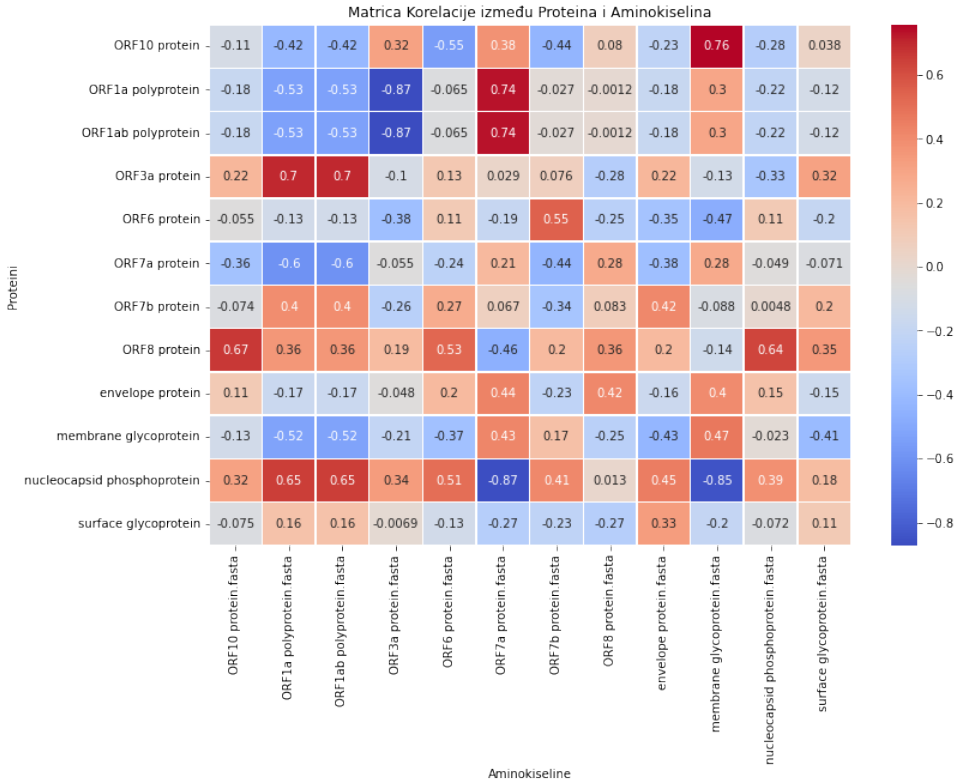
Izračunali smo i prikazali matricu korelacije između proteina i aminokiselina na osnovu procenata mutacija i na osnovu pozicija. Za čuvanje podataka koristili smo DataFrame iz biblioteke Pandas koji će

sadržavati rezultate korelacije. Indeksi su proteini, kolone su aminokiseline, a vrednosti će biti Pearson koeficijenti korelacije (pomoću funkcije pearsonr iz biblioteke scipy.stats).
Korelacija se odnosi na meru statističke zavisnosti između dva skupa podataka. Pearson koeficijent korelacije meri linearnu zavisnost između dva skupa podataka, i ima vrednosti između -1 (potpuna negativna korelacija) i 1 (potpuna pozitivna korelacija). Vrednost 0 znači odsustvo linearnog odnosa.
Korišćenjem biblioteke seaborn se pravi Heatmap matrice korelacije, gde boja označava jačinu korelacije.

Iz matrice korelacija, primećujemo da protein ORF8 ima nisku korelaciju sa svim aminokiselina. Ovo može ukazivati na to da mutacije u tom proteinu nisu usko povezane sa specifičnim aminokiselinama. Moguće je da su mutacije difuzne i ne pokazuju jasnu tendenciju da se koncentrišu na određene pozicije u aminokiselinskom nizu.
Dok za proteine ORF11a, ORF1ab i ORF10 primećujemo visoku korelaciju sa svim aminokiselinama. Ovo može sugerisati na postojanje određenih aminokiselinskih pozicija koje su često mutirane u sklopu tog proteina. Ako su mutacije fokusirane na određene delove aminokiselinskog niza, to može ukazivati na važnost tih regiona za funkciju proteina.
Osim toga, visoka korelacija između određenog proteina i aminokiselina može ukazivati na specifičnu interakciju ili vezivanje tog proteina sa određenim aminokiselinama.



Korelacija između pozicija proteina i aminokiselina može pružiti uvid u strukturu i funkciju proteina, jer pozicije koje su visoko korelisane mogu biti ključne za određene funkcije proteina ili za održavanje njegove strukture. Sa druge strane, pozicije sa niskom korelacijom mogu ukazivati na delove proteina koji su manje međusobno povezani ili imaju različite uloge u funkciji proteina.



4 Korišćeni alati i biblioteke

Program je implementiran u programskom jeziku Python u popularnom okruženju Jupyter Notebook. Koristili smo različite bioinformatičke biblioteke. Bio biblioteka u Pythonu, poznata i kao Biopython, predstavlja popularan set alatki i modula za rad sa biološkim podacima u programskom jeziku Python. Ova biblioteka pruža različite funkcionalnosti koje olakšavaju analizu i manipulaciju biološkim sekvencama, rad s biološkim bazama podataka, kao i implementaciju različitih bioinformatičkih algoritama.

4.1 Biopython biblioteka

Evo nekoliko ključnih aspekata i modula unutar Biopython biblioteke:

1. **Seq i SeqRecord klase:** - `Seq` klasa predstavlja niz bioloških podataka, kao što su nukleotidne ili aminokiselinske sekvence. Ova klasa omogućava manipulaciju i analizu bioloških podataka na nivou sekvenci. - `SeqRecord` klasa se koristi za čuvanje podataka o sekvencama zajedno sa dodatnim informacijama poput imena, opisa i drugih metapodataka.

2. **SeqIO modul:** - Ovaj modul omogućava učitavanje i snimanje sekvenci u različitim formatima, uključujući FASTA, GenBank, Clustal i drugi. Pomaže u olakšavanju radnji s biološkim podacima, posebno kada radite s različitim izvorima podataka. Referentni izolat učitan je pomoću BioPython's 'SeqIO' modula, zajedno sa ostalim sekvencama iz Srbije.

3. **Bio.SeqUtils modul:** - Pruža različite funkcije za analizu sekvenci, uključujući izračunavanje udela G+C, prevođenje nukleotidnih sekvenci u aminokiseline, pronalaženje start i stop kodona, itd.

4. **Bio.Align modul:** - Sadrži alate za poravnanje sekvenci. Različiti algoritmi poravnanja kao što su ClustalW, MUSCLE, i drugi su dostupni kroz ovaj modul.

5. **Bio.Entrez modul:** - Omogućava pristup NCBI (National Center for Biotechnology Information) bazama podataka, kao što su GenBank, PubMed, i druge. Pruža funkcionalnosti za pretragu, preuzimanje i analizu bioloških informacija sa NCBI resursa.

6. **Bio.PDB modul:** - Koristi se za rad s podacima o trodimenzionalnoj strukturi proteina. Ovaj modul omogućava analizu i manipulaciju PDB (Protein Data Bank) fajlovima.

7. **Bio.Graphics modul:** - Pruža alate za vizualizaciju bioloških podataka, uključujući crtanje sekvenci, dijagrama poravnanja i grafova.

Ovi moduli čine samo deo Biopython biblioteke, koja nudi još mnogo drugih funkcionalnosti. Više informacija i detalja možete pronaći na zvaničnoj veb stranici Biopython projekta.

4.2 MAFFT i JalView

Alat MAFFT je upotrebljen za poravnanje nukleotidnih sekvenci, a za prikaz rezultata korišćen je alat JalView.

MAFFT:

MAFFT (Multiple Alignment using Fast Fourier Transform) je alat za poravnanje više sekvenci nukleotida ili aminokiselina. Glavna funkcija MAFFT-a je da stvori poravnanje koje ukazuje na homologije između različitih sekvenci. Evo nekih ključnih karakteristika i informacija o MAFFT-u:

1. **Brzo poravnanje:** MAFFT koristi brze heurističke metode, uključujući FFT-NS-2 (Fast Fourier Transform) i iterative metode, kako bi efikasno poravna više sekvenci. Ovo ga čini pogodnim za analizu velikih skupova podataka.

2. **Iterativni pristup:** Iterativni pristup poravnanju u MAFFT-u pomaže u poboljšanju kvaliteta poravnanja. Algoritam prolazi kroz više faza poboljšanja kako bi konvergirao ka optimalnom poravnanju.

3. **Podrška za različite vrste sekvenci:** MAFFT podržava poravnanje nukleotidnih sekvenci (DNA i RNA) kao i aminokiselinskih sekvenci. Ovo ga čini svestranim alatom za različite vrste analiza.

4. **Različite metode poravnanja:** MAFFT nudi nekoliko različitih metoda poravnanja, uključujući L-INS-i (najprecizniji algoritam), FFT-NS-2 (brza heuristička metoda) i E-INS-i (za evolucijski daleko srodne sekvence). Korisnicima omogućava odabir odgovarajuće metode u zavisnosti od specifičnosti problema.

5. **Interfejsi komandne linije i grafički interfejsi:** MAFFT se može koristiti putem komandne linije, što ga čini pogodnim za automatizaciju u okviru skriptiranja. Takođe postoje i grafički interfejsi koji olakšavaju upotrebu za korisnike sa manje iskustva u radu sa komandnom linijom.

JalView:

JalView je vizualizacijski alat za analizu poravnanja sekvenci. Njegov primarni cilj je olakšati analizu evolucijskih odnosa između sekvenci i identifikaciju konzerviranih i promenljivih regiona. Evo nekoliko ključnih karakteristika JalView-a:

1. **Vizualizacija poravnanja:** JalView pruža interaktivnu vizualizaciju poravnanja sekvenci. Korisnici mogu jasno videti konzervirane regione, insercije, delecije i druge karakteristike poravnanja.

2. **Analiza evolucijskih odnosa:** Alat omogućava korisnicima da analiziraju evolucijske odnose između sekvenci pomoću različitih metoda, uključujući izračunavanje identiteta, sličnosti i udaljenosti.

3. **Dodatne informacije o sekvencama:** JalView prikazuje dodatne informacije o sekvencama, uključujući svojstva pojedinih aminokiselina ili nukleotida. Takođe podržava prikazivanje strukturnih informacija o proteinima.

4. **Bojenje i označavanje:** Korisnici mogu prilagoditi boje i oznake kako bi istakli određene osobine u poravnanju. Ovo je posebno korisno za analizu specifičnih regiona ili karakteristika.

5. **Integracija sa drugim alatima:** JalView se može integrisati sa drugim alatima za analizu sekvenci i struktura. Takođe podržava uvoz i izvoz podataka u različitim formatima.

6. **Interaktivno označavanje:** Korisnici mogu interaktivno označavati i uređivati sekvence direktno iz interfejsa JalView-a.

7. **Podrška za različite formate poravnanja:** Alat podržava različite formate poravnanja, uključujući Clustal, FASTA, Stockholm i druge.

4.3 Instalacije

Potrebni alati i biblioteke i njihova instalacija na Linux operativnom sistemu.

4.3.1 Alat MAFFT

Instalacija: sudo apt install mafft
Pozivanje iz terminala, za njegovo korišnjenje potrebno je da fajl ispunjava odredjene uslove, npr. da sekvence počinju znakom »"(Ostale uslove možete pročitati u dokumentaciji za ovaj alat.)
mafft spojene_sekvence.fasta > poravnate_sekvence.fasta

4.3.2 Python biblioteke

- **Biopython**: Za rad s biološkim podacima, uključujući analizu sekvenci.
  `pip install biopython`
- **Matplotlib**: Za crtanje grafika.
  `pip install matplotlib`
- **NumPy**: Za rad s numeričkim podacima.
  `pip install numpy`
- **SciPy**: Za statističke analize.
  `pip install scipy`
- **Pandas**: Za rad s podacima u obliku DataFrame.
  `pip install pandas`
- **Seaborn**: Za poboljšanje vizualizacija.
  `pip install seaborn`

5 Zaključak

Ova istraživanja su pružila dubok uvid u genetsku raznolikost lokalnih sojeva SARS-CoV-2 u populaciji Srbije. Kroz analizu nukleotidnih sekvenci, identifikaciju granica proteina i detaljnu analizu mutacija na nukleotidnom i aminokiselinskom nivou, istraživanje je doprinelo razumevanju genetičkih karakteristika ovog virusa u lokalnom kontekstu.

Ključna otkrića obuhvataju identifikaciju najčešće i najređe mutiranih pozicija u genomu, analizu regiona sa najmanjom i najvećom stopom mutacija, kao i vizualizaciju evolutivnih odnosa između sekvenci. Korišćenjem alata MAFFT za poravnanje i JalView za vizualizaciju, istraživanje je demonstriralo efikasnost bioinformatičkih pristupa u analizi velikog broja sekvenci.

6 Literatura

Literatura

[1] Wei-jie Guan i dr. "Clinical Characteristics of Coronavirus Disease 2019 in China". U: *New England Journal of Medicine* 382.18 (2020.), str. 1708–1720.

[2] Chaolin Huang i dr. "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China". U: *The Lancet* 395.10223 (2020.), str. 497–506.

[3] Stephen A Lauer i dr. "The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application". U: *Annals of Internal Medicine* 172.9 (2020.), str. 577–582.

[4] Puja Mehta i dr. "COVID-19: consider cytokine storm syndromes and immunosuppression". U: *The Lancet* 395.10229 (2020.), str. 1033–1034.

[5] Ani Nalbandian i dr. "Post-acute COVID-19 syndrome". U: *Nature Medicine* 27.4 (2021.), str. 601–615.

[6] D Wrapp, N Wang, KS Corbett i dr. "Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation". U: *Science* 367.6483 (2020.), str. 1260–1263.

[7] Joseph T Wu i dr. "Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China". U: *Nature Medicine* 26.4 (2020.), str. 506–510.

[8] P Zhou, XL Yang, XG Wang i dr. "A pneumonia outbreak associated with a new coronavirus of probable bat origin". U: *Nature* 579.7798 (2020.), str. 270–273.

1. https://mafft.cbrc.jp/alignment/software/
2. https://www.gisaid.org/
3. https://biopython.org/
4. https://matplotlib.org/
5. https://numpy.org/
6. https://www.scipy.org/
7. https://pandas.pydata.org/
8. https://seaborn.pydata.org/