



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	16
5.1	Edit rastojanje	16
5.2	Hijerarhijsko sakupljajuće klasterovanje	17
5.3	Vizuelizacija klastera	19
5.4	Poređenje rezultata površinskih proteina sa p-adičnim rastojanjem	23
6	Zaključak	25
	Reference	26

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, 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 njihovog edit rastojanja. Ovakav pristup omogućava otkrivanje grupa sličnih sekvenci koje mogu ukazati na zajedničko poreklo ili funkcionalnu povezanost.

U implementaciji i analizi, koristićemo sledeće biblioteke: Biopython za rad sa biološkim podacima [1], scikit-learn za primenu hijerarhijskog klasterovanja i PCA [2], NumPy za numeričke proračune [3], pandas za obradu i analizu podataka [4], matplotlib za kreiranje grafikona [5] i seaborn za unapređene statističke vizualizacije [6].

Potrebne biblioteke možete instalirati pokretanjem sledeće komande iz terminala:

```
pip install biopython scikit-learn numpy pandas matplotlib seaborn
```

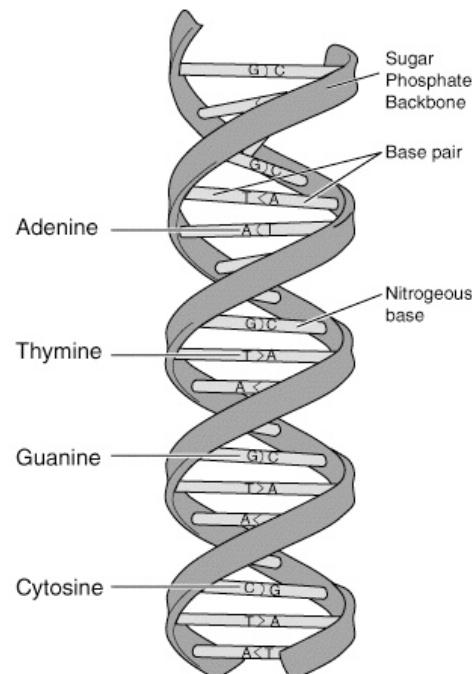
Ceo kod istraživanja dostupan je na GitHub platformi [7, 8].

2 Genetski kod

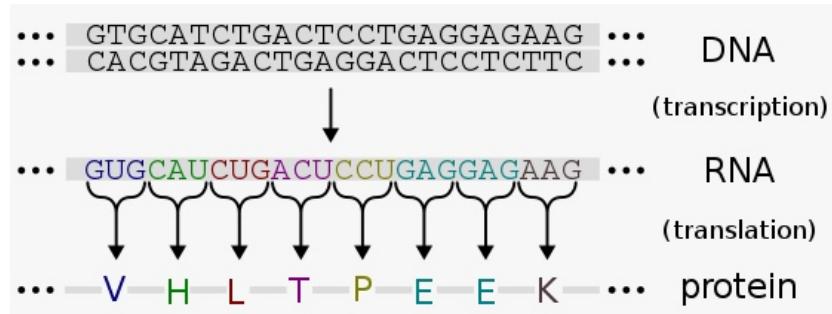
Sveukupna genetska informacija jednog organizma naziva se genom, a sva genetska informacija nalazi se u molekulu DNK. Svaki funkcionalni region molekula DNK naziva se gen. Gen je fizička i funkcionalna jedinica nasleđivanja koja prenosi naslednu poruku iz generacije u generaciju, a čini ga celovit deo 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.

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 najde 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 1: DNK molekul



Slika 2: Proces prevodenja sekvene DNK molekula u protein

Genetski kod je eksperimentalno otkriven sredinom 1960-ih [9], što nam omogućava da razumemo kako funkcioniše u praksi, ali njegovo teorijsko razumevanje i dalje nije potpuno. Skoro sve žive vrste koriste isti genetski kod, poznat kao standardni genetski kod, dok samo mali broj organizama pokazuje male varijacije u ovom kodu.

	U	C	A	G					
U	UUU UUC UUA UUG	Phe Leu	UCU UCC UCA UCG	Ser	UAU UAC UAA UAG	Tyr Stop Stop	UGU UGC UGA UGG	Cys Stop Trp	U C A G
C	CUU CUC CUA CUG	Leu	CCU CCC CCA CCG	Pro	CAU CAC CAA CAG	His Stop	CGU CGC CGA CGG	Arg	U C A G
A	AUU AUC AUA AUG	Ile	ACU ACC ACA ACG	Thr	AAU AAC AAA AAG	Asn Stop	AGU AGC AGA AGG	Ser Stop	U C A G
G	GUU GUC GUA GUG	Val	GCU GCC GCA GCG	Ala	GAU GAC GAA GAG	Asp Glu	GGU GGC GGA GGG	Gly	U C A G

First position (5' end)

Third position (3' end)

Amino acid names:

- Ala = alanine Gln = glutamine Leu = leucine Ser = serine
- Arg = arginine Glu = glutamate Lys = lysine Thr = threonine
- Asn = asparagine Gly = glycine Met = methionine Trp = tryptophan
- Asp = aspartate His = histidine Phe = phenylalanine Tyr = Tyrosine
- Cys = cysteine Ile = isoleucine Pro = proline Val = valine

Slika 3: Standardni genetski kod

3 Preprocesiranje podataka

Nakon što smo se upoznali sa osnovama genetskog koda i procesima 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 Biotehnoške Informacije (NCBI) [10].

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 regije koji mogu biti prepisani i prevedeni u proteinske sekvene.

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, što nam je pomoglo u određivanju pozicija potencijalno kodirajućih delova RNK.

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

    # Pronalaženje 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 pronalaženje ORF-ova

Nakon identifikacije ORF-ova, možemo prevesti RNK sekvence u sekvence aminokiselina koristeći standardni genetski kod. Zatim upoređujemo dobijene proteinske sekvene sa referentnim proteinskim sekvencama kako bismo identifikovali koje sekvene 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)

    # Prevođenje ORF-ova u sekvene 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] = genom[virus][start:end].transcribe()

```

Slika 5: Poređenje dobijenih proteinskih sekvenci sa referentnim proteinima

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 dopuna 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

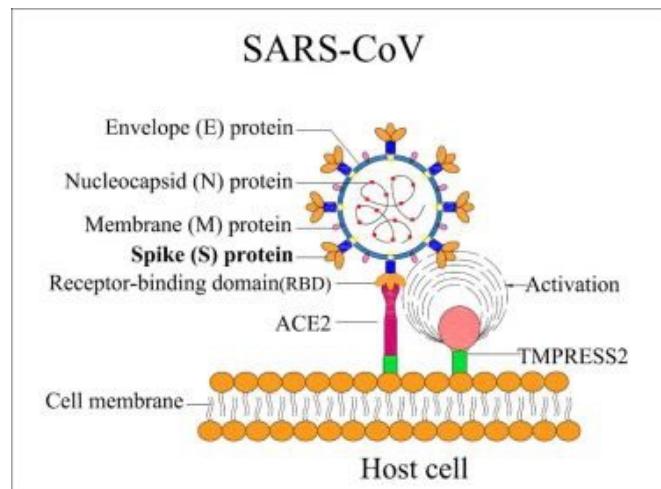
Ceo kod korišćen za preprocesiranje podataka možete pogledati u datoteci `preprocesiranje.ipynb`. Rezultati koji su dobijeni su smešteni u datoteći `rnk.fasta` koja se nalazi u direktorijumu `proteini`, poddirektorijumu `podaci`.

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.

Na primer, spike protein SARS-CoV-1 virusa se vezuje za ACE2 receptor na ćelijama domaćina. ACE2 receptor se široko nalazi na površini ljudskih ćelija, posebno u respiratornom traktu, što objašnjava visoku stopu infektivnosti ovog virusa. Sa druge strane, MERS-CoV koristi DPP4 receptor za ulazak u ćelije domaćina, što rezultira drugačijom patogenezom i manjom 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, neki koronavirusi, poput BCoV i HCoV-OC43, koriste hemaglutinin-esterazu (HE) protein za ulazak u ćelije domaćina. Ova specifičnost omogućava BCoV-u da inficira goveda, dok HCoV-OC43 inficira ljude.



Slika 7: Vezivanje SARS-CoV za ćeliju domaćina

Korišćenjem p-adičnog i Hamingovog rastojanja, istražićemo razlike i sličnosti između ovih proteina na molekularnom nivou. Kod koji koristimo se nalazi u datoteci `povrsinski_proteini_rastojanja.ipynb`.

4.1 P-adično rastojanje

P-adično rastojanje je ključni koncept u našoj analizi površinskih proteina i predstavlja matematički alat za modeliranje genetskog koda. Ovaj pristup, predložen 2006. godine od strane Dragovića i Dragovića [11], istražuje bliskost kodona koji kodiraju istu aminokiselinu u ultrametričnom prostoru, uvodeći p-adični prostor kodona za $p = 5$ i $p = 2$.

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,

$$x = x_0 + x_1p + x_2p^2 + \dots + x_kp^k \equiv x_0x_1x_2\dots x_k$$

,

$$y = y_0 + y_1p + y_2p^2 + \dots + y_kp^k \equiv y_0y_1y_2\dots y_k$$

gde su $x_i \in \{0, 1, \dots, p - 1\}$ i $y_i \in \{0, 1, \dots, p - 1\}$ cifre brojeva u odgovarajućoj p -adičnoj bazi, a \equiv je oznaka za drugačiji zapis broja x , odnosno y . Tada se udaljenost između x i y računa kao:

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

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, to 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 $A + U = C + G = 5$.

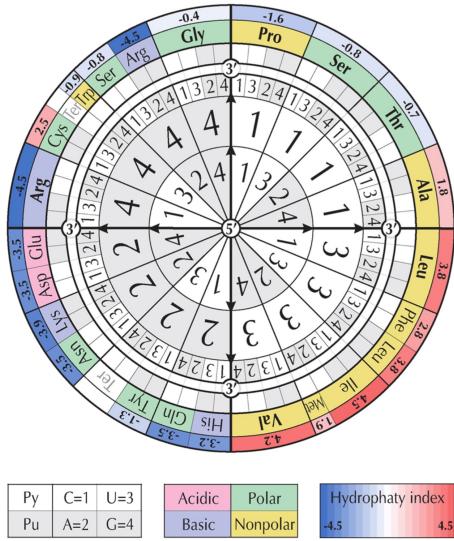
Na kraju, uzete su sledeće vrednosti za nukleotide:

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

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: Dodata p-adičnih brojeva kodonima



Slika 9: Standardni genetski kod preko p-adičnih brojeva

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

$$d_5(a, b) = \sum_{i=0}^n |a_{3i+1}a_{3i+2}a_{3i+3} - b_{3i+1}b_{3i+2}b_{3i+3}|_5,$$

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

```
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 sekvene 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. Postoje dva glavna pristupa primene 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 istih aminokiselina, što može imati uticaj na strukturu, ali ne i na funkciju proteina. Upoređivanjem aminokiselina možemo bolje razumeti funkcionalne posledice genetiskih varijacija.

```
def hamingovo_rastojanje(a, b): # a i b su sekvene 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 da koristimo metode poput p-adičnog i Hamingovog rastojanja , sekvene koje se porede moraju da budu iste dužine. U ovom radu korišćena su dva načina dopunjavanja kraćih sekvenci:

1. dopunjavanje najučestalijim kodonom: kraća sekvenca se dopunjava najučestalijim kodonom koji se pojavljuje u njoj do dužine duže sekvene.
2. dopunjavanje najnajučestalijim nukleotidom: kraća sekvenca se dopuna najučestalijim nukleotidom koji se pojavljuje u njoj do dužine duže sekvene.

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.

Dobijeni rezultati računanjem p-adičnog rastojanja:

P-adična 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

Slika 12: P-adična rastojanja nakon dopune najučestalijim kodonom

P-adična 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

Slika 13: P-adična rastojanja nakon dopune najučestalijim nukleotidom

P-adična 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

Slika 14: P-adična rastojanja nakon dopune nulama

Dobijeni rezultati 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

Slika 15: Hamingova rastojanja nakon dopune najučestalijim kodonom

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

Slika 16: Hamingova rastojanja nakon dopune najučestalijim nukleotidom

Dobijeni rezultati 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

Slika 17: Hamingova rastojanja nakon dopune najučestalijim kodonom

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

Slika 18: Hamingova rastojanja nakon dopune najučestalijim nukleotidom

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 funkcionalisanja ovih metrika. P-adična rastojanja uzimaju u obzir poziciju prve razlike u nukleotidima unutar kodona, dok Hamingova rastojanja jednostavno broje pozicije na kojima su kodoni različiti, bez obzira na važnost tih razlika.

Ove razlike ukazuju na to da Hamingova rastojanja pružaju grublju sliku sličnosti, dok p-adična rastojanja daju detaljniji uvid u specifične evolucione promene.

Kada poređimo 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.

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 većina ostalih beta-koronavirusi koriste ACE2 ili slične receptore, 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 meri razliku između dva niza karaktera određujući minimalan broj operacija (umetanja, brisanja, zamene) potrebnih za pretvaranje jednog niza u drugi. Dinamičko programiranje omogućava efikasno pronađenje najkraćeg niza operacija, precizno mereći razlike čak i između sekvenci različitih dužina.

U bioinformatičkim istraživanjima, edit rastojanje je ključan alat za razumevanje genetskih promena, evolutivnih odnosa i funkcionalnih sličnosti među različitim organizmima ili sekvencama.

```
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 19: Funkcija za računanje edit rastojanja

Kod za računanje edit rastojanja se nalazi u `edit_rastojanje.ipynb` datoteci, a dobijeni rezultati zabeleženi su u `edit_rastojanja.txt` datoteci.

Ukupan broj referentnih proteina za naših pet koronavirusa iznosi 56. Za svaku RNK sekvencu koja kodira protein, vršimo poređenje sa svakom drugom, što dovodi do ukupno 1540 kombinacija za računanje edit rastojanja.

S obzirom na to da nekoliko sekvenci ima dužinu veću od 20000 nukleotida, proces računanja edit rastojanja bio je dugotrajan i zahtevan i završen je nakon 14 sati. Zbog ove složenosti, ne preporučujemo da pokrećete ovaj kod.

5.2 Hijerarhijsko sakupljajuće klasterovanje

Hijerarhijsko sakupljajuće klasterovanje (engl. Hierarchical Agglomerative Clustering) je tehnika 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 koristeći edit rastojanje kao meru sličnosti između svakog para proteina, primenili smo algoritam hijerarhijskog sakupljajućeg klasterovanja na osnovu ove matrice kako bismo grupisali proteine u klastere.

```
def Klasterovanje(edit_rastojanja, broj_klastera):
    klasterovanje = AgglomerativeClustering(n_clusters=broj_klastera, linkage='average', metric='precomputed')
    klasteri = klasterovanje.fit_predict(edit_rastojanja)

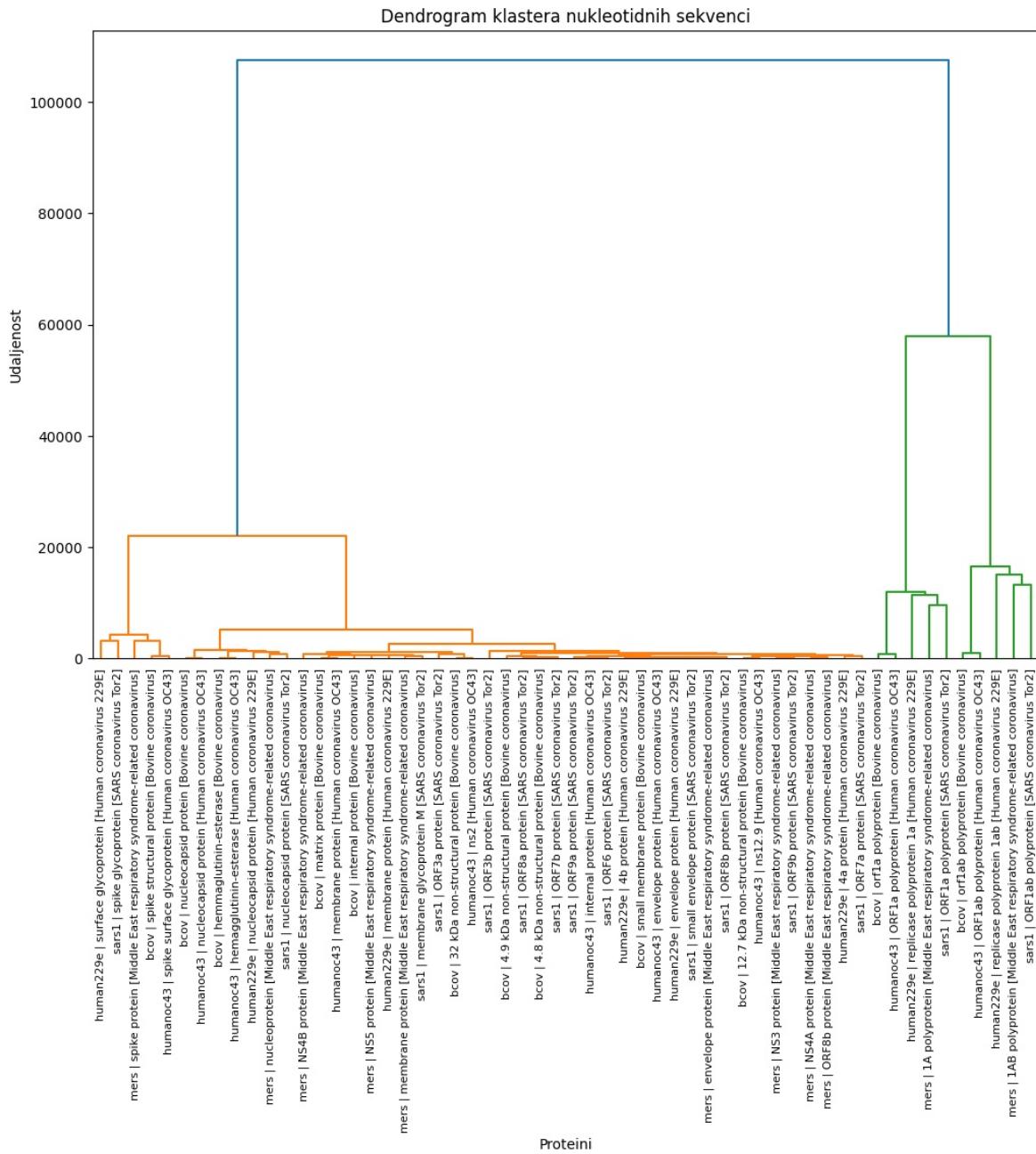
    return klasteri

def nacrtaj_dendogram(edit_rastojanja, proteini_nazivi):
    linkage_matrix = linkage(matrica_rastojanja, method='average')

    # IsCRTavanje dendrograma
    plt.figure(figsize=(12, 8))
    dendrogram(linkage_matrix, labels=list(proteini_nazivi), leaf_font_size=8)
    plt.title('Dendrogram klastera nukleotidnih sekvenca')
    plt.xlabel('Proteini')
    plt.ylabel('Udaljenost')
    plt.show()
```

Slika 20: Funkcije za klasterovanje podataka i crtanje dendograma

Ceo kod se nalazi u datoteci `hijerarhijsko_klasterovanje.ipynb`.



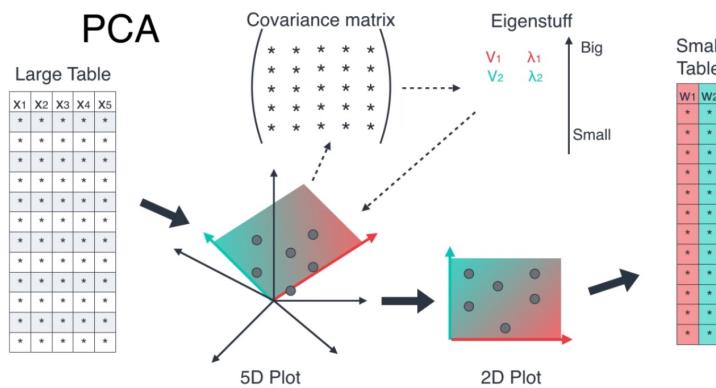
Slika 21: Dendrogram klastera

Nakon detaljne analize dendrograma, zaključili smo da postoji podela na tri glavne grupe, što je rezultiralo odlukom da koristimo tri klastera kao optimalan broj.

5.3 Vizuelizacija klastera

Kako bismo dobijene klasterovane podatke vizuelizovali i bolje ih razumeli, primenili smo analizu glavnih komponenti (engl. Principal Component Analysis - PCA) na matricu edit rastojanja.

PCA je tehnika smanjivanja dimenzionalnosti koja transformiše originalne podatke u novi skup linearno nezavisnih promenljivih, poznatih kao glavne komponente, koje zadržavaju maksimalnu varijansu podataka.

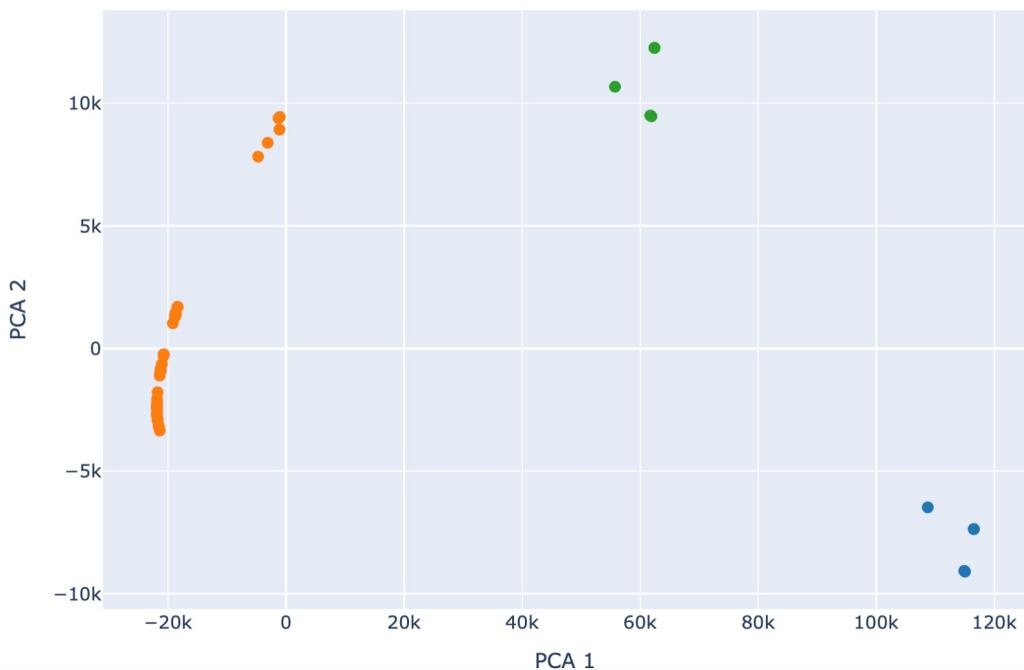


Slika 22: Princip smanjenja dimezionalnosti podataka

Primena PCA na matricu edit rastojanja omogućila je smanjenje dimenzionalnosti na dve glavne komponente, čime smo postigli dvodimenzionalnu reprezentaciju podataka. Ova dvodimenzionalna matrica omogućava nam vizualizaciju klastera u dvodimenzionalnom prostoru i lakšu identifikaciju obrasca i grupisanja među proteinima.

Interaktivnu vizuelizaciju klastera možete pogledati u datoteci `klasteri.html`, a sliku istog u nastavku.

Vizuelizacija klastera



Slika 23: Vizuelizacija klastera

Na osnovu vizuelizacije klastera, možemo zaključiti sledeće:

- U zelenom klasteru nalaze se ORF1a polyprotein koronavirusa SARS-CoV-1, BCoV i HCoV-OC43, 1A polyprotein MERS-CoV-a i replicase polyprotein 1a virusa HCoV-229E. Ovi proteini su deo ne-struktturnih proteina koji su ključni za replikaciju i sintezu virusnog RNK genoma i proteina.
- Plavi klaster obuhvata proteine: ORF1ab i 1AB polyproteine i replicase polyprotein 1ab. Kao i proteini u zelenom klasteru, ovi proteini su ne-strukturni i osnovni 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 i struktura, ali njihovo grupisanje zajedno ukazuje na određene zajedničke karakteristike.

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 sugerije da, uprkos funkcionalnoj raznolikosti, dele određene evolutivne ili strukturne karakteristike koje ih čine sličnijima jedne drugima u poređenju sa proteinima iz drugih klastera.

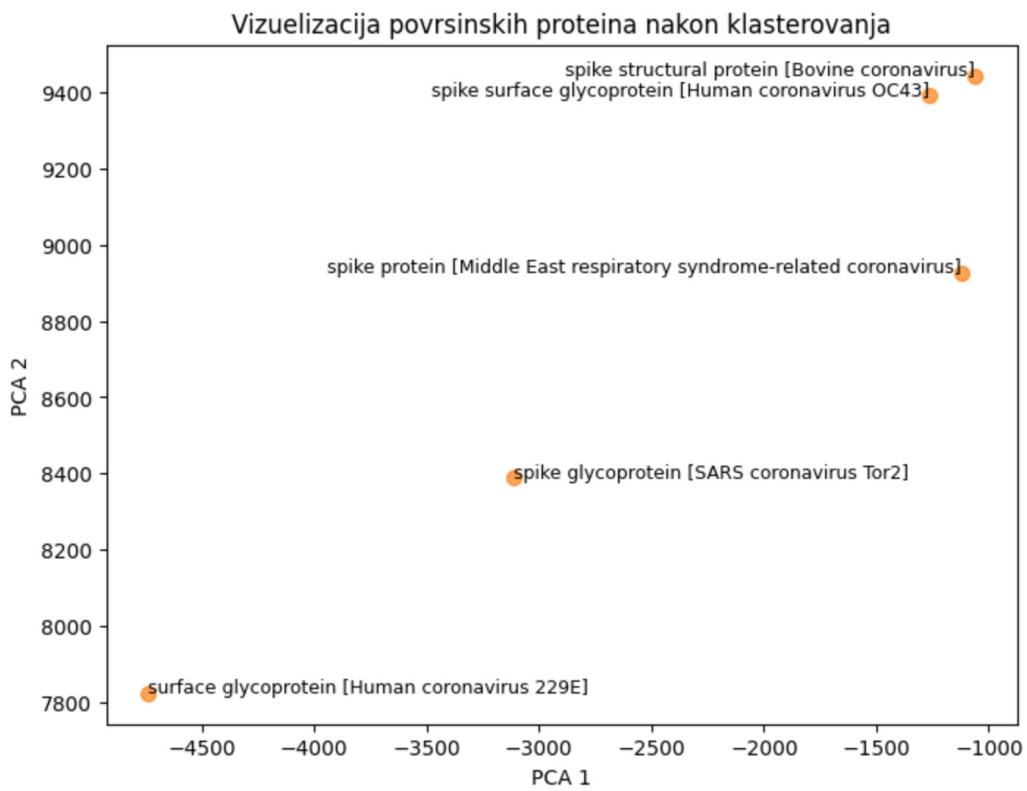
Posebno je važno napomenuti da su proteini koji imaju slične funkcije, iako potiču od različitih koronavirusa, međusobno bliži nego proteini sa različitim funkcijama.

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 izdvajao podklaster površinskih proteina, što ukazuje na to da zaista postoji stvarna biološka osnova za tu grupaciju.

Osim toga, dendrogram potvrđuje grupisanje površinskih proteina u određen podklaster (videti sliku 21), što je dodatna podrška postojanju takve grupacije.

Kada bismo “uvećali” podklaster u kome se nalaze površinski proteini, mogli bismo detaljnije videti njihovu raspodelu. Međutim, ovo je raspored prikazan na matrici na koju je primenjena PCA, čime je smanjena njena dimenzija sa 56x56 na 56x2.



Slika 24: Vizuelizacija podklastera

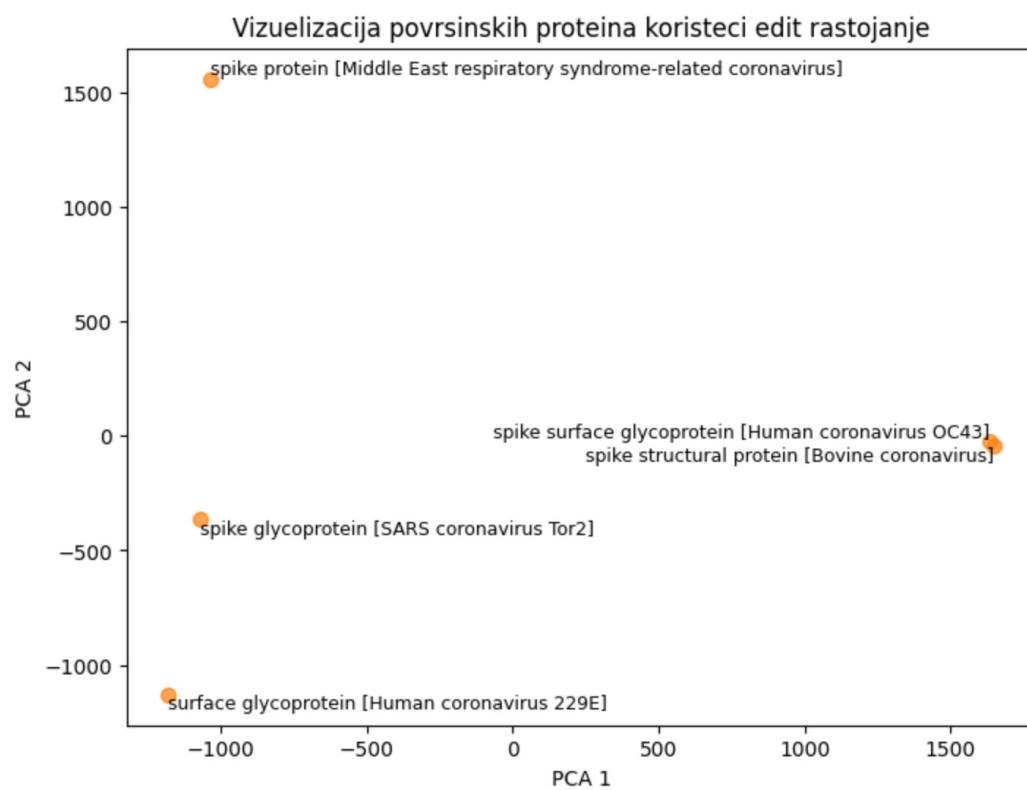
Dobijeni rezultati računanjem edit rastojanja (prikaz samo za površinske proteine):

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

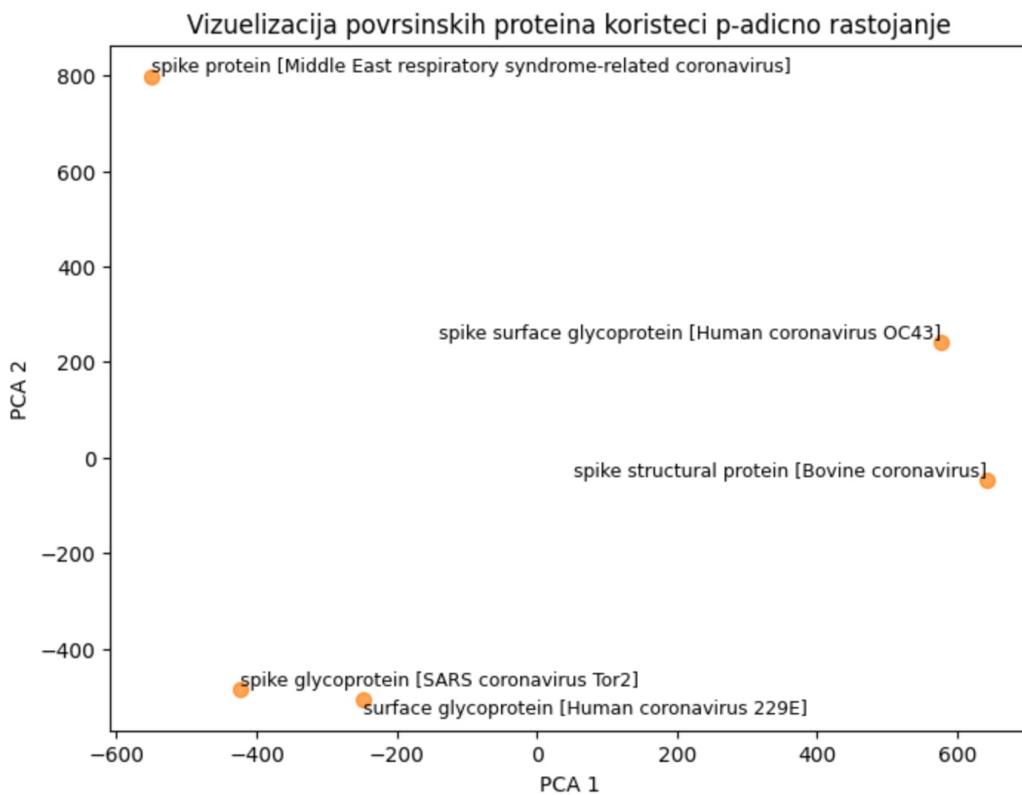
Slika 25: Edit rastojanja

5.4 Poređenje rezultata površinskih proteina sa p-adičnim rastojanjem

Kako bismo bolje i preciznije mogli vizuelno da uporedimo rezultate klasificiranja i računanja edit rastojanja sa p-adičnim rastojanjem, napravili smo novu matricu koja sadrži samo edit rastojanja između površinskih proteina. Na tako dobijenu matricu smo primenili PCA radi vizuelizacije. Takođe, PCA smo primenili i na jednu od matrica p-adičnih rastojanja (u datoteci `povrsinski_proteini_rastojanja.ipynb`). Ovim smo obezbedili smanjenje dimenzionalnosti oba skupa podataka sa 5x5 na 5x2 i dobili sličnije iskrivljene slike zbog PCA.



Slika 26: Vizuelizacija površinskih proteina koristeći edit rastojanje



Slika 27: Vizuelizacija površinskih proteina koristeći p-adično rastojanje

Analizirajući dobijene rezultate, primećujemo da iako su numeričke vrednosti udaljenosti značajno različite između ove dve metode, obrasci udaljenosti među površinskim proteinima su slični.

Oba skupa rezultata pokazuju slične obrasce među određenim koronavirusima. Površinski proteini Human coronavirus 229E i SARS coronavirus Tor2, kao i Bovine coronavirus i Human coronavirus OC43, pokazuju tendenciju da budu međusobno bliži nego što su u odnosu na površinski protein iz Middle East respiratory syndrome-related coronavirus.

Izbor odgovarajuće metode zavisi od specifičnih zahteva istraživanja i potreba analize. KlastEROVANJE na osnovu edit rastojanja može pružiti korisne uvide u strukturu sličnosti kada je potrebno sagledati globalne genetske razlike. Sa druge strane, p-adična rastojanja su od velike važnosti kada je potrebno detaljno istražiti lokalne mutacije koje mogu imati značajan biološki uticaj, na primer, na funkciju proteina.

6 Zaključak

U ovom seminarском раду istraživali smo uticaj p-adičnog rastojanja na genetski kod koronavirusa i poredili dobijene rezultate sa rezultatima korišćenjem već tradicionalnih metoda u bioinformatici.

Hamingovo rastojanje na kodonima pruža brzu i relativno preciznu analizu proteina, dok klasterovanje na osnovu edit rastojanja omogućava sporiju, ali detaljniju analizu, fokusiranu na strukturne razlike.

Ključna prednost p-adičnih rastojanja je u tome što p-adična metrika pridaje najveći značaj razlici na prvom, a najmanji na poslednjem nukleotidu u kodonu, što je u skladu sa standardnim genetskim kodom. Zbog toga ova metoda omogućava dublje razumevanje veze između kodona i aminokiselina, što je od suštinskog značaja za proučavanje funkcionalnih karakteristika proteina.

Na osnovu ovih nalaza, možemo zaključiti da p-adično rastojanje može postati ključan alat u bioinformatici zbog svoje sposobnosti da otkrije lokalne genetske razlike i pruži dublje razumevanje strukture genoma.

Reference

- [1] Cock, P.J.A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), 1422-1423. <https://doi.org/10.1093/bioinformatics/btp163>.
- [2] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. <https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>.
- [3] Harris, C.R., et al. (2020). Array programming with NumPy. *Nature*, 585, 357-362. <https://doi.org/10.1038/s41586-020-2649-2>.
- [4] McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 51-56. <https://doi.org/10.25080/Majora-92bf1922-00a>.
- [5] Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 90-95. <https://doi.org/10.1109/MCSE.2007.55>.
- [6] Waskom, M., et al. (2020). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021. <https://doi.org/10.21105/joss.03021>.
- [7] Jelisaveta Gavrilović. (2024). IP2-bioinformatics. <https://github.com/jelisavetagavrilovic/IP2-bioinformatics>.
- [8] Marko Paunović. (2024). IP2-Bioinformatics. <https://github.com/MarkoPaunovic14/IP2-Bioinformatics>.
- [9] Bernfield, M.R., & Nirenberg, M.W. (1965). RNA codewords and protein synthesis. *Science*, 147(3657), 479-484. <https://doi.org/10.1126/science.147.3657.479>.
- [10] National Center for Biotechnology Information. (s.d.). Databases. <https://www.ncbi.nlm.nih.gov/genbank/>.
- [11] Dragovich, B., Dragovich, A., 2006. A p-adic model of DNA sequence and genetic code. <https://arxiv.org/abs/q-bio/0607018v1>