Univerzitet u Beogradu
Matematički fakultet

ISTRAŽIVANJE UTICAJA P-ADIČNOSTI NA GENETSKI KOD KORONAVIRUSA: ANALIZA SEKVENCI I KLASTEROVANJE

Seminarski rad iz predmeta Istraživanje podataka 2

Profesor:
Nenad Mitić

Studenti:
Jelisaveta Gavrilović 188/2020
Marko Paunović 104/2020

Beograd, maj 2024.

Sadržaj

1 Uvod **2**

2 Genetski kod **3**

3 Preprocesiranje podataka **5**

4 Analiza površinskih proteina **7**
4.1 P-adično rastojanje 8
4.2 Hamingovo rastojanje 11
4.3 Analiza rezultata 11

5 Klasterovanje **15**
5.1 Edit rastojanje 15
5.2 Hijerarhijsko sakupljajuće klasterovanje 16
5.3 Vizuelizacija klastera 18
5.4 Poređenje rezultata površinskih proteina sa p-adičnim rastojanjem 22

6 Zaključak **24**

1 Uvod

Genetski kod predstavlja osnovu biološke informacije u svim živim organizmima. On je ključan za razumevanje procesa genetičke ekspresije i sinteze proteina, te ima glavnu ulogu u prenošenju genetičke informacije.

Koronavirusi su velika porodica virusa koji mogu izazvati bolesti kod životinja i ljudi. Kod ljudi, poznato je da nekoliko koronavirusa izaziva respiratorne infekcije koje mogu varirati od blagih prehlada do ozbiljnijih bolesti kao što su Middle East Respiratory Syndrome Coronavirus (MERS), Severe Acute Respiratory Syndrome Coronavirus (SARS) i COVID-19, bolest uzrokovana novim koronavirusom SARS-CoV-2. Zbog brzine širenja i potencijalno teških posledica po zdravlje, istraživanje genetske strukture koronavirusa je od presudnog značaja za razvoj dijagnostičkih alata, terapija i vakcina.

Mi ćemo se fokusirati na analizu genetskog koda nekoliko značajnih koronavirusa: SARS-CoV (uzročnik SARS-a), MERS-CoV (uzročnik MERS-a), Bovine coronavirus (BCoV), Human coronavirus 229E (HCoV-229E) i Human coronavirus OC43 (HCoV-OC43). Ovi virusi predstavljaju različite grupe koronavirusa sa značajnim genetskim i epidemiološkim karakteristikama.

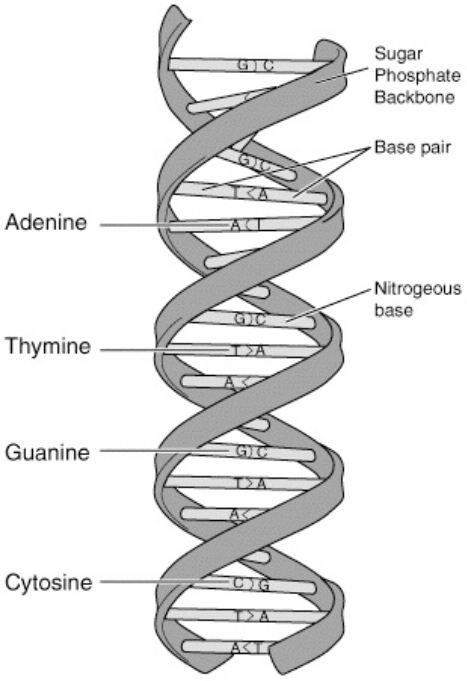
Cilj ovog rada je analizirati uticaj p-adičnosti na razlike u genetskom kodu između ovih različitih vrsta koronavirusa. P-adična analiza pruža novi pristup proučavanju genetskog koda, omogućavajući otkrivanje skrivenih obrazaca i struktura u genetskim sekvencama. Kombinacija p- adičnih rastojanja i Hammingovih rastojanja omogućava detaljno poređenje genetskih sekvenci na molekularnom nivou, što može pružiti uvid u evolutivne procese i funkcionalne razlike.

Osim toga, istraživanje će se fokusirati i na upotrebu hijerarhijskog klasterovanja za grupisanje genetskih sekvenci na osnovu njihove sličnosti. Ovakav pristup omogućava otkrivanje grupa sličnih sekvenci koje mogu ukazati na zajedničko poreklo ili funkcionalnu povezanost.

2 Genetski kod

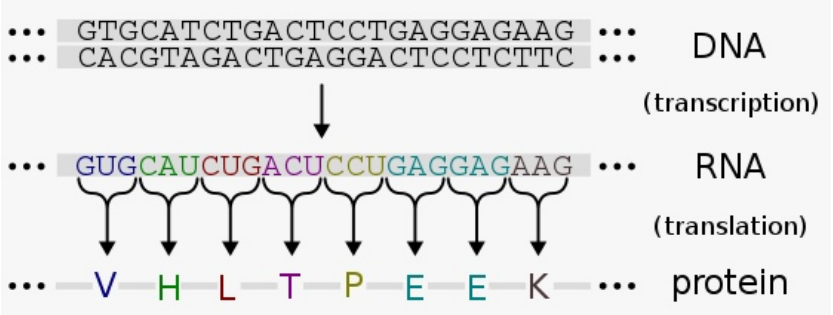
Sveukupna genetska informacija jednog organizma naziva se genom, a sva genetska informacija nalazi se u molekulu DNK. Svaki funkcionalni region molekula DNK naziva se gen. Gen je fizička i funkcionalna jedinica nasleđivanja koja prenosi naslednu poruku iz generacije u generaciju, a čini ga celovit deo DNK potreban za sintezu jednog proteina ili jednog molekula RNK. Svaki gen se putem procesa transkripcije prevodi u odgovarajući molekul RNK, koji se procesom translacije prevodi u sekvencu aminokiselina.

Genetski kod je jezik za prenošenje genetske poruke od DNK (gena) do proteina i sadržan je u redosledu baza na lancu DNK. Celokupan genetski kod sastoji se od jedinstvenog kombinovanja četiri tipa nukleotida DNK. Svaki nukleotid se sastoji od podgrupe koju čine fosfatna grupa, šećer dezoksiriboze i jedna od četiri moguće azotne baze, koje su grupisane u dve kategorije: purini i pirimidini. Purinske baze Adenin (A) i Guanin (G) su veće i sastoje se od dva aromatična prstena. Pirimidinske baze Citozin (C) i Timin (T) su manje i sastoje se od jednog aromatičnog prstena. Jedinica genetskog koda je niz od tri nukleotida (triplet) DNK i on se u celini komplementarno prenosi, transkripcijom, na informacionu RNK. Kod molekula RNK, Timin je zamenjen Uracilom (U) i šećer dezoksiriboze je zamenjen šećerom riboze. Triplet na informacionoj RNK naziva se **kodon** i predstavlja šifru za jednu aminokiselinu, dok niz kodona šifruje polipeptidni lanac.



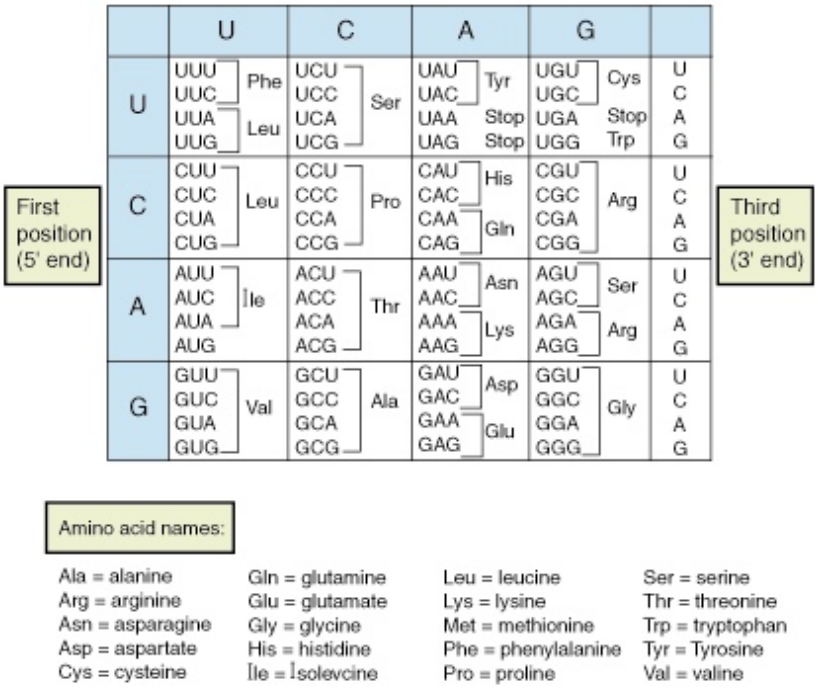
Slika 1: DNK molekul

Početak translacije zahteva prisustvo male ribozomalne jedinice koja se vezuje za start kodon na i-RNK, što zauzvrat označava gde i-RNK počinje da kodira određeni protein. U 98% slučajeva ovaj kodon je AUG. Proces elongacije traje sve dok ribozom ne naiđe na jedan od tri moguća stop kodona: UAA, UAG ili UGA, kada se translacija završava. Tada se zaustavlja sinteza polipeptidnog lanca i protein se oslobađa u citoplazmu.



Slika 2: Proces prevođenja sekvence DNK molekula u protein

Genetski kod je eksperimentalno dešifrovan sredinom 1960-ih (Bernfield i Nirenberg, 1965) i obično je predstavljen odgovarajućom tabelom, gde je eksplicitno navedena veza između kodona i aminokiselina, kao i stop signali. Dakle, možemo reći da eksperimentalno znamo šta je genetski kod. Međutim, ne možemo reći da postoji njegova potpuna teorijska deskripcija i razumevanje. Skoro sva živa bića koriste isti genetski kod, odnosno genetsku šifru koja se naziva standardni genetski kod, mali broj organizama koristi veoma male varijacije standardnog koda.



Slika 3: Standardni genetski kod

3 Preprocesiranje podataka

Nakon što smo se upoznali sa osnovama genetskog koda i procesom transkripcije i translacije, preprocesiranje referentnih genoma i proteina predstavlja naredni korak za pripremu podataka za detaljnu analizu. Svi podaci su preuzeti iz javno dostupne baze podataka Nacionalnog Centra za Biotehnološke Informacije (NCBI).

Jedan od najvažnijih koraka preprocesiranja je identifikacija otvorenih okvira za čitanje (ORF-ova). ORF-ovi su nizovi nukleotida u RNK koji potencijalno kodiraju proteine. Pronalaženjem ORF-ova, identifikovali smo regione koji mogu biti prepisani i prevedeni u proteinske sekvence.

Zašto je ovo važno?

Kodiranje proteina započinje identifikacijom odgovarajućih sekvenci u RNK koje sadrže informaciju o proteinskoj strukturi. Međutim, prema standardnom genetskom kodu jedna aminokiselina može biti kodirana različitim kombinacijama kodona. To znači da istu proteinsku sekvencu možemo prevesti iz više različitih RNK sekvenci. Identifikacija ORF-ova omogućila nam je precizno mapiranje ovih kodirajućih delova RNK tako što nam je pomogla u određivanju pozicije potencijalno kodirajućih delova RNK.

```python
def pronadji_orf(rnk_sekvenca, minimum, maksimum):
    start_kodon = 'AUG'
    stop_kodoni = ['UAA', 'UAG', 'UGA']
    orfovi = []

    # Pronalazenje svih pozicija start kodona
    start_pozicije = [i for i in range(len(sekvenca) - 2) if sekvenca[i:i+3] == start_kodon]

    for start_poz in start_pozicije:
        orf = ''
        for i in range(start_poz, len(sekvenca) - 2, 3):
            kodon = sekvenca[i:i+3]
            if kodon in stop_kodoni:
                if len(orf) >= minimum  and  len(orf) <= maksimum:
                    orfovi.append((start_poz, i+3, orf))
                break
            if len(orf) > maksimum:
                break
            orf += kodon

    return orfovi
```

Slika 4: Funkcija za pronalazenje ORF-ova

Nakon što identifikujemo ORF-ove, možemo prevesti RNK sekvence u sekvence aminokiselina, koristeći standardni genetski kod. Zatim upoređujemo dobijene proteinske sekvence sa referentnim proteinskim sekvencama kako bismo identifikovali koje sekvence kodiraju poznate proteine.

```
proteinske_sekvence_u_genomima = {}
for virus, genom in genomi.items():
    proteinske_sekvence_u_genomima[virus] = {}
    # Transkripcija DNK u RNK
    rnk_genom = Seq(genom).transcribe()

    # Kako bismo ubrzali proces poredjenja, izdvajacemo samo ORF-ove koji nisu kraci od najkraceg
    # i nisu duzi od najduzeg referentnog proteina
    minimum, maksimum = opseg_duzina_orfova(proteini)

    # Pronalazenje ORF-ova u sekvenci
    orfovi = pronadji_orf(rnk_genom, minimum, maksimum)

    # Prevodjenje ORF-ova u sekvence aminokiselina i poredjenje sa referentnim proteinima
    for start, end, orf_sekv in orfovi:
        sekv_aminokiselina = orf_sekv.translate()

        for naziv, protein in proteini[virus].items():
            if protein == sekv_aminokiselina:
                proteinske_sekvence_u_genomima[virus][naziv] = genomi[virus][start:end].transcribe()
```

Slika 5: Potencijalni proteini

Iako smo uspeli da identifikujemo većinu proteina ovom metodom, neki proteini nisu mogli biti automatski pronađeni. Za te proteine koristili smo dodatne informacije iz NCBI baze podataka kako bismo ručno odredili njihove pozicije u genomima. Ova ručna intervencija osigurava da imamo kompletan skup proteinskih sekvenci za dalju analizu.

```
# bcov
proteinske_sekvence_u_genomima['bcov']['orf1ab polyprotein [Bovine coronavirus]'] = \
        (genomi['bcov'][210:13332] + genomi['bcov'][13331:21491]).transcribe()

# human229e
proteinske_sekvence_u_genomima['human229e']['replicase polyprotein 1ab [Human coronavirus 229E]'] = \
        (genomi['human229e'][292:12520] + genomi['human229e'][12519:20565]).transcribe()

# humanoc43
proteinske_sekvence_u_genomima['humanoc43']['ORF1ab polyprotein [Human coronavirus OC43]'] = \
        (genomi['humanoc43'][209:13340] + genomi['humanoc43'][13339:21493]).transcribe()

# mers
proteinske_sekvence_u_genomima['mers']['1AB polyprotein [Middle East respiratory syndrome-related coronavirus]'] = \
        (genomi['mers'][278:13433] + genomi['mers'][13432:21511]).transcribe()

# sars1
proteinske_sekvence_u_genomima['sars1']['ORF1ab polyprotein [SARS coronavirus Tor2]'] = \
        (genomi['sars1'][264:13392] + genomi['sars1'][13391:21482]).transcribe()
```

Slika 6: Ručno dodati proteini

4 Analiza površinskih proteina

Površinski proteini koronavirusa, posebno spike (S) proteini, igraju ključnu ulogu u inficiranju domaćina. Oni omogućavaju virusu da se veže za receptore na ćelijama domaćina, što je prvi korak u procesu ulaska virusa u ćeliju. Razlike u ovim proteinima među različitim vrstama koronavirusa rezultiraju različitim receptorima na koje se virus vezuje, što utiče na patogenezu i prenosivost virusa.

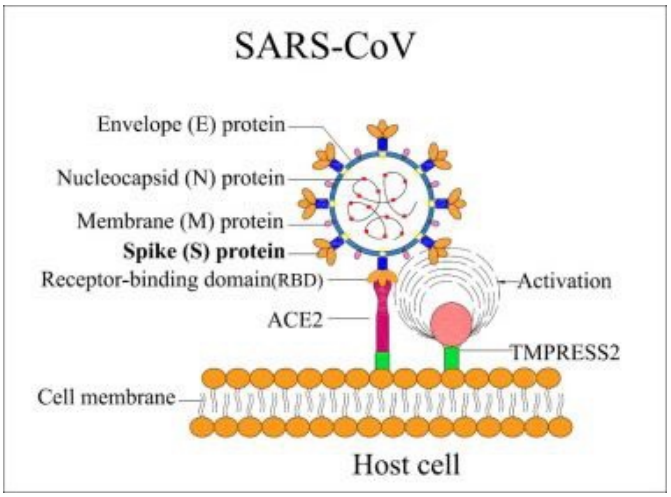
Osnovne razlike među vrstama koronavirusa:

Spike protein SARS-CoV-1 virusa se vezuje za ACE2 receptor na ćelijama domaćina. ACE2 receptor se nalazi na površini mnogih ljudskih ćelija, uključujući ćelije respiratornog trakta, što omogućava efikasnu infekciju i visoku stopu prenosa.

Nasuprot tome, spike protein MERS-CoV koristi DPP4 receptor za ulazak u ćelije domaćina. Ovaj receptor se razlikuje od ACE2 receptora, što rezultira različitom patogenezom i nižom stopom prenosa u poređenju sa SARS-CoV-1.

Spike protein HCoV-229E virusa se vezuje za APN receptor na ćelijama domaćina. Ova specifičnost omogućava HCoV-229E da efikasno inficira ljude, ali obično izaziva blaže respiratorne infekcije.

Pored spike proteina, BCoV i HCoV-OC43 virusi koriste hemaglutinin-esteraza (HE) protein za ulazak u ćelije domaćina. HE protein pomaže u razgradnji sluznog sloja, što omogućava spike proteinu da se efikasno veže za ćelijske receptore. Ova specifičnost omogućava BCoV-u da inficira goveda, dok HCoV-OC43 inficira ljude.



Slika 7: Vezivanje SARS-CoV za ćeliju domaćina

4.1 P-adično rastojanje

P-adično rastojanje je ključni koncept u našoj analizi površinskih proteina i koristićemo ga kao matematički alat za modeliranje genetskog koda. Ovaj pristup, predložen 2006. godine (Dragović i Dragović, 2006), istražuje bliskost kodona koji kodiraju istu aminokiselinu u ultrametričnom prostoru, uvodeći p-adični prostor kodona za p = 5 i p = 2. Link do rada.

Sada ćemo ukratko objasniti šta su p-adični brojevi.

P-adični brojevi predstavljaju cele brojeve u specifičnom sistemu zasnovanom na prostom broju p. Ovaj pristup omogućava određivanje udaljenosti između brojeva korišćenjem p-adične norme, koju označavamo: $|x|_p$.

Neka su x i y dva cela broja,

$$x = x_0 + x_1 p + x_2 p^2 + \ldots + x_k p^k \equiv x_0 x_1 x_2 \ldots x_k$$

,

$$y = y_0 + y_1 p + y_2 p^2 + \ldots + y_k p^k \equiv y_0 y_1 y_2 \ldots y_k$$

gde su $x_i \in \{0, 1, \ldots, p-1\}$ i $y_i \in \{0, 1, \ldots, p-1\}$ cifre brojeva u odgovarajućoj p-adičnoj bazi. Tada se udaljenost između x i y računa kao:

$$d_p(x,y) = |x-y|_p = \begin{cases} 1 & , x_0 \neq y_0 \\ \frac{1}{p} & , x_0 = y_0,\ x_1 \neq y_1 \\ \frac{1}{p^2} & , x_0 = y_0,\ x_1 = y_1,\ x_2 \neq y_2 \\ \vdots & \\ \frac{1}{p^k} & , x_0 = y_0,\ \ldots,\ x_{k-1} = y_{k-1},\ x_k \neq y_k \end{cases}$$

U našem kontekstu, p-adični pristup omogućava analizu "bliskosti"između kodona koji kodiraju istu aminokiselinu u genetskom kodu.

Korišćenjem Dragovićevog modela koji se oslanja na 5-adično rastojanje, prvo su pridruženi odgovarajući brojevi kodonima. Ova konstrukcija brojeva temelji se na prirodnim karakteristikama nukleotida:

Kao što je ranije rečeno, postoje tri pirimidinske (C, T, U) i tri purinske (A, G, I) azotne baze, gde I označava inozin koji se može koristiti kao deo antikodona transportne RNK (tRNK) kako bi se omogućilo uparivanje sa više različitih kodona, što doprinosi fleksibilnosti i efikasnosti procesa translacije. Budući da su Timin (T) i Uracil (U) praktično ekvivalentni, ostaje pet

nukleotida kojim treba pridružiti odgovarajućih pet cifara (jer p = 5) uzimajući u obzir da su purini (A, G) i pirimidini (C, U) međusobno sličniji nego purin u poređenju sa pirimidinom. Ova sličnost je prirodno opisana kroz 2-adično rastojanje.

Pošto je inozin poseban slučaj njemu je dodeljen broj 0, te nas to dovodi do toga da je $A \equiv 2$ i $G \equiv 4$ ili $G \equiv 2$ i $A \equiv 4$ jer $d_2(0,2) = d_2(4,2) = \frac{1}{2}$. Tada bi trebalo da bude $C \equiv 1$ i $U \equiv 3$ ili $U \equiv 1$ i $C \equiv 3$. Zbog uparivanja baza (A, U) i (C, D) trebalo bi da važi $A + U = C + G = 5$.

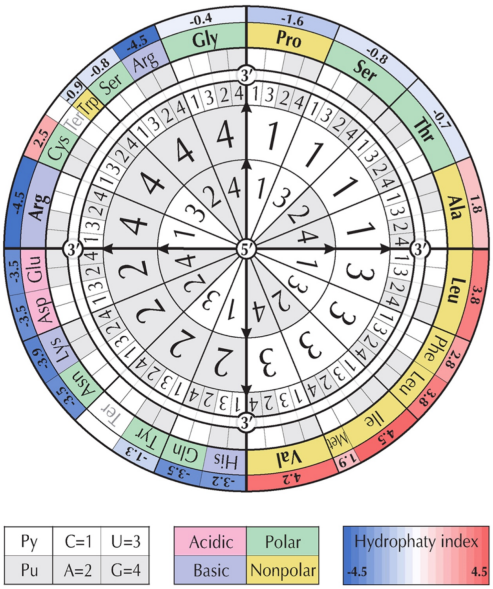
Na kraju, uzeta je identifikacija nukleotida:

$$I \equiv 0, C \equiv 1, A \equiv 2, U(T) \equiv 3, G \equiv 4.$$

Na osnovu identifikacije nukleotida, kodonima iz standardnog genetskog koda dodeljeni su brojevi:

CCC **111** Pro	ACC **211** Thr	UCC **311** Ser	GCC **411** Ala
CCU **113** Pro	ACU **213** Thr	UCU **313** Ser	GCU **413** Ala
CCA **112** Pro	ACA **212** Thr	UCA **312** Ser	GCA **412** Ala
CCG **114** Pro	ACG **214** Thr	UCG **314** Ser	GCG **414** Ala
CAC **121** His	AAC **221** Asn	UAC **321** Tyr	GAC **421** Asp
CAU **123** His	AAU **223** Asn	UAU **323** Tyr	GAU **423** Asp
CAA **122** Gln	AAA **222** Lys	UAA **322** Ter	GAA **422** Glu
CAG **124** Gln	AAG **224** Lys	UAG **324** Ter	GAG **424** Glu
CUC **131** Leu	AUC **231** Ile	UUC **331** Phe	GUC **431** Val
CUU **133** Leu	AUU **233** Ile	UUU **333** Phe	GUU **433** Val
CUA **132** Leu	AUA **232** Met	UUA **332** Leu	GUA **432** Val
CUG **134** Leu	AUG **234** Met	UUG **334** Leu	GUG **434** Val
CGC **141** Arg	AGC **241** Ser	UGC **341** Cys	GGC **441** Gly
CGU **143** Arg	AGU **243** Ser	UGU **343** Cys	GGU **443** Gly
CGA **142** Arg	AGA **242** Ter	UGA **342** Trp	GGA **442** Gly
CGG **144** Arg	AGG **244** Ter	UGG **344** Trp	GGG **444** Gly

Slika 8: Dodela p-adičnih projeva kodonima



Slika 9: Standardni genetski kod preko p-adičnih brojeva

Nakon što kodone prevedemo u brojeve, p-adično rastojanje možemo da izračunamo formulom:

$$d_5(a,b) = |a_1a_2a_3 - b_1b_2b_3|_5 + |a_4a_5a_6 - b_4b_5b_6|_5 + \ldots + |a_{3n+1}a_{3n+2}a_{3n+3} - b_{3n+1}b_{3n+2}b_{3n+3}|_5,$$

gde su a i b dve RNK sekvence sa $n+1$ kodona, $n = 0, 1, 2, \ldots$

```python
def p_adicno_rastojanje_kodona(a, b, p):
    a, b = str(a), str(b)

    if a[0] != b[0]:
        return 1
    elif a[1] != b[1]:
        return 1/p
    elif a[2] != b[2]:
        return 1/(p**2)

    return 0


def p_adicno_rastojanje(p_rnk1, p_rnk2, p=5):
    p_rastojanje = 0

    for (a, b) in zip(p_rnk1, p_rnk2):
        p_rastojanje += p_adicno_rastojanje_kodona(a, b, p)

    return p_rastojanje
```

Slika 10: Funkcije za računanje p-adičnog rastojanja

4.2 Hamingovo rastojanje

Hamingovo rastojanje je metoda za merenje razlika između dve sekvence iste dužine. Ova metoda broji pozicije na kojima se odgovarajući simboli razlikuju.

U kontekstu genetskog koda, Hamingovo rastojanje se koristi za upoređivanje sekvenci DNK ili RNK, posebno njihovih kodona. Postoje dva glavna pristupa u primeni Hamingovog rastojanja:

- upoređivanje RNK sekvenci preko kodona: analizom se utvrđuje koliko se kodoni razlikuju na istim pozicijama u različitim RNK sekvencama. Ova metoda je korisna za identifikaciju mutacija ili varijacija u genetskim sekvencama koje mogu uticati na funkciju proteina.

- upoređivanje aminokiselina: fokus je na direktnom poređenju aminokiselina koje su kodirane kodonima. Promene u kodonima mogu dovesti do kodiranja različitih aminokiselina, ali i različiti kodoni mogu da kodiraju istu aminokiselinu, što može imati značajan uticaj na strukturu i funkciju proteina. Upoređivanjem aminokiselina možemo bolje razumeti funkcionalne posledice genetiskih varijacija.

```
def hamingovo_rastojanje(a, b): # a i b su sekvence kodona ili aminokiselina
    a = np.array(list(a))
    b = np.array(list(b))

    haming = 0
    for x, y in zip(a, b):
        if str(x) != str(y):
            haming += 1

    return haming
```

Slika 11: Funkcija za računanje Hamingovog rastojanja

4.3 Analiza rezultata

Kako bismo mogli koristiti metode poput Hamingovog i p-adičnog rastojanja za kvantitativnu analizu, sekvence koje se porede moraju da budu iste dužine. U ovom radu korišćene su metode dopunjavanja kraćih sekvenci:

1. dopunjavanje najfrekventnijim kodonom: kraća sekvenca se dopunjava najfrekventnijim kodonom koji se pojavljuje u njoj do dužine duže sekvence,

2. dopunjavanje najfrekventnijim nukleotidom: kraća sekvenca se dopunjava najfrekventnijim nukleotidom koji se pojavljuje u njoj do dužine duže sekvence.

Dodatno, za računanje p-adičnog rastojanja koristili smo i treću metodu gde smo kraću sekvencu nakon prevođenja u p-adične brojeve, dopunili “000”. Kako p-adična vrednost ”000”ne postoji u standardnom genetskom kodu, to predstavlja dobar indikator da su na tim pozicijama sekvence različite.

Rezultati koji su dobijeni računanjem p-adičnog rastojanja:

P-adicna rastojanja kada krace sekvence dopunjujemo najfrekventnijim kodonom

	spike structural protein [Bovine coronavirus]	surface glycoprotein [Human coronavirus 229E]	spike surface glycoprotein [Human coronavirus OC43]	spike protein [Middle East respiratory syndrome-related coronavirus]	spike glycoprotein [SARS coronavirus Tor2]
spike structural protein [Bovine coronavirus]	0.00	1029.32	966.44	1095.76	1042.68
surface glycoprotein [Human coronavirus 229E]	1029.32	0.00	1031.20	1058.64	957.36
spike surface glycoprotein [Human coronavirus OC43]	966.44	1031.20	0.00	1061.76	1059.76
spike protein [Middle East respiratory syndrome-related coronavirus]	1095.76	1058.64	1061.76	0.00	1047.88
spike glycoprotein [SARS coronavirus Tor2]	1042.68	957.36	1059.76	1047.88	0.00

P-adicna rastojanja kada krace sekvence dopunjujemo najfrekventnijim nukleotidom

	spike structural protein [Bovine coronavirus]	surface glycoprotein [Human coronavirus 229E]	spike surface glycoprotein [Human coronavirus OC43]	spike protein [Middle East respiratory syndrome-related coronavirus]	spike glycoprotein [SARS coronavirus Tor2]
spike structural protein [Bovine coronavirus]	0.00	1033.92	966.44	1095.76	1042.68
surface glycoprotein [Human coronavirus 229E]	1033.92	0.00	1031.52	1056.96	954.08
spike surface glycoprotein [Human coronavirus OC43]	966.44	1031.52	0.00	1061.76	1059.76
spike protein [Middle East respiratory syndrome-related coronavirus]	1095.76	1056.96	1061.76	0.00	1047.88
spike glycoprotein [SARS coronavirus Tor2]	1042.68	954.08	1059.76	1047.88	0.00

P-adicna rastojanja kada krace sekvence dopunjujemo nulama

	spike structural protein [Bovine coronavirus]	surface glycoprotein [Human coronavirus 229E]	spike surface glycoprotein [Human coronavirus OC43]	spike protein [Middle East respiratory syndrome-related coronavirus]	spike glycoprotein [SARS coronavirus Tor2]
spike structural protein [Bovine coronavirus]	0.00	1076.84	969.00	1098.32	1071.08
surface glycoprotein [Human coronavirus 229E]	1076.84	0.00	1074.08	1095.76	976.68
spike surface glycoprotein [Human coronavirus OC43]	969.00	1074.08	0.00	1061.76	1086.20
spike protein [Middle East respiratory syndrome-related coronavirus]	1098.32	1095.76	1061.76	0.00	1071.24
spike glycoprotein [SARS coronavirus Tor2]	1071.08	976.68	1086.20	1071.24	0.00

Rezultati koji su dobijeni računanjem Hamingovog rastojanja poredivši kodone:

Hamingova rastojanja kada krace sekvence dopunjujemo najfrekventnijim kodonom

	spike structural protein [Bovine coronavirus]	surface glycoprotein [Human coronavirus 229E]	spike surface glycoprotein [Human coronavirus OC43]	spike protein [Middle East respiratory syndrome-related coronavirus]	spike glycoprotein [SARS coronavirus Tor2]
spike structural protein [Bovine coronavirus]	0.0	1317.0	1213.0	1334.0	1319.0
surface glycoprotein [Human coronavirus 229E]	1317.0	0.0	1300.0	1322.0	1218.0
spike surface glycoprotein [Human coronavirus OC43]	1213.0	1300.0	0.0	1316.0	1326.0
spike protein [Middle East respiratory syndrome-related coronavirus]	1334.0	1322.0	1316.0	0.0	1321.0
spike glycoprotein [SARS coronavirus Tor2]	1319.0	1218.0	1326.0	1321.0	0.0

Hamingova rastojanja kada krace sekvence dopunjujemo najfrekventnijim nukleotidom

	spike structural protein [Bovine coronavirus]	surface glycoprotein [Human coronavirus 229E]	spike surface glycoprotein [Human coronavirus OC43]	spike protein [Middle East respiratory syndrome-related coronavirus]	spike glycoprotein [SARS coronavirus Tor2]
spike structural protein [Bovine coronavirus]	0.0	1320.0	1213.0	1334.0	1319.0
surface glycoprotein [Human coronavirus 229E]	1320.0	0.0	1304.0	1328.0	1220.0
spike surface glycoprotein [Human coronavirus OC43]	1213.0	1304.0	0.0	1316.0	1326.0
spike protein [Middle East respiratory syndrome-related coronavirus]	1334.0	1328.0	1316.0	0.0	1321.0
spike glycoprotein [SARS coronavirus Tor2]	1319.0	1220.0	1326.0	1321.0	0.0

Rezultati koji su dobijeni računanjem Hamingovog rastojanja poredivši aminokiseline:

Hamingova rastojanja kada krace sekvence dopunjujemo najfrekventnijim kodonom

	spike structural protein [Bovine coronavirus]	surface glycoprotein [Human coronavirus 229E]	spike surface glycoprotein [Human coronavirus OC43]	spike protein [Middle East respiratory syndrome-related coronavirus]	spike glycoprotein [SARS coronavirus Tor2]
spike structural protein [Bovine coronavirus]	0.0	1269.0	1163.0	1283.0	1267.0
surface glycoprotein [Human coronavirus 229E]	1269.0	0.0	1254.0	1282.0	1162.0
spike surface glycoprotein [Human coronavirus OC43]	1163.0	1254.0	0.0	1260.0	1294.0
spike protein [Middle East respiratory syndrome-related coronavirus]	1283.0	1282.0	1260.0	0.0	1269.0
spike glycoprotein [SARS coronavirus Tor2]	1267.0	1162.0	1294.0	1269.0	0.0

Hamingova rastojanja kada krace sekvence dopunjujemo najfrekventnijim nukleotidom

	spike structural protein [Bovine coronavirus]	surface glycoprotein [Human coronavirus 229E]	spike surface glycoprotein [Human coronavirus OC43]	spike protein [Middle East respiratory syndrome-related coronavirus]	spike glycoprotein [SARS coronavirus Tor2]
spike structural protein [Bovine coronavirus]	0.0	1275.0	1163.0	1284.0	1267.0
surface glycoprotein [Human coronavirus 229E]	1275.0	0.0	1263.0	1284.0	1165.0
spike surface glycoprotein [Human coronavirus OC43]	1163.0	1263.0	0.0	1260.0	1294.0
spike protein [Middle East respiratory syndrome-related coronavirus]	1284.0	1284.0	1260.0	0.0	1269.0
spike glycoprotein [SARS coronavirus Tor2]	1267.0	1165.0	1294.0	1269.0	0.0

Ono što prvo možemo primetiti jesu slični obrasci u rastojanjima između površinskih proteina koronavirusa, što ukazuje na pouzdanost naše analize, nezavisno od specifične metode dopunjavanja kraćih sekvenci. Ovo je važno jer pokazuje da rezultati analize ostaju konzistentni i relevantni bez obzira na sitnije razlike u metodama obrade podataka.

Takođe, primećujemo da su rastojanja dobijena Hamingovim rastojanjem nešto veća u odnosu na rastojanja dobijena p-adičnim rastojanjem, što se može objasniti razlikama u načinu funkcionisanja ovih metrika. P-adična rastojanja uzimaju u obzir poziciju razlika u kodonu, dok Hamingova rastojanja jednostavno broje različite pozicije na kojima su kodoni različiti bez obzira na njihovu važnost. Ove razlike ukazuju na to da Hamingova rastojanja pružaju grublju sliku sličnosti, dok p-adična rastojanja daju detaljniji uvid u specifične evolucione promene.

Kada poredimo aminokiseline umesto kodona, dobijamo manja rastojanja. Ovo je očekivano jer različiti kodoni mogu kodirati istu aminokiselinu (degeneracija genetskog koda), smanjujući broj pozicija na kojima su sekvence različite. To pokazuje da poređenje na nivou aminokiselina može biti korisno za uvid u funkcionalnu sličnost.

Analizom rezultata, vidimo da postoje nekoliko parova proteina koji pokazuju veću sličnost:

- Spike structural protein [Bovine coronavirus] i Spike surface glycoprotein [Human coronavirus OC43].
- Spike glycoprotein [SARS coronavirus Tor2] i Surface glycoprotein [Human coronavirus 229E].

Bliskost između površinskih proteina Bovine coronavirusa i Human coronavirusa OC43 je razumljiva jer oba virusa pripadaju istoj virusnoj grupi beta-koronavirusa i za ulazak u ćeliju domaćina koriste dodatno HE protein. Slično, bliskost između SARS-CoV-1 i Human coronavirusa 229E može biti povezana sa sličnim mehanizmima interakcije sa receptorima domaćina.

Međutim, veću razliku u rastojanjima ima protein:

- Spike protein [Middle East respiratory syndrome-related coronavirus] sa spike proteinima ostalih koronavirusa.

Ove udaljenosti se mogu objasniti različitim evolutivnim stazama i funkcionalnim adaptacijama koje su ovi virusi prošli. MERS-CoV je poznat po svojoj specifičnosti prema DPP4 receptoru, dok ostali beta-koronavirusi koriste ACE2 receptor, što ukazuje na značajne razlike u strukturi i funkciji njihovih spike proteina.

5 Klasterovanje

Nakon analize sličnosti među površinskim proteinima koronavirusa korišćenjem p-adičnog i Hamingovog rastojanja, sada prelazimo na analizu kroz primenu edit rastojanja i klasterovanja. Kroz ovaj proces, cilj nam je da identifikujemo zajedničke karakteristike i evolutivne veze među proteinima.

5.1 Edit rastojanje

Edit rastojanje je mera različitosti između dva niza karaktera koja se koristi u bioinformatičkim analizama genetskih sekvenci. Ova metrika meri minimalni broj jednostrukih operacija ("edita") koji su potrebni da se jedan niz pretvori u drugi. Operacije koje se uzimaju u obzir su umetanje, brisanje i zamena karaktera.

Računanje edit rastojanja ponekad zahteva primenu dinamičkog programiranja radi efikasnog pronalaženja najkraćeg niza operacija koje su potrebne za transformaciju jednog niza u drugi. Ova tehnika omogućava precizno merenje razlike između sekvenci čak i kada su sekvence različitih dužina.

U bioinformatičkim istraživanjima, edit rastojanje je ključni alat za razumevanje genetskih promena, evolutivnih odnosa i funkcionalne sličnosti između različitih organizama ili sekvenci.

```python
def edit_rastojanje(str1, str2):
    duzina_str1 = len(str1)
    duzina_str2 = len(str2)

    # Inicijalizujemo matricu za čuvanje rastojanja
    rastojanja = [[0] * (duzina_str2 + 1) for _ in range(duzina_str1 + 1)]

    # Inicijalizujemo prvi red i prvu kolonu
    for i in range(duzina_str1 + 1):
        rastojanja[i][0] = i
    for j in range(duzina_str2 + 1):
        rastojanja[0][j] = j

    # Popunjavamo matricu rastojanja
    for i in range(1, duzina_str1 + 1):
        for j in range(1, duzina_str2 + 1):
            if str1[i - 1] == str2[j - 1]:
                cena = 0
            else:
                cena = 1
            rastojanja[i][j] = min(rastojanja[i - 1][j] + 1,        # brisanje
                                   rastojanja[i][j - 1] + 1,        # ubacivanje
                                   rastojanja[i - 1][j - 1] + cena) # zamena

    return rastojanja[duzina_str1][duzina_str2]
```

Slika 12: Računanje edit rastojanja

Ukupan broj referentnih proteina za naša 5 koronavirusa je 56. Imajući u vidu da za svaki referentni protein vršimo poređenje sa svakim, to nas dovodi do ukupno, 1540 kombinacija za računanje edit rastojanja. S obzirom na to da nekoliko RNK sekvenci koje kodiraju proteine imaju dužinu i preko 20000 nukleotida, proces računanja edit rastojanja bio je dugotrajan i zahtevan, a završio se nakon oko 14 sati, tako da ne preporučujemo da pokrećete kod.

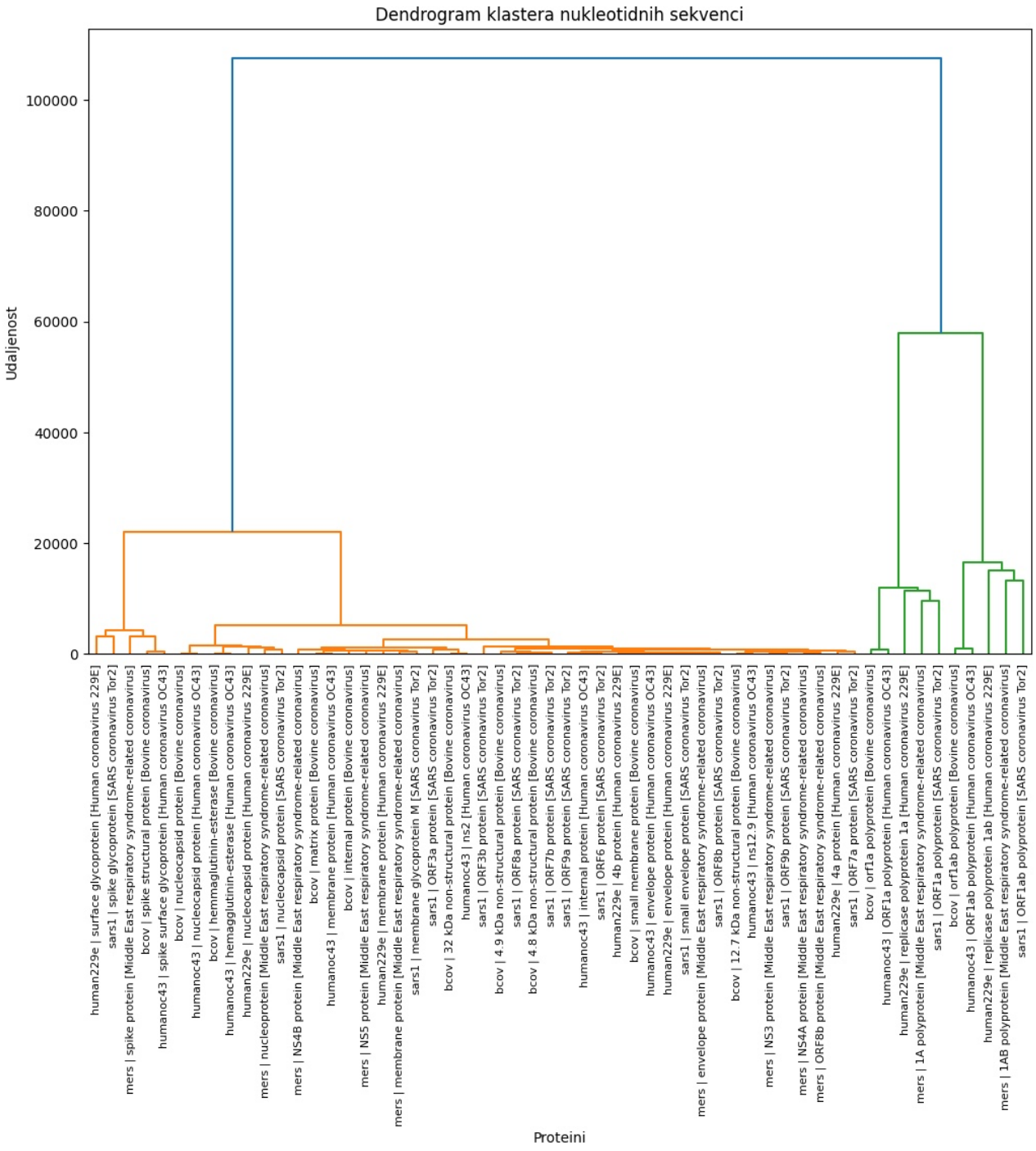
Dobijeni rezultati zabeleženi su u fajlu edit rastojanja, pružajući osnovu za dalje analize i interpretaciju genetskih sličnosti i razlika među koronavirusnim proteinima.

5.2 Hijerarhijsko sakupljajuće klasterovanje

Hijerarhijsko sakupljajuće klasterovanje (engl. Hierarchical Agglomerative Clustering) je tehnika klasterovanja koja grupiše podatke u hijerarhijsku strukturu. Ova tehnika počinje sa svakim podatkom kao odvojenim klasterom i zatim iterativno spaja najbliže klastere dok ne ostane samo jedan klaster koji sadrži sve podatke.

Ključni korak u analizi hijerarhijskog sakupljajućeg klasterovanja je interpretacija dendrograma, grafičke reprezentacije hijerarhijske strukture klastera. Dendrogram prikazuje način na koji su podaci grupisani u klastere i omogućava vizuelnu analizu sličnosti među grupama.

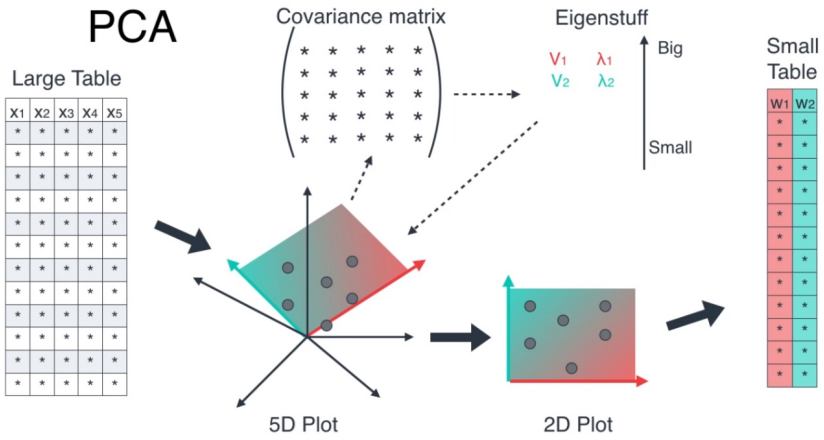
Nakon što smo dobili matricu rastojanja korišćenjem edit rastojanja kao mere sličnosti između svakog para proteina, primenili smo algoritam hijerarhijskog sakupljajućeg klasterovanja na osnovu ove matrice kako bismo grupisali proteine u klastere. Nakon detaljne analize dendograma, zaključili smo da postoji podela na tri glavne grupe, što je rezultiralo odlukom da koristimo 3 klastera kao optimalan broj.



Slika 13: Dendrogram klastera

5.3 Vizuelizacija klastera

Kako bismo dobijene klasterovane podatke vizuelizovali i lakše ih interpretirali, primenili smo analizu glavnih komponenti (engl. Principal Component Analysis - PCA) na matricu edit rastojanja. PCA je tehnika redukcije dimenzionalnosti koja transformiše originalne podatke u novi skup linearno nezavisnih promenljivih, poznate kao glavne komponente, koje zadržavaju maksimalnu varijansu podataka.

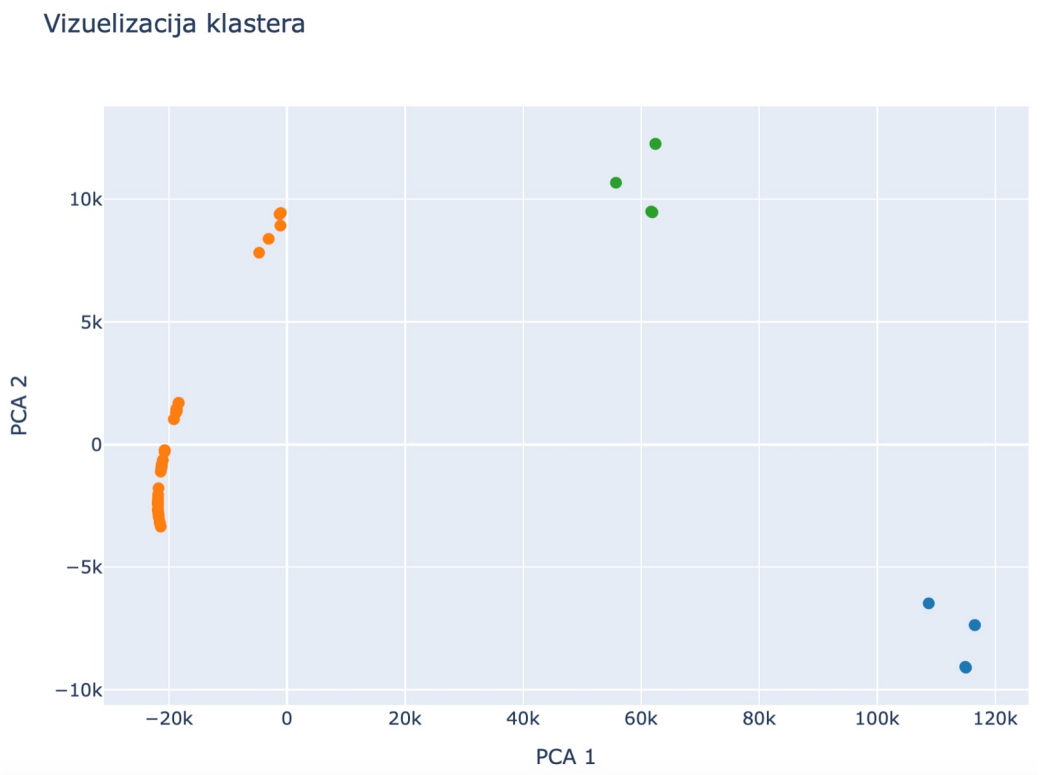


Slika 14: Princip smanjenja dimezionalnosti podataka

Koristeći PCA, sveli smo dimenzionalnost naših podataka na dve glavne komponente radi vizualizacije u dvodimenzionalnom prostoru, tj. grafikonskog prikazivanja podataka.

U kodu smo primenili PCA na matricu edit rastojanja, nakon čega smo dobili dvodimenzionalnu matricu koja predstavlja transformisane podatke. Ovo nam omogućava da i vizualno identifikujemo obrasce i grupisanja među proteinima.

Interaktivnu vizuelizaciju klastera možete pogledati ovde, a sliku istog u nastavku.



Slika 15: Vizuelizacija klastera

Na osnovu vizuelizacije klastera, možemo zaključiti sledeće:

- U zelenom klasteru nalaze se ORF1a i 1A polyproteini i replicase polyprotein 1a različitih koronavirusa. Ovi proteini su deo ne-strukturnih proteina koji su ključni za replikaciju i sintezu virusnog RNK genoma i proteina.
- Plavi klaster obuhvata ORF1ab i 1AB polyproteine i replicase polyprotein 1ab. Kao i proteini u zelenom klasteru, ovi proteini su ne-strukturni i esencijalni za procese transkripcije i replikacije virusne RNK.
- Narandžasti klaster obuhvata raznovrsne proteine koji su ključni za različite aspekte životnog ciklusa koronavirusa. On sadrži proteine različitih funkcija, struktura i interakcija, ali njihovo grupisanje zajedno ukazuje na određene zajedničke karakteristike među koronavirusima. Ovaj klaster uključuje:
  - Ne-strukturne proteine (NS3 protein, NS4A protein, NS4B protein, NS5 protein) koji su ključni za replikaciju virusa i modulaciju imunskog odgovora domaćina.

– Strukturne proteine poput envelope proteina, membrane proteina i nucleocapsid proteina, koji su važni za formiranje strukture virusnih čestica i održavanje njihovog integriteta.
– Površinske proteine, koji se grupišu u jednom podklasteru.

Pored ovih, narandžasti klaster uključuje i ostale proteine koji mogu imati različite uloge u infekciji, replikaciji i patogenezi virusa. Njihovo zajedničko grupisanje sugeriše da, uprkos funkcionalnoj raznolikosti, dele određene evolutivne ili strukturne karakteristike koje ih čine sličnijima jedne drugima u poređenju sa proteinima iz drugih klastera.

Posebno je važno napomenuti da su proteini koji imaju slične funkcije, i ako potiču od različitih koronavirusa, medjusobno bliži nego proteini sa različitim funkcijama. Ovo ukazuje na visok stepen konzervacije funkcionalno važnih proteina među različitim vrstama koronavirusa

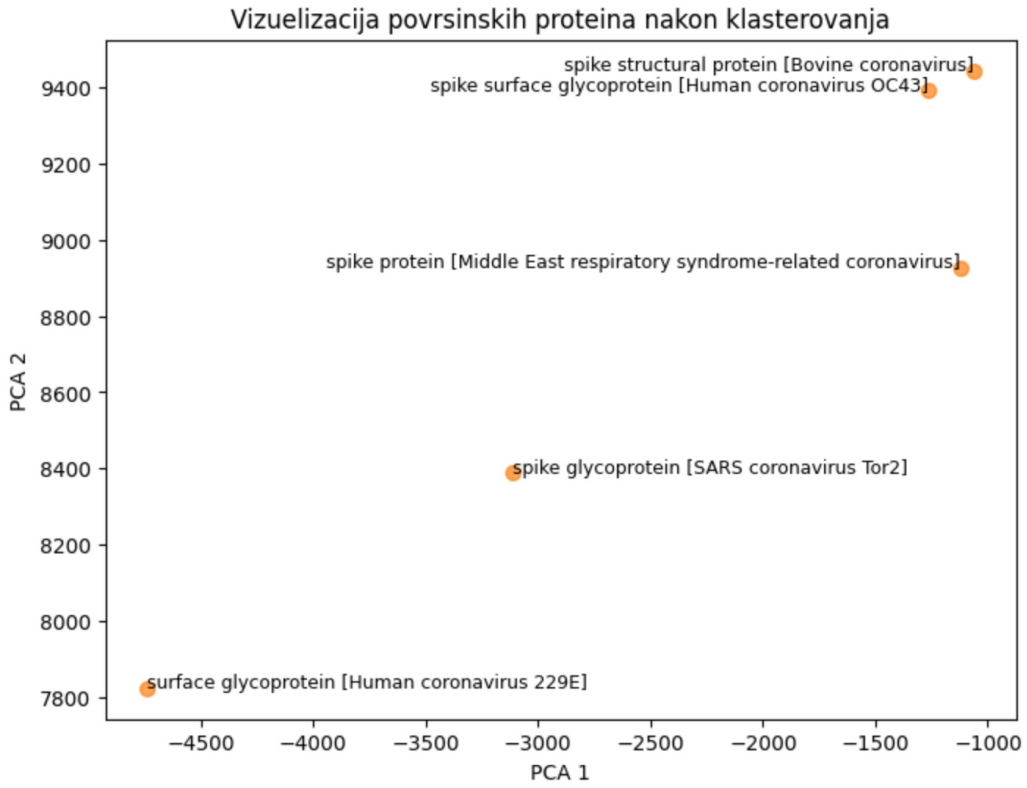
Da li zaista postoji podklaster ili se pojavljuje samo vizuelno zbog PCA?

Smanjenje dimenzionalnosti može dovesti do pojave da su neki podaci grupisani ili razdvojeni na način koji nije u potpunosti u skladu sa biologijom. U kontekstu analize površinskih proteina koronavirusa, iako možemo primetiti određene obrasce grupisanja koristeći PCA, važno je imati na umu da slika koju dobijamo malo iskrivljena. Ipak, smanjenjem dimenzionalnosti se jasno izdvajio podklaster površinskih proteina, što ukazuje na to da postoji stvarna biološka osnova za tu grupaciju.

Osim toga, dendrogram potvrđuje grupisanje površinskih proteina u određen podklaster (videdti sliku 13), što je dodatna podrška postojanju takve grupacije. Dendrogram nam pruža informacije o sličnosti između proteina na osnovu udaljenosti i načina njihovog povezivanja u klaster, što potvrđuje pretpostavku.

Znamo da su korona virusi jako infektivni i da lako prelaze sa jednog domaćina na drugog. To čine upravo pomoću površinskih proteina koji poseduju strukture koje se vezuju za receptore na površini ćelija domaćina i lako spajaju svoju membranu sa membranom domaćina sto omogućava ubacivanje virusnog genetskog materijala u ćeliju i njeno inficiranje. Pošto površinski proteini dele klaster sa drugim proteinima znamo da nisu toliko različiti od njih, ali pošto su izdvojeni u podklaster znači da se dovoljno razlikuju da bi ta razlika bila uočljiva. Ta razlika i jesu posebne strukture za prodiranje kroz ćelijsku membranu.

Kada bismo zumirali podklaster u kome se nalaze površinski proteini, mogli bismo detaljnije videti njihovu raspodelu. Međutim, treba napomenuti da je ovaj raspored prikazan na matrici na koju je primenjena PCA, čime je smanjena dimenzija sa 56x56.



Slika 16: Vizuelizacija podklastera

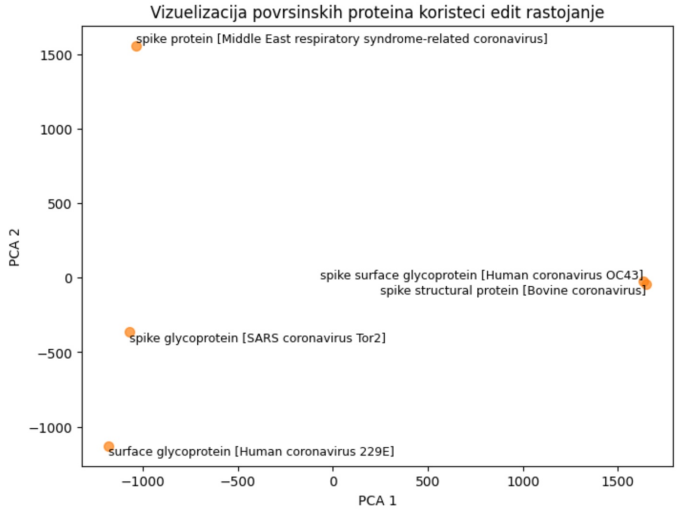
Rezultati koji su dobijeni računanjem edit rastojanja (prikaz samo za površinske proteine):

Edit rastojanja povrsinskih proteina

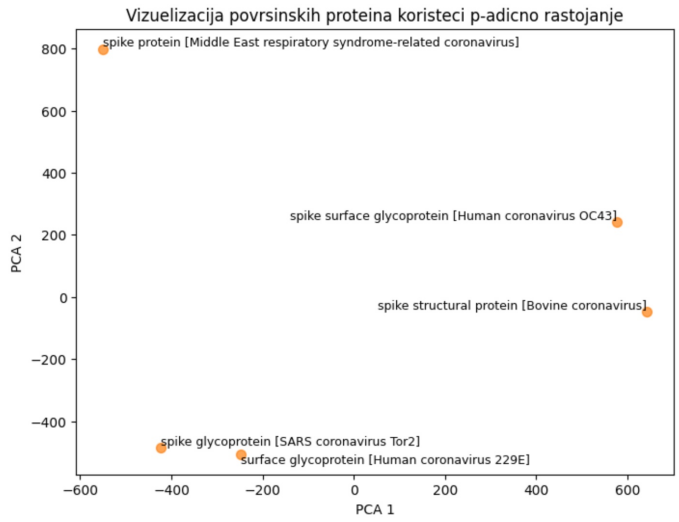
	spike structural protein [Bovine coronavirus]	surface glycoprotein [Human coronavirus 229E]	spike surface glycoprotein [Human coronavirus OC43]	spike protein [Middle East respiratory syndrome-related coronavirus]	spike glycoprotein [SARS coronavirus Tor2]
spike structural protein [Bovine coronavirus]	0.0	1930.0	276.0	1916.0	1879.0
surface glycoprotein [Human coronavirus 229E]	1930.0	0.0	1925.0	1952.0	1799.0
spike surface glycoprotein [Human coronavirus OC43]	276.0	1925.0	0.0	1887.0	1867.0
spike protein [Middle East respiratory syndrome-related coronavirus]	1916.0	1952.0	1887.0	0.0	1869.0
spike glycoprotein [SARS coronavirus Tor2]	1879.0	1799.0	1867.0	1869.0	0.0

5.4 Poređenje rezultata površinskih proteina sa p-adičnim rastojanjem

Kako bismo bolje i preciznije mogli da uporedimo sa p-adičnim rastojanjem, na rastojanja izmedju površinskih proteina ćemo sada primeniti PCA i time dimenziju smanjiti sa matrice 5x5 (PCA primenjujemo na matricu rastojanja u kojoj se nalaze samo rastojanja površinski proteini). Isto ćemo uraditi i kod p-adicnog rastojanja.



Slika 17: Vizuelizacija površinskih proteina koristeći edit rastojanje



Slika 18: Vizuelizacija površinskih proteina koristeći p-adično rastojanje

Razmatrajući rezultate koje smo dobili korišćenjem p-adičnog i edit rastojanja za analizu površinskih proteina koronavirusa, dolazimo do zaključka da ove dve metode pružaju različite, ali komplementarne uvide u genetske razlike među virusima. Iako su numeričke vrednosti udaljenosti značajno različite između ove dve metode, primećujemo sličnosti u obrascima udaljenosti između površinskih proteina.

Oba skupa rezultata pokazuju slične obrasce bliskosti između određenih koronavirusa. U oba skupa rezultata, površinski proteini iz Human coronavirus 229E i SARS coronavirus Tor2, kao i Bovine coronavirus i Human coronavirus OC43, imaju tendenciju da budu međusobno bliži nego što su u odnosu na površinski protein iz Middle East respiratory syndrome-related coronavirus.

Korišćenje oba pristupa, p-adičnog i edit rastojanja, omogućilo nam je da dobijemo detaljniji i sveobuhvatniji uvid u genetske sličnosti i razlike među površinskim proteinima koronavirusa. Dok p-adična rastojanja pružaju precizan uvid u lokalne promene koje mogu imati značajan biološki uticaj, edit rastojanja omogućavaju širu analizu ukupnih genetskih razlika. Ovo dovodi do nešto većih vrednosti rastojanja. Ovo dovodi do većih vrednosti jer uzima u obzir sve razlike bez specifičnog fokusa na lokalizovane mutacije.

6 Zaključak

U ovom seminarskom radu istražili smo uticaj p-adičnog, Hamingovog i edit rastojanja na genetski kod koronavirusa analizom rastojanja među proteinima i klasterovanjem. Naši rezultati su pokazali da postoji značajna povezanost između p-adičnog rastojanja i varijacija u sekvencama proteina koronavirusa kao i velika sličnost između p-adičnog rastojanja i edit rastojanja, koje je već standard u bioinformatici.

Ovi nalazi su značajni jer mogu doprineti boljem razumevanju genetskog koda virusa, evolucije virusa i pomoći u razvoju ciljanih terapija.

Međutim, istraživanje ima svoja ograničenja. Naši rezultati su zasnovani na ograničenom uzorku sekvenci. Takođe, korišćene metode analize imaju svoja ograničenja koja bi mogla biti prevaziđena u budućim istraživanjima.

Za buduća istraživanja preporučujemo proširenje uzorka uključivanjem novih vrsta virusa i upotrebu naprednijih algoritama za analizu rezultata. Dalja istraživanja bi mogla dodatno produbiti naše dosadašnje znanje u ovoj oblasti.

Naši rezultati mogu pomoći prilikom razvijanja novih vakcina, personalizovang lečenja i pomoći nam u borbi protiv budućih pandemija.

Razmatrajući p-adicna rastojanja u analizi površinskih proteina koronavirusa, dolazimo do zaključka da ova metoda nosi sa sobom značajne prednosti i potencijal za napredak u bioinformatici.

Jedna od ključnih prednosti p-adičnih rastojanja jeste što uzimaju u obzir lokalne razlike u sekvencama, čime se osigurava da se značajne genetske promene ne previde. Ono najveći značaj pridaje razlici na prvom, a najmanji razlici na poslednjem nukleotidu u kodonu što je u saglasnosti sa standardnim genetskim kodom (Slika 9). Ovo je posebno važno u slučaju virusa, koji su poznati po brzim evolutivnim promenama i mutacijama kako bi izbegli imunološki odgovor domaćina.

Uzimajući u obzir sve ove faktore, p-adična rastojanja imaju potencijal da postanu važan alat u bioinformatici za analizu genoma, posebno u kontekstu evolucije virusa i razvoja terapija. Njihova sposobnost da pruže detaljan uvid u genetske razlike na lokalnom nivou čini ih dragocenim resursom za istraživače koji žele da dublje razumeju strukturu i funkciju genoma.

Zaključili smo da p-adično rastojanje može da parira tradicionalnom edit rastojanju prilikom poređenja sekvenci i da p-adično rastojanje može biti vredan alat u bioinformatici.