

**UNIVERZITET U BEOGRADU
TEHNOLOŠKO-METALURŠKI FAKULTET**

ZAVRŠNI RAD

**BIOINFORMATIČKA ANALIZA UTICAJA
VARIJABILNOG REGIONA SEKVENCE 16S
rRNA KODIRAJUĆEG GENA NA
REZULTAT IDENTIFIKACIJE BAKTERIJA**

Sonja Jović

Beograd, septembar 2021.

DATUM ODBRANE:_____

OCENA RADA:_____

OCENA ODBRANE:_____

SREDNJA OCENA:_____

MENTOR:

Dr Mirjana Rajilić – Stojanović, docent

ČLAN KOMISIJE:

Dr Suzana Dimitrijević – Branković, redovni profesor

KANDIDAT:

Sonja Jović

Sadržaj

Sažetak	1
Uvod	2
1. Teorijski deo	3
1.1 Tipovi ćelija mikroorganizama	3
1.2 Klasičan pristup identifikaciji bakterija	3
1.3 Sistematika mikroorganizama	5
1.4 Karl Vouz i filogenetsko stablo	5
1.5 Moderan pristup identifikaciji bakterija	6
1.6 Bioinformatička obrada podataka	9
1.6.1 Eliminisanje nepravilnosti	9
1.6.2 Formiranje klastera (OTU)	10
1.7 Bioinformatički alati	11
1.7.1 QIIME	12
1.7.2 UPARSE	12
1.7.3 mothur	12
1.7.4 MEGAN CE	12
1.7.5 DADA2	13
2. Eksperimentalni deo	14
2.1 Materijal i metode	14
2.1.1 Mothur naredbe korišćene u analizi	15
2.1.2 Vizualizacija podataka	17
2.2 Rezultati i diskusija	18
2.2.1 Taksonomska identifikacija korišćenjem SILVA referentne baze	18
2.2.2 Taksonomska identifikacija korišćenjem Greengenes referentne baze	24
Zaključak	28
Literatura	29

Sažetak

U ovom završnom radu izvršena je identifikacija bakterija na osnovu sekvence hipervarijabilnih regiona u okviru gena koji kodira sintezu 16S ribozomalnu RNK (rRNK). U ovom radu korišćene su sekvence ukupno 2533 bakterija od kojih 1216 potiče od kultivisanih vrsta koje nastanjuju intestinalni trakt čoveka, dok preostalih 1317 potiče iz bakterija intestinalnog trakta koje su detektovane u studijama baziranim na sekvenci 16S rRNK. Za sve korišćene sekvence određena je taksonomska pozicija korišćenjem prilagođene SILVA baze sekvenci 16S rRNK. Sekvenca 16S rRNK sadrži devet hipervarijabilnih regiona (V1-V9). Analizirani set podataka imao je nekompletne sekvence pa je bioinformatička analiza urađena na osnovu sekvence regiona od V1 do V8.

Analiza je izvršena bioinformatičkom metodom, korišćenjem *mothur* alata za obradu podataka, dok su se za dodeljivanje taksonomije koristile SILVA referentna baza i Greengenes. Ekstrakcija tabelarnih podataka i vizuelizacija izvršeni su u programskom jeziku *Python*. Odstupanja od referentnih podataka javljaju se zbog kvaliteta i dužine ulaznih sekvenci, stepena varijabilnosti regiona, kao i izbora referentne baze.

Uvod

Bioinformatika je naučna oblast koja spaja molekularnu biologiju, genetiku, matematiku i statistiku i informatiku. Podaci velikih razmera obrađuju se korišćenjem programskih alata, a biološki problemi sagledavaju se sa kompjuterske tačke gledišta [1].

Bioinformatika se bavi prikupljanjem, obradom i analizom bioloških podataka u koje se najčešće ubrajaju sekvence DNK i sekvence aminokiselina za određivanje funkcije gena i proteina, uspostavljanje evolucionih odnosa i predviđanje trodimenzionalnih oblika proteina [2].

Glavni alati rada jednog bioinformatičara uključuju softverske programe i internet. Osnovna aktivnost – analiza sekvenci DNK i proteina vrši se preko dostupnih softvera namenjenih za tu primenu i dostupnih baza podataka. Softveri se nalaze na internetu u vidu javno dostupnih projekata (eng. *Open source*) što znači da korisnici mogu besplatno da pristupe programu. Uvidom u kod programa korisnici učestvuju u daljem unapređenju i poboljšanju koda. Kompjuterske baze podržavaju brzu asimilaciju i sadrže upotrebljive formate za softversku obradu i upravljenje podacima. Pristup datim bazama je besplatan. Međutim, zbog raznovrsne prirode novih podataka ne postoji jedinstvena, sveobuhvatna baza za pristup ovim informacijama [3].

Analiza sekvence gena koji kodira 16S ribozomalne RNK (rRNK) koristi se za identifikaciju mikroorganizama zbog svojih specifičnih osobina:

1. Gen 16S rRNK se sastoji od 9 varijabilnih regiona, koji variraju između vrsta i visoko konzerviranih regiona, koji nisu promenljivi.
2. Ribozomi su ćelijske organele prisutne u svim ćelijama i svim tipovima ćelija i sastoje se od velike i male subjednice. Mala subjednica ribozoma izgrađena je od ribozomalne RNK koja se kodira kao 16S rRNK kod bakterija.

Upravo ove osobine definišu specifičnost gena 16S rRNK i svrstavaju ga u gen markere [4]. Sekvenca 16S rRNK se koristi za preciznu identifikaciju bakterija u kombinaciji sa metodama klasične mikrobiologije i osnov je više molekularnih metoda za analizu kompleksinih ekosistema.

Baze podataka koje se najčešće koriste za identifikaciju bakterija na osnovu 16s rRNK su SILVA, Greengenes i EzBioCloud referentna baza [5].

Glavna prednost ovog načina identifikacije je smanjen uticaj faktora koji utiču na morfologiju i citologiju ćelije, kao što je starost ćelije i spoljašnjih faktora koji definišu posebne laboratorijske uslove za gajenje i odvijanje ćelijskih metaboličkih procesa [6, 7].

U cilju evaluacije preciznosti bioinformatičkih alata u kombinaciji sa različitim bazama sekvenci u ovom radu izvršeno je poređenje načina identifikacije bakterija na osnovu različitih hipervarijabilnih regiona sekvence 16S rRNK, već identifikovanih bakterija mikrobiote [8] bioinformatičkom analizom programskim alatom *mothur* [9] korišćenjem SILVA i Greengenes baza sekvenci.

1. Teorijski deo

1.1 Tipovi ćelija mikroorganizama

Mikroorganizmi su grupa živih bića vidljiva samo mikroskopom. Golim okom se mogu videti grupe mikroorganizama u obliku kolonija, taloga ili zamućenja. Kroz istoriju utvrđeno je da se mikroorganizmi nalaze svuda oko nas. Najveći broj mikroorganizama su jednoćelijski organizmi, ali se mogu pronaći i višećelijski u obliku slabo diferenciranih tkiva [10].

Morfologija i citologija su oblasti mikrobiologije koje se bave proučavanjem izgleda ćelije i njenom unutrašnjom građom. Intenzivan razvoj ovih oblasti počinje sa otkrićem elektronskog mikroskopa, kada je na osnovu građe ćelije utvrđeno da postoje dva tipa ćelija prokariotski i eukariotski tip. Osnovna razlika ova dva tipa ćelija je u postojanju organizovanog jedra. Eukariotski tip ćelije, pored organizovanog jedra u svojoj unutrašnjosti sadrži i veći broj specijalizovanih organela (*tabela 1*) [4, 10].

Ribozomi su prisutni u oba tipa ćelija. Veličina ribozoma se razlikuje između mikroorganizama i iznosi 70S (S – Svedbergova jedinica) kod bakterija i arheja i 80S kod eukariota. Ribozomi su izgrađeni od dve subjedinice (male i velike). Svaka subjedinica predstavlja kompleks ribozomalne RNK i proteina. Kod bakterijskih ćelija ribozom je izgrađen od 50S (velika subjedinica) i 30S (mala subjedinica). Subjedinica 30S formirana je od kompleksa proteina i 16S ribozomalne rRNK. Eukariotske ćelije su izgrađene od 40S (male subjedinice) i 60S (velike subjedinice) i njihovu malu subjedinicu formira skup proteina i 18S ribozomalne rRNK [4, 11].

Tabela 1: Karakteristike prokariotske i eukariotske ćelije [10]

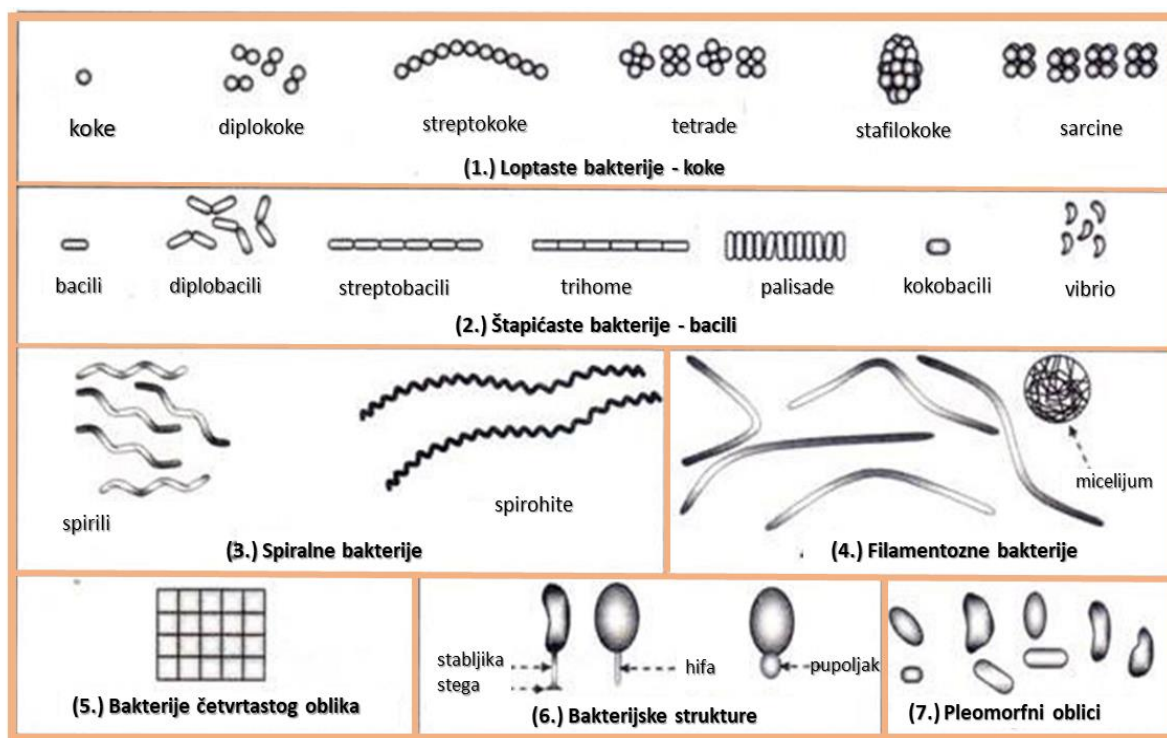
Karakteristike ćelije	Prokariotska	Eukariotska
Veličina (prečnik)	0,3-2 µm	2-20 µm
Genetičke sturkture		
Lokacija	Nukleoid	Jedro
Jedarna membrana	/	+
Broj hromozoma	1-4 (svi identični)	>1 (svi različiti)
Citoplazmatična sturktura		
Ćelijski zid	Ista hemijska jedinjenja	Ako postoji, sastav se razlikuje među organizmima
Citoskelet	+	+
Organele	/	+
Ribozomi	70S	80S

1.2 Klasičan pristup identifikaciji bakterija

Identifikacija mikroorganizama tradicionalnim metodama predstavlja identifikaciju prema njihovim osobinama kao što su oblik, veličina ćelija i kolonija, miris i boja. Veličina bakterija varira od 0,2 do 10 µm. Postoje tri osnovna oblika [10]:

- sferni oblik – koke,
- izduženi štapčasti oblik – bacili i
- spiralni oblik – spirale.

Većina bakterija je monomorfna što znači da prilikom razmnožavanja zadržava svoj oblik i građu. Ovakav tip bakterija je samim tim i lakše identifikovati. Međutim postoje i bakterije koje nemaju tačno definisan oblik pa se u čistim kulturama mogu javiti u različitim veličinama i oblicima, takve bakterije su pleomorfne (*slika 1*) [12].



Slika 1: Oblici bakterija [13]

Selektivno bojenje bakterija obezbeđuje bolji uvid u morfologiju ćelija. Boje poseduju takve karakteristike da se mogu afinitetno vezivati za ciljane delove ćelije. Najčešće korišćene metode su acidorezistentno bojenje i bojenje po Gramu. Bojenje po Gramu se zasniva na razlici u građi ćelijskog zida i činjenici da se bakterije sa spoljašnjim peptidoglukanskim slojem i tejhonom kiselinom stabilno boje. Ovo su takozvane Gram pozitivne bakterije. Bojenje omogućava da se, osim podele na Gram pozitivne i Gram negativne bakterije, utvrde veličina, oblik ćelije, kao i postojanje, pozicija i oblik spora [12].

Pored morfoloških razlika prema kojima se razvrstavaju bakterije, identifikacija se može izvršiti i uz pomoć biohemijskih testova. Posmatranjem ponašanja ćelija i praćenjem njihovog rasta u posebno definisanim medijumima bakteriju je moguće biohemijski okarakterisati i bliže definisati. Određuje se mogućnost fermentacije različitih supstrata, kao što su glukoza, laktoza ili manitol, zatim se analiziraju proizvodi metabolizma - da li dolazi do proizvodnje gasa i/ili kiseline koji se detektuju tačno definisanim indikatorima. Biohemijskim testovima može se ispitati i aktivnost enzima kao što su oksidaze, lipaze, ureaze i drugi [12].

Pored biohemijskih testova za identifikaciju bakterija mogu da se koriste i serološki testovi kojima se određuje stvaranje antitela na dejstvo antigena. Antigen može biti deo bakterijske ćelije koji će u ispitivanom uzorku izazvati pojavu antitela, ali može biti i cela ćelija [12].

Tradicionalno se koriste identifikacioni ključevi koji predstavljaju niz dijagnostičkih i diskriminativnih testova i šema kojima se bliže određuje taksonomska grupa. Prilikom određivanja viših taksonomskih grupa koriste se postojanije – morfološke karakteristike, dok se prilikom određivanja nižih grupa koriste biohemijski testovi. Ovakav pristup identifikaciji i klasifikaciji bakterija zahteva značajne resurse i vreme za analize. Neophodno je obezbediti rast bakterija van njihovog prirodnog okruženja uspostavljanjem ispravnih parametara stanja. Za preciznija određivanja vrste potrebno je ponoviti testove u većem broju uz variranje parametara stanja, na osnovu čega se određuje ponašanje mikroorganizma u različitim uslovima [10].

1.3 Sistematika mikroorganizama

Sistematika mikroorganizama, kao i svih živih bića bavi se proučavanjem diverziteta organizama, njihovim odnosom i interakcijom. Sistematika povezuje taksonomiju i filogeniju [10].

Filogenija proučava evoluciju živih bića kroz istoriju. Bazirana je na ispitivanju genetskog materijala i njegove promene tokom evolucije. Na osnovu filogenije, osim što dobijamo odgovor zašto je genetski materijal nekog organizma danas takav, možemo da predvidimo i kako će se sekvenca menjati u budućnosti [3].

Taksonomija se bavi proučavanjem klasifikacije, identifikacije i nomenklature živih bića. Klasifikacija se bavi grupisanjem mikroorganizma u taksonomske grupe prema karakteristikama posmatranog mikroorganizma. Identifikacija ima zadatak da prepozna nepoznati mikroorganizam i svrsta ga u poznate taksonomske jedinice, nomenklatura dodeljuje naziv nepoznatom mikroorganizmu. Taksonomske jedinice ili taksoni predstavljaju odgovarajuće homogene grupe mikroorganizama koji su srodni tj. potiču od zajedničkog pretka. Taksonomske jedinice su organizovane u hijerarhijski posebne nivoe u zavisnosti od morfoloških, fizioloških, biohemijskih i genetičkih svojstva: domen (kao najviša jedinica [14]), tip, klasa, red, porodica, rod, vrsta. Što su veće razlike na genetskom nivou to je veća udaljenost povezivanja [15].

Na svakom nižem nivou, taksonomska jedinica uže definiše svojstva, tako da nivo vrste predstavlja najprecizniju klasifikaciju. Vrsta, kod prokariotskog tipa ćelije, se može uže odrediti u vidu soja ukoliko dolazi do određenih odstupanja u genetskom materijalu iste vrste, ili u vidu klonova kada je genetski materijal potpuno identičan. Soj predstavlja genetsko odstupanje jedne vrste, sa visokim stepenom fenotipske sličnosti i kao takav može se svrstati u najnižu taksonomsku jedinicu za primenu u raznim granama industrije [10, 15].

Ocem taksonomije smatra se Karl Line (*Carl Linnaeus*, 1707-1778), koji je uveo binominalni sistem nomenklature i živi svet podelio u dva carstva: biljke i životinje. Prvi je napravio podelu po hijerarhijskim nivoima, a glavne karakteristike razdvajanja bile su pokretnost ćelije i mogućnost vršenja fotosinteze [3, 10].

Pronalaskom elektronskog mikroskopa, sistematika živih bića uvažava dihtomiju prokariota-eukariota [16] i deli živi svet u pet kraljevstva: kraljevstvo *Procaryote* (*Monere*) – u koje ulaze sve prokariote, dok su eukariote svrstane u sledeća četiri kraljevstva: biljke (*Plantae*), životinje (*Animalia*), gljive (*Fungi*) i protisti (*Protista*). Kraljevstvo protista obuhvata jednoćelijske alge i protozoe [10]. Poslednju revoluciju u sistematici živog sveta uveo je Karl Vouz (*Carl Woese*) oslanjajući se na sekvencu ribozomalne RNK [14].

1.4 Karl Vouz i filogenetsko stablo

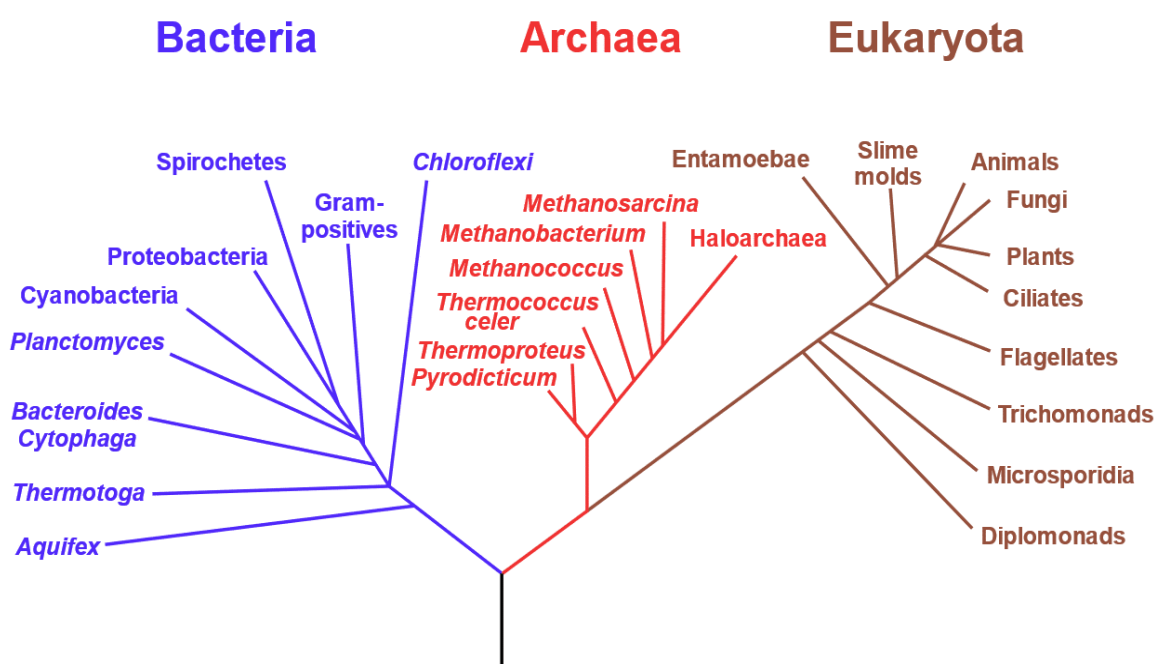
Karl Vouz (*Carl Woese*, 1928-2012) je američki mikrobiolog i biofizičar. Sa svojim saradnicima, Vouz je ustanovio da je dotadšnji način podele živog sveta prema dihtomiji prokariote-eukariote nepravilan. Prema prirodi i tipu ćelija nije moguće direktno porediti prokariote sa eukariotama. Predložio je nove taksonomske grupe više od carstva - domene (*urkingdom*). Fenotipe koji odgovaraju prokariotskom tipu ćelije razvrstao je u grupe *eubacteria* i *archaebacteria*, a fenotipe koji odgovaraju eukariotskom tipu ćelije je svrstao je u grupu *eukarya* [16]. Do ovog zaključka došao je analizom sekvence ribozomalne RNK, za koju je ustanovljeno da predstavlja odličan molekulski hronometar. Ustanovio je razlike sekundarnih struktura molekula RNK male subjedinice ribozoma između grupa *eubacteria* i *archaebacteria*, kao i jednostavni otisak (*eng. sequence signature*) ova dva kraljevstva [4].

Usledila je rekonstrukcija ranije uspostavljene podele živog sveta na pet carstava. Carstvo *Monera* prema svojim karakteristikama značajno odstupa od ostala četiri carstva i kao takvo ne može biti na istom taksonomskom nivou. Predložio je da se nova sistematika bazira na osnovu analize molekulskih sekvenci, dok se klasični pristup sistematizaciji koristi kao potvrda novom [14, 16]. Vouz i njegov tim

su predložili uvođenje nove taksonomske grupe – domen. *Eubacteria* svrstana je u domen bakterija, *archeobacteria* u domen arheja i ćelije koje imaju definisano jezgro (*eukarya*) u domen eukarija [14].

Filogenetsko stablo dobijeno je komparativnom analizom sekvence gena koji kodira sintezu male subjedinice rRNK. Ovi geni su univerzalno raspoređeni, sadrže konzervirane i hipervarijabilne regione čiji se sastav baza menjao različitim brzinama tokom evolucije. Dužine grana proporcionalne su broju tih promena [4, 17].

Vouz je takođe uveo i pojam univerzalnog pretka (*LUCA - last common ancestor*), koji se nalazi u korenu filogenetskog stabla svog živog sveta (slika 2). Osnove u postojanju univerzalnog pretka pronalazi u nivoima organizacija genetičkog materijala živih bića. Po njegovoj pretpostavci genetski materijal pretka imao bi jednostavniji oblik od prokarija. Njegovi geni bili bi odvojene fizičke jedinice koji pored svoje replikativne forme sadrže i funkcionalnu [4]. Ideja da je moguće postojanje ovakve vrste genetskog materijala javila sa otkrićem da RNK ima i katalitička svojstva, odnosno da može sadržati i genetsku i enzimsku funkciju [18].



Slika 2: Univerzalno filogenetsko stablo dobijeno komparativnom analizom rRNK sekvenci [17]

1.5 Moderan pristup identifikaciji bakterija

Jedan od najprimenjenijih načina identifikacije mikroorganizama danas jeste analiza sekvence gena 16S rRNK kod bakterija i arheja, i gena 18S rRNK kod eukarija. Ribozomalna RNK je gradivna jedinica koja ulazi u sastav male subjedinice ribozoma (eng. *SSU – small subunit*). Gen 16S rRNK dobar je filogenetski marker jer se, pored toga što je sveprisutan, sastoji iz konzerviranih (očuvanih), varijabilnih i hipervarijabilnih delova. Sastav konzerviranog dela je takav da se nije mnogo menjao sa vremenom kroz evoluciju. Mutacije u varijabilnim i hipervarijabilnim delovima su bile mnogo češće, pa se analizom sekvence celog gena ili njegovih hipervarijabilnih regiona može odrediti evolutivna udaljenost i povezanost mikroorganizama. Sekvenca hipervarijabilnih regiona pogodna je za taksonomsku identifikaciju. Dužina gena koji kodira 16S rRNK je približno 1550 bp i sastoji se od devet hipervarijabilnih regiona (V1 – V9) [19, 20]. Istraživanja pokazuju da se analizom sekundarne strukture 16S rRNK može detaljnije istražiti njena funkcija kao i da se može osigurati pozicija homologija u

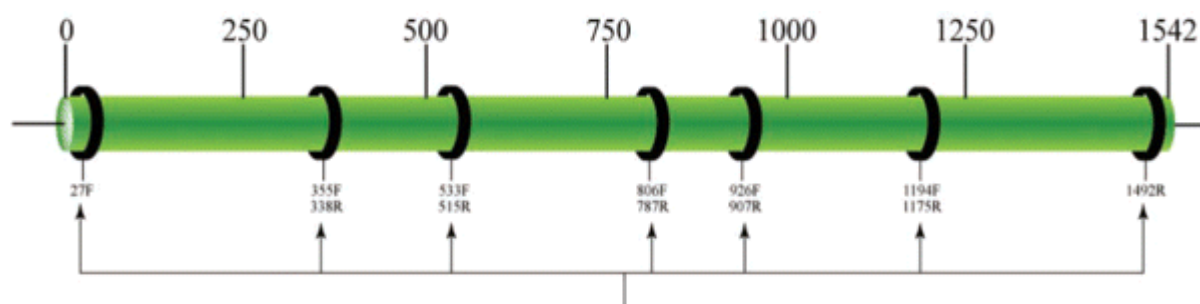
višestrukim poravnanjima sekvenci i filogenetskim analizama [20]. Analiza sekvenci 16S rRNK se koristi i za kvantifikaciju određene bakterije u ekosistemu, ali mana ovom pristupu je što većina bakterija sadrži nekoliko kopija kodona za sintezu 16S rRNK što dovodi do nemogućnosti da se odredi apsolutna zastupljenost bakterija u uzorku [21].

Nukleotidne sekvence opisane su procentima sličnosti. Bakterije čije sekvence gena 16S rRNK imaju 97 % sličnosti, pripadaju istoj vrsti. Za nivo soja procenat sličnosti mora biti ≥ 99 %. Ovaj procenat opada sa definisanjem viših taksonomskih nivoa. Tako, da bi se bakterije svrstale u istu porodicu, sličnost sekvenci treba da je ≥ 82 %, dok za klasu i tip važi procenat sličnosti od ≥ 78 % i ≥ 75 %, respektivno [20]. Međutim, treba uzeti u obzir da postoje bakterije koje, iako su pripadnici različite vrste, imaju sličnost sekvence 16S rRNK gena do 100 %, stoga se one ne mogu precizno identifikovati korišćenjem sekvence ovog filogenetskog markera. U tom slučaju mogu se koristiti drugi gen markeri kao što su *rpoB* gen (gen koji kodira – subjedinicu bakterijske RNK polimeraze), *tuf* gen (elongacioni faktor Tu) i dr [7, 20].

U genomu živih bića, između sekvenci male i velike subjediniče, nalazi se region interni transkribovani spejser (ITS), koji predstavlja repetitivnu nuklearnu jedinicu i sastoji se iz brzo evoluirajućeg ITS 1 gena i konzerviranog regiona 5,8S i dela sa umerenom evolutivnom brzinom ITS2. Zbog svojih osobina, ovaj gen je pogodan za filogenetske analize i najčešće se koristi kod gljiva [22].

Proces identifikacije bakterija započinje prikupljanjem uzoraka i obezbeđivanjem odgovarajućih uslova za njihovo skladištenje. Zatim sledi ekstrakcija genetskog materijala iz kojeg želimo da dobijemo informacije. Sledeći korak je priprema biblioteke podataka koja podrazumeva izbor hipervarijabilnog regiona koji se sekvencira, njegovo targetiranje uz pomoć univerzalnih prajmera i PCR amplifikacija [23].

Univerzalni prajmer je PCR prajmer sintetisan kao komplementaran oligonukleotid konzerviranim regionima molekula 16S rRNK gena. Ovi prajmeri su dizajnirani tako da se vezuju za krajnje delove konzerviranih regiona tako da omogućavaju umnožavanje hipervarijabilnih regiona čija se sekvenca analizira (slika 3). Dostupni su dobro proučavani setovi prajmera za umnožavanje hipervarijabilnih regiona. Međutim, činjenica je da setovi prajmera, iako zahvataju veliku većinu, ne zahvataju sve grupe bakterija u uzorku i rezultati analize mogu dovesti do pogrešnih zaključaka o sastavu proučavane zajednice [23]. Na slici 3 crni krugovi predstavljaju konzervirane delove koji su targetirani od strane prajmera [7].



Slika 3: Linearni oblik 16S rRNK [7]

Sekvenciranjem se određuje primarna struktura gena odnosno tačan redosled nukleotida u molekulu. Do danas su razvijene različite metode koje su značajno skratile vreme sekvenciranja i smanjile troškove ove procedure [21].

Prva korišćena metoda sekvenciranja bila je Sangerova dideoksi metoda, razvijena 1977. godine. Metoda je bila vremenski zahtevna i znatno skuplja od kasnije razvijenih metoda. Sa porastom potražnje sekvenciranja došlo je do razvoja novih tehnologija sekvenciranja, koje su u mogućnosti da paralelno sekvenciraju veliki broj segmenata DNK [24]. Različite metode sekvenciranja sledeće generacije (*eng. Next generation sequencing, NGS*) razvijene su od strane više kompanija. Ograničenje kod ovih metoda je nemogućnost sekvenciranja celokupne sekvence 16S rRNK, već se pristupa sekvenciranju regiona definisanim univerzalnim prajmerima [25]. Jedna od prvih platformi bila je Roche 454, poznata još i kao 454 pirosekvenciranje, koja je prvobitno napravljena za sekvenciranje celih genoma [25]. Danas je

najčešće u upotrebi Illumina MiSeq platforma čiji se rezultat sekvenciranja dobija u vidu kraćih čitanja od oko 250 baznih parova [21, 24]. Pored Illumine u upotrebi su i IonTorrent, AB SOLiD, PacBio i druge NGS platforme [25].

Bitno je istaći da i minimalne promene u proceduri - od sakupljanja uzorka do statističke obrade podataka - dovode do odstupanja rezultata analize, tako da se mogu dobiti različiti rezultati za naizgled slične studije [23]. Veliki broj dostupnih naučnih radova ukazuje na razlike u rezultatima zbog postojanja nestandardizovanih procedura za pripremu i analizu sekvenci. Zato je neophodno uzeti u obzir što više faktora koji mogu uticati na varijabilnost rezultata. Takođe, da bi procedure mogle međusobno da se uporede, neophodno je detaljno zabeležiti svaki korak analitičkog toka [21].

1.6 Bioinformatička obrada podataka

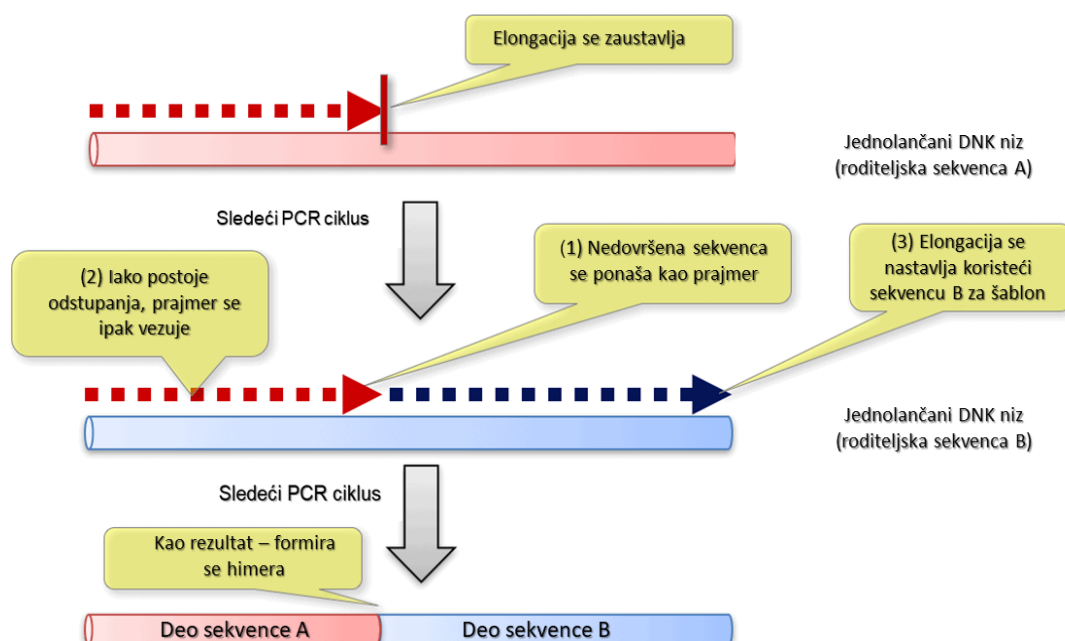
1.6.1 Eliminisanje nepravilnosti

Nakon sekvenciranja prvi korak obrade podataka podrazumeva smanjenje ili kompletno uklanjanje grešaka koje su se javile tokom PCR amplifikacije i sekvenciranja. Ove greške obuhvataju sekvence sa lošim čitanjima, sekvence sa nukleotidnim bazama koje su dvosmislene i imaju nisku ocenu kvaliteta ili sekvence koje nisu dovoljno duge. Jedna od nepravilnosti koja se javlja prilikom sekvenciranja je da sekvencer greškom stvori duge homopolimere [23, 25].

Postoji nekoliko pristupa za definisanje grešaka tokom sekvenciranja koje su zasnovane na merenju stope greške. Referentna vrednost greške dobija se merenjem greške pri sekvenciranju zajednice koja simulira sastav uzorka, čije su sekvence poznate [25].

Takođe, prilikom PCR amplifikacije, dolazi do nastanka himeričnih sekvenci. Himere su produkti PCR amplifikacije u kojima je došlo do kombinacije različitih delova sekvenci koje potiču od različitih roditeljskih nizova. Kod ovih sekvenci u toku PCR reakcije nije došlo do potpune amplifikacije, već se sekvenca prevremeno odvajaju od roditeljske i u daljem toku amplifikacije ponaša se kao prajmer (*slika 4*) [23]. Himerične sekvence je teško otkriti upravo zato što ne predstavljaju nepravilne sekvence odnosno nemaju standardne greške koje se javljaju prilikom sekvenciranja. Razvijen je veliki broj algoritama za detekciju himera kao što su *ChimeraSlayer*, *Uchime*, *Persus* i drugi [25, 26].

Način detekcije himeričnih sekvenci razvrstava se na dva tipa - prema poređenju sa referentnom bazom (*eng. reference-based*) i *de novo* metoda [27]. Prva metoda filtrira potencijalno himerične strukture i poredi ih sa posebno prilagođenom bazom podataka bez himeričnih sekvenci. *De novo* metoda je bazirana na činjenici da su roditeljske sekvence prošle kroz minimum jedan više PCR ciklus nego himerična sekvenca i činjenici da će roditeljska sekvenca biti češće prisutna među podacima [27].



Slika 4: Mehanizam formiranja himeričnih sekvenci [28]

1.6.2 Formiranje klastera (OTU)

Nakon provere ulaznih podataka i filtriranja loših sekvenci vrši se identifikacija. Postupak identifikovanja bakterija u upitnom uzorku najčešće se vrši na tri načina (*slika 5*):

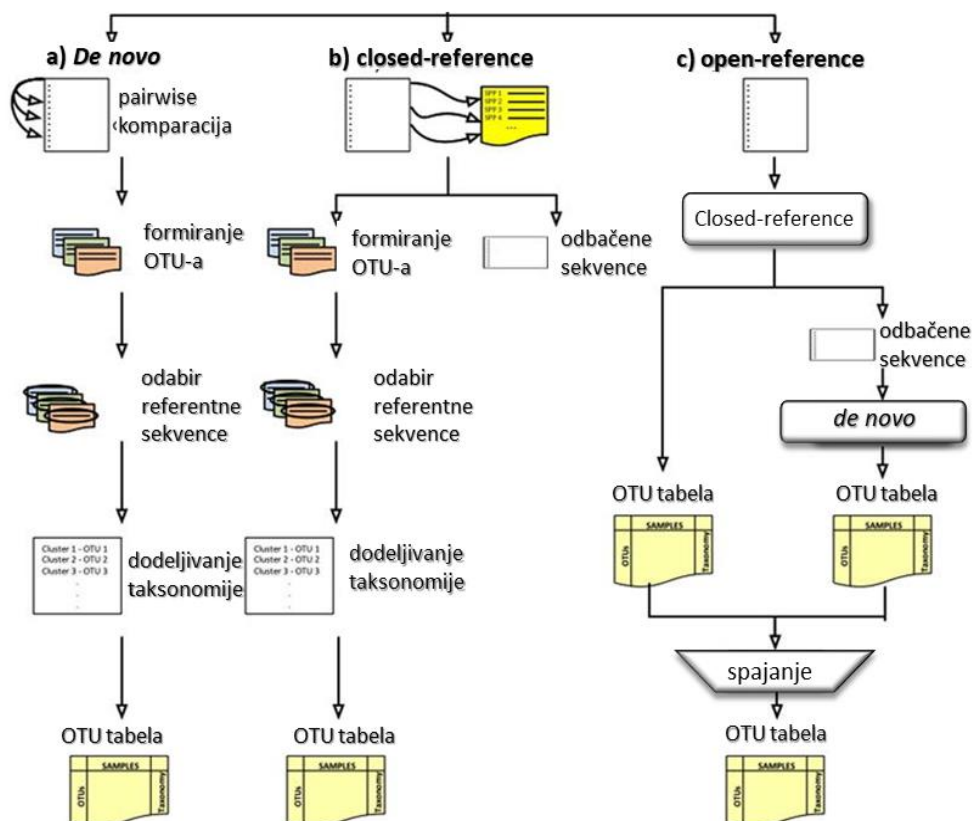
- **CLOSED-REFERENCE METODA** [29]: Dobijene sekvence porede sa sekvencama u javnim bazama podataka i na osnovu sličnosti sa bazom formiraju taksonomske grupacije (*eng. Operational taxonomic units, OTUs*) – takozvani filotipski metod. (*slika 5b*) [23]. Metoda je brza i nije kompjuterski zahtevna. Algoritam koji je trenutno najviše u upotrebi je *Naivni Bajes Klasifikator*, koji od dobijene sekvence formira kraće – do 8 nukleotida i poredi ih sa referentnom sekvencom [30].

Mana ovog pristupa je što je grupisanje OTU-a na ovaj način direktno zavisno od kvaliteta referentne baze. Dakle, ako je određeni broj sekvenci nov ne može se klasifikovati, zbog nedostatka informacija u bazi podataka. Za oblasti istraživanja koje su dovoljno dobro i dugo ispitivane, kao što je ljudska mikrobiota, baze podataka su redovno ažurirane i pogodne za ovakvu analizu. Međutim, ukoliko se radi o nedovoljno ispitanim staništima, za koje su baze podataka nepotpune, može se dogoditi da eksperimentalna sekvenca ostane neodređena [21, 23].

- **DE NOVO METODA**: Metoda se zasniva primarno na formiranju klastera (*de novo* OTU-a). Klasteri u ovoj metodi predstavljaju formiranu grupu sekvenci visoke sličnosti (*slika 5a*) [29, 31]. Prag koji se uzima za određivanje na nivou vrste je 97 %, što znači da sekvence moraju imati sličnost ≥ 97 % da bi se svrstale u istu vrstu, odnosno ≥ 99 % kako bi se svrstale u isti soj. [7] Reprezentativna sekvenca se zatim, iz svakog formiranog *de novo* OTU-a poredi sa referentnom bazom. Dakle, sekvence su grupisane i formirani klasteri ne zavise direktno od kvaliteta referentne baze. Međutim, unutar jedne jedinice mogu se naći različite taksonomske grupe [23]. Razvijene su mnoge metode za konstruisanje *de novo* OTU-a, ali u svim metodama ovi klasteri poseduju karakteristike definisane u granicama trenutno posmatranog seta podataka. Definisanje *de novo* OTU-a direktno zavisi od relativne zastupljenosti vrsta u posmatranom uzorku i samim tim definisane klastere u različitim setovima podataka nije moguće porediti. [29, 32]

Algoritmi koji se koriste za klasifikovanje sekvenci u jedan OTU su *najbliži sused*, *prosečni sused* i *najdalji sused* [9, 29, 31]. Najbliži sused zahteva da ispitivana sekvenca zadovoljava postavljeni procenat sličnosti (npr. ≥ 97 %) sa makar jednom sekvencom iz OTU-a sa kojim se poredi. Prosečni sused zahteva da procenat sličnosti ispitivane sekvence bude u okviru proseka sličnosti, dobijenog poređenjem sa svakom sekvencom iz definisanog OTU-a. Najdalji sused zahteva da ispitivana sekvenca zadovoljava postavljeni procenat sličnosti sa svakom sekvencom u definisanom OTU-u [9, 29].

- **OPEN-REFERENCE METODA** [29]: Ova metoda predstavlja kombinaciju prethodne dve (*slika 5c*). Dakle, deo upitnih sekvenci koje nisu pronađene u referentnoj bazi, obrazuje klastere na osnovu međusobne sličnosti (97 % za nivou vrste) i samo na ove sekvence primenjuje se *de novo* metoda.



Slika 5: Šematski prikaz formiranja operativnih taksonomskih jedinica [29]

Pored ova tri načina identifikacije, koja su najzastupljenija u praksi, danas je u upotrebi i klasifikacija na osnovu varijanti amplikovanih sekvenci (*eng. Amplicon sequence variants, ASVs*). Razvoj boljih metoda određivanja grešaka sekvenciranja i amplifikacije omogućava dobru kontrolu varijanti amplikovanih sekvenci tako da je moguća podela sekvenci na nivou razlike jednog nukleotida. Određivanje i razlikovanje ASV-a u odnosu na greške nastale sekvenciranjem ili u procesu amplifikacije, zasniva se na činjenici da će biološka sekvenca (sekvenca ekstrahovana direkno iz uzorka) biti najzastupljenija u setu podataka nakon amplifikacije i sekvenciranja. Ovaj način odbacuje potrebu za grupacijom sekvenci prema međusobnoj sličnosti i nije zavisen od posmatranog seta podataka [32].

Nakon izvršenog klasterovanja, izlazni podatak analize predstavlja formirana tabela zastupljenosti (OTU tabela) (slika 5). Ove tabele koriste se za dalje analize kojima se procenjuje diverzitet, odnosno složenost zajednice. Diverzitet predstavlja bogatstvo (broj različitih taksonomskih grupa i zavisi od zastupljenosti svake od taksonomskih grupa u zajednici). Alfa diverzitet odnosi se na raznolikost vrsta u posmatranoj zajednici, dok se beta diverzitet odnosi na raznolikost vrsta između zajednica. Oba tipa diverziteta proračunavaju se na osnovu već utvrđenih matematičkih algoritama [23].

1.7 Bioinformatički alati

Sa razvojem tehnologija sekvenciranja i porastom količine sekvenciranih podataka došlo je do razvoja softvera i alata kojim se omogućuje preciznija analiza diverziteta ekosistema, identifikacija mikroorganizama, njihove funkcije i evolucije. Postoji više vrsta bioinformatičkih alata koji su vremenom razvijeni za analizu podataka dobijenih sekvenciranjem hipervarijabilnih regiona sekvence 16S rRNA gena NGS metodama i u ovom poglavlju biće predstavljeni neki od njih.

1.7.1 QIIME

QIIME (*Quantitative Insights Into Microbial Ecology*) predstavlja alat dizajniran da ponudi niz instrukcija podeljenih prema fazama analize (*eng. pipeline*). Implementiran je kao kolekcija komandnih linija koji korisniku omogućavaju da neprerađene sekvence preradi do kvalitetnih grafika - paket pruža mogućnost vizualizacije rezultata analiza. Kao ulazni podaci koriste se sirove sekvence nukleinskih kiselina koje mogu da potiču od virusa, arheja, bakterija i gljiva [33]. QIIME je razvila laboratorija Robina Najta (*eng. Knight lab*), alat je besplatan za korišćenje (open-source) i ispraćen online instrukcijama i forumima [29, 34]. Nedavno je izašla nova verzija alata - QIIME2, koja pruža dodatne funkcije obrade sekvenci, statističke obrade i vizualizacije, a bazirana je na *plug-in* arhitekturi, odnosno pruža mogućnost korisniku da upotrebi dodatne softverske funkcije koje nisu bile deo primarne verzije [35].

1.7.2 UPARSE

UPARSE je softverski paket dizajniran za analizu sekvence gena 16S rRNK. Konstruisan je od strane Roberta Edgara (*eng. Robert C. Edgar*) [36]. UPARSE proverava kvalitet sekvenci, filtrira ih na osnovu određenih parametara, uklanja smetnje i greške koje se javljaju prilikom sekvenciranja i PCR amplifikacije i zatim vrši klasterovanje preostalih sekvenci u OTU-e. Klasterovanje se vrši paralelno sa proverom postojanja himeričnih sekvenci korišćenjem UPARSE-OTU algoritma, što dramatično povećava tačnost analize [36].

1.7.3 mothur

Mothur predstavlja projekat čija je težnja da obezbedi jedinstven open-source softver dostupan za bioinformatičke analize mikrobni zajednica. Mothur softver nastao je inkorporacijom prethodnih verzija, SONS i DOTUR, uz poboljšanja kao što su dodatni kalkulatori za računanje α - i β - diverziteta, računanje distanci uparenih sekvenci (*eng. pairwise distances*), kao i vizualizacije u vidu denograma i toplotnih karata (*eng. heat maps*) [9].

Softver je razvijen od strane dr. Patrika Šlosa (*eng. Patrick Schloss*) i Šlos laboratorije (*eng. the Schloss lab*) [9]. Ovaj softver kao i QIIME, radi preko pozivanih komandnih linija u komandnom prozoru. Isti parametri za filter-kontrolu sekvenci su korišćeni i u QIIME-u, ali se mothur češće koristi pri analizi nekultivisanih bakterija koje potiču iz više različitih sredina. Šlos laboratorija je razvila poseban algoritam za podelu sekvenci u OTU-e, OptiClust, koji je implementiran u mothur funkcije [37]. Zbog mogućnosti obrade podataka nastalih u različitim DNK sekvencerima (Sanger, 454 pirosekvenciranje, MiSeq, IonTorrent i drugih) najviše je korišćen bioinformatički alat. Kao i za QIIME, mothur poseduje svoje forume i online instrukcije kako bi se korisnici bolje upoznali sa mothur-ovom standardnom operativnom procedurom [3].

1.7.4 MEGAN CE

MEGAN CE (*Metagenom Analyzer Community Edition*) je softverski paket namenjen za analizu sekvenci mikrobioma, dodeljivanje taksonomije i određivanje funkcije gena. Softver poseduje ugrađene pakete koji omogućavaju vizualizaciju rezultata analize. MEGAN CE razvio je Centar za Bioinformatiku koji pripada Univerzitetu Tübingen. Dodeljivanje taksonomije bazirano je na taksonomiji Nacionalnog centra za biotehnoške informacije (NCBI), ali se mogu koristiti i druge baze. Uz paket je integrisan i alat za poravnanje DIAMOND i server Megan Server, kojim se omogućuje skladištenje podataka velike memorije [38]. Poslednja izvedena verzija MEGAN LR obezbeđuje kvalitetniju analizu dugih sekvenci [39].

1.7.5 DADA2

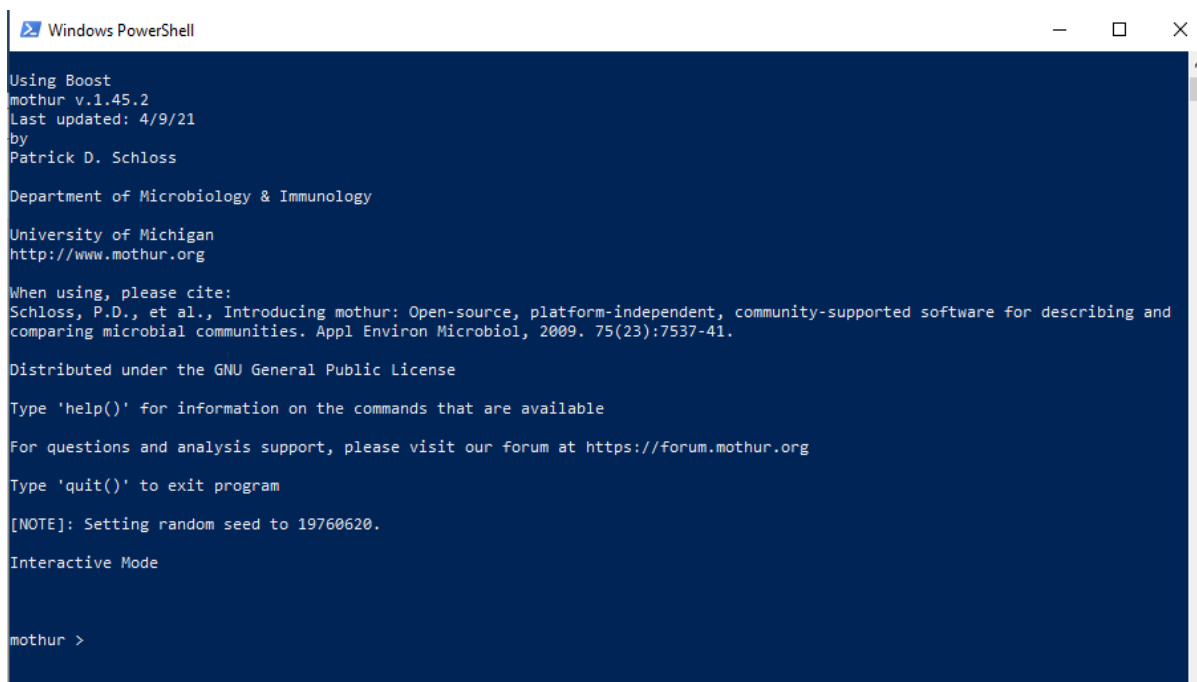
DADA2 (*The Divisive Amplicon Denoising Algorithm*) je paket, programskog jezika *R* čija je uloga da modeluje i ispravlja greške u amplikonima nastalim Illumina sekvenciranjem. Predstavlja poboljšani algoritam ranije verzije DADA i svoju analizu bazira na formiranje varijanti amplikovanih sekvenci. Paket sadrži kompletan tok analize PCR amplikona Illumina sekvenci - spajanje formiranih čitanja sekvenci, filtriranje po kvalitetu i dereplikovanje i formiranja ASV-a [40].

2. Eksperimentalni deo

2.1 Materijal i metode

Za analizu uticaja hipervarijabilnih regiona 16S rRNK kodirajućeg gena na kvalitet identifikacije i klasifikacije bakterija, odabrane su sekvence hipervarijabilnih regiona od V1 do V8 bakterija opisanih u studiji [8] koja prikazuje klasifikovane vrste koje naseljavaju ljudski gastrointestinalni trakt. Sekvence su opisne i skladištene u FASTA formatu. Kako je ljudski gastrointestinalni trakt dobro proučavana oblast, formiranje klastera i dodeljivanje taksonomije je izvršeno prema filotipskoj metodi (*closed-reference*).

Analiza je rađena u bioinformatičkom alatu mothur verzija 1.45.2. (slika 6).



```
Windows PowerShell

Using Boost
mothur v.1.45.2
Last updated: 4/9/21
by
Patrick D. Schloss

Department of Microbiology & Immunology
University of Michigan
http://www.mothur.org

When using, please cite:
Schloss, P.D., et al., Introducing mothur: Open-source, platform-independent, community-supported software for describing and
comparing microbial communities. Appl Environ Microbiol, 2009. 75(23):7537-41.

Distributed under the GNU General Public License

Type 'help()' for information on the commands that are available

For questions and analysis support, please visit our forum at https://forum.mothur.org

Type 'quit()' to exit program

[NOTE]: Setting random seed to 19760620.

Interactive Mode

mothur >
```

Slika 6: Prikaz pokrenutog alata mothur iz komandnog prozora

Vizuelizacija podataka urađena je u programskom jeziku *Python* uz korišćenje biblioteke *matplotlib.pyplot*.

Za dodeljivanje taksonomije korišćene su dve referentne baze: SILVA baza podataka i Greengenes. Bazama je pristupljeno preko mothur internet strane, na kojoj se nalaze prilagođene baze podataka za poravnanje sa analiziranim sekvencama, proveru himeričnih sekvenci i dodeljivanje taksonomije [41, 42].

SILVA referentna baza je ispraćena redovnim ažuriranjima i jedna je od trenutno korišćenih najtačnijih referentnih baza [5, 9]. Sa druge strane, verzija Greengenes baze podataka, koja se nalazi na pomenutoj internet stranici, poslednji put ažurirana je 2013. godine što je čini relativno starom bazom podataka [5].

Mothur predstavlja jedan od najviše citiranih bioinformatičkih alata. Nalazi se u vidu open-source paketa, dostupog sa online priručnikom o upotrebi i detaljno opisanih nizova komandi za analizu 16S rRNK gena i ITS rRNK gena, ali i analize proteinskih i virusnih sekvenci [9].

2.1.1 Mothur naredbe korišćene u analizi

Mothur procedura (eng. *pipeline*) za analizu sekvence 16S rRNK gena bazirana je na univerzalnom načinu bioinformatičke obrade podataka i sadrži funkcije za kontrolu kvaliteta i filtriranje sekvenci i funkcije za klasterovanje i dodelu taksonomije. Ove funkcije predstavljaju implementaciju već postojećih algoritama koji su našli svoju primenu u bioinformatičkim analizama.

Analiza sekvenci hipervarijabilnih regiona 16S rRNK gena, korišćenih u ovom radu izvršena je kroz sledeće naredbe [43]:

1. `> unique.seqs ()` – naredba koja pronalazi jedinstvene sekvence. Uzorci mikrobioma koji se analiziraju mogu da sadrže veliki broj identičnih sekvenci. Kako bi se ubrzala dalja analiza, prvo se pronalaze jedinstvena čitanja, a zatim dopisuje koliko puta su svaka od ovih međusobno različitih čitanja pronađena u originalnom skupu podataka. Ova funkcija pravi dve izlazne datoteke. Jedna sadrži imena jedinstvenih čitanja i njihovih sekvenci u FASTA formatu, dok je druga datoteka tabelarnog tipa u *names* formatu, sa dve kolone. U prvoj koloni sadrži imena jedinstvenih čitanja – reprezentativnu sekvencu, a u drugoj koloni prikazuje skup svih ostalih imena sekvenci koja su identična sa reprezentativnom.
2. `> count.seqs ()` – naredba koja kao ulazni parametar uzima prethodno napravljenu datoteku u *names* formatu i formira tabelu sa imenima jedinstvenih sekvenci i brojem njihovog ponavljanja. Ovakva datoteka skladištena je u *count_table* formatu.
3. `> align.seqs ()` – naredba koja vrši poravnanje definisane FASTA datoteke sa referentnom bazom. Naredba omogućava postavljanje dodatnih parametara kao što su:
 - Način pretrage šablon sekvence iz referentne baze. Mogućnosti su *kmer*, *blast* i *suffix*, u ovom radu korišćen je način *kmer*, koji je podrazumevani način pretrage ukoliko se ne definiše suprotno. Pokazao se kao najbrži način.
 - Metod poravnanja definiše na koji način će se ispitivana sekvenca poravnati sa najbližijom sekvencom iz referentne baze. Postoji više metoda poravnanja koje se koriste, kao što su *blastn*, *goth* i *needleman*, a u ovom radu korišćen je *needleman* metod (*Needleman-Wunsch* algoritam).
 - Definisanje stepena negativnog uticaja (eng. *penalty score*) praznina i nepodudaranja nukleotida na krajnju ocenu poravnanja. Mogućnosti su: *match*, *mismatch*, *gapopen*, i *gapextend*.

Naredba *align.seqs()* kreira tri izlazne datoteke. Datoteka formata *align* sadrži jedinstvena čitanja koja su poravnata sa svojim šablonom iz referentne baze i kompletnu sekvencu tih čitanja. *Align* format je jedan vid FASTA formata. Datoteka formata *accnos* sadrži spisak jedinstvenih sekvenci koje nisu uspele da se poravnaju sa referentnim sekvencama. Zabeležene su sve sekvence koje nisu korisne za dalju analizu. Poslednja, treća datoteka je *align.report*, koja je tabelarnog tipa i sadrži kompletne informacije o izvršenom poravnanju (ocenu poravnanja za svaku sekvencu, način pretrage, ime šablona i njegovu dužinu sekeve i dr., slika 7).

	QueryName	Query Length	TemplateName	Template Length	Search Method	Search Score	Alignment Method	Query Start	Query End	Template Start	Template End
1											
2	Unl46593	60	AP011765.UncDel53	1475	kmer	100	needleman	1	60	1237	1296
3	AgoJeu2	60	CP017580.HFSP833	1444	kmer	100	needleman	1	60	1204	1263
4	Unl33279	60	AM405748.Unc65074	1441	kmer	100	needleman	1	60	1201	1260
5	Unl47407	60	FPLO01006500.GJ6Z2	1446	kmer	100	needleman	1	60	1208	1267
6	Unl33516	59	AM406445.Unc67609	1437	kmer	100	needleman	1	59	1198	1256
7	Unl47674	60	AM405728.Unc74152	1430	kmer	100	needleman	1	60	1190	1249
8	Unl52645	59	AM406496.Unc54200	1451	kmer	100	needleman	1	59	1214	1272
9	GdtPamel	60	LT900217.I9BUrol3	1430	kmer	100	needleman	1	60	1190	1249
10	RumFla24	59	AB824503.Unc40191	1433	kmer	100	needleman	1	59	1196	1254

Slika 7: Prikaz izveštaja *align.report* za prvih deset sekvenci, nakon izvršenog poravnanja regiona V8

4. `> summary.seqs ()` – naredba sumira kvalitet sekvenci. Može se koristiti sa nepravilnim ili poravnanim sekvencama. U toku ove analize, korišćena je za sumiran prikaz poravnanja izvršenog pozivanjem naredbe `align.seqs()`. Inputi su prethodno formirana datoteka `align` i poslednje formirana datoteka `count_table`. Uvođenjem datoteke `count_table`, kao jedan od parametara obezbeđuje se obuhvat kompletnog spiska sekvenci. Ova naredba omogućava da vidimo kvalitet poravnanja direktno iz komandnog prozora (slika 8). Izlaz funkcije u komandnom prozoru sadrži početnu i krajnju poziciju u šablonu baze, informaciju o postojanju dvosmislenih nukleotida i broj očitanih homopolimera. Formirana izlazna datoteka je u `summary` formatu i sadrži navedene informacije za svaku posmatranu sekvencu.

```
mothur >
summary.seqs(fasta=HITestV1.unique.align, count=HITestV1.count_table)
```

Using 4 processors.

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	0	0	0	0	1	1
2.5%-tile:	1143	1787	22	0	2	64
25%-tile:	1143	1793	38	0	3	634
Median:	1143	1793	42	0	3	1267
75%-tile:	1143	1793	52	0	4	1900
97.5%-tile:	1184	1793	62	0	5	2470
Maximum:	43105	43116	167	6	7	2533
Mean:	1515	2155	43	0	3	
# of unique seqs:			1849			
total # of seqs:			2533			

It took 5 secs to summarize 2533 sequences.

Output File Names:
HITestV1.unique.summary

Slika 8. Prikaz `summary.seqs()` funkcije na V1 hipervarijabilnom regionu korišćenom u analizi

5. `> screen.seqs ()` - naredba omogućuje da se zadrže sekvence koje ispunjavaju određene kriterijume. Kriterijume definišemo kroz parametre funkcije i oni mogu biti dužina posmatranih sekvenci, broj dvosmislenih nukleotida, dužina homopolimera kao i početna i krajnja pozicija poravnanja. Pored ovih parametara neophodno je definisati korišćenu FASTA datoteku, `count_table` datoteku i `summary` datoteku dobijenu iz prethodne `summary.seqs()` naredbe.
6. `> filter.seqs ()` – naredba uklanja kolone iz formiranog poravnanja na osnovu dodeljenih kriterijuma. Najčešće se uklanjaju kolone koje su kompletno ispunjene oznakama poput „-“ ili „-“, jer ne sadrže informacije po kojima se mogu odrediti sličnosti među sekvencama. Ulazni podaci su FASTA datoteka (može biti u vidu `align` formata) i definisani kriterijumi. Izlazna datoteka je skladištena je kao `filter.fasta` format.
7. `> pre.cluster ()` – naredba predstavlja implementaciju algoritma pseudo-jedne veze (eng. *pseudo-single linkage*) [44]. Glavni cilj je da se spoje sekvence koje se razlikuju za definisan broj nukleotida. Algoritam prvo sortira sekvence prema zastupljenosti, zatim uzima najzastupljenije sekvence i formira skup svih sekvenci koje se od najzastupljenije razlikuju za definisani broj nukleotida. Sekvence koje se razlikuju za jedan nukleotid na 100 nukleotida su najverovatnije greške nastale tokom sekvenciranja i ne predstavljaju pravu biološku varijaciju. Definisano odstupanje za analizu izvršenu u ovom redu je 2, što predstavlja podrazumevanu vrednost za dati parametar. Izlazne datoteke su `precluster.fasta` i `precluster.count_table`.
8. `> chimera.vsearch ()` – predstavlja jednu od naredbi u `mothur` alatu, uz pomoć koje se vrši provera postojanja himeričnih sekvenci. Himerične sekvence predstavljaju veštačke tvorevine nastale tokom PCR amplifikacije i neophodno ih je ukloniti pre dodeljivanja taksonomije kako se ne bi zamenile za novu vrstu [25]. *Vsearch* predstavlja bioinformatički alat za obradu podataka sekvenci populacije [45]. Algoritam za pretragu himeričnih sekvenci u *Vsearch* alatu implementiran je u `chimera.vsearch` naredbi `mothur-a`. Za analizu u ovom radu korišćen je de novo metod pretrage himeričnih sekvenci, pa su ulazni parametri `fasta` i `count_table` datoteka iz prethodnog koraka. Izlazni podaci sadrže datoteku u `chimeras` formatu [46].

9. `> classify.seqs ()` – je naredba koja dodeljuje taksonomiju analiziranim sekvencama. Za dodeljivanje taksonomije korišćen je Naivni Bajes klasifikator. Ulazni podaci su datoteka FASTA formata i datoteka *taxonomy* formata, koja se nalazi u setu podataka dobijenih uz referentne baze [41, 42]. Jedan od parametara je i *cutoff* parametar kojim se definiše minimalna butstrep (eng. *bootstrep*) vrednost. Ideja je da se jedna sekvenca podeli na kraće podsekvence (eng. *subsets*) i klasifikuje na osnovu njih. Postupak se ponovi 100 puta po sekvenci – ovaj postupak predstavlja butstrep metodu. Povratna informacija je vrednost izražena u procentima i za jednu sekvencu predstavlja najčešću klasifikaciju prema podsekvencama (butstrep vrednost). Dakle ukoliko definišemo *cutoff* = 80, svaki nivo klasifikovan ispod te vrednosti biće označen kao *unclassified*. Izlazni podaci predstavljaju *taxonomy* datoteku – koja sadrži pripisanu taksonomiju za svaku sekvencu, *tax.summary* datoteku – koja sadrži broj sekvenci koje su pronađene na svakom nivou i *accnos* datoteku – koja sadrži sve sekvence kojima je dodeljen nepoznat takson.
10. `> remove.lineage ()` – naredba kojom se uklanjaju sekvence koje ne pripadaju definisanim kriterijumima. U zavisnosti od analize istraživač može zahtevati uklanjanje konkretne takse prema oblasti koju istražuje. U ovoj analizi uklonjene su sekvence koje mogu poticati od eukarija, hloroplasta i mitohondrija, ali i sekvence koje su klasifikovane kao nepoznati takson iz prethodne naredbe. Ulazni parametri, pored definisanog taksona, su *fasta* i *count_table* datoteka i *taxonomy* datoteka iz prethodnog koraka.
11. `> phylotype ()` – naredba je korišćena za grupisanje OTU-a na osnovu dodeljene taksonomije. Ulazni parametar je *taxonomy* datoteka, koja je dobijena u prethodnom koraku, što znači da ne sadrži nepoznate takson, kao ni sekvence mitohondrija, hloroplasta eukarija i arheja. Pored taksonomije, može se definisati i nivo taksonomije koji želimo da uključimo – od 1 do 6/7, gde 1 predstavlja nivo roda, kao najniže takson u SILVA referentnoj bazi, odnosno nivo vrste kada je reč o Greengenes referentnoj bazi. Ova funkcija daje izlazne datoteke tipa *list*, *rabund* i *sabund*, koje sadrže neophodne podatke za vizuelizaciju rezultata analize.
12. `> classify.otu ()` - naredba koja povezuje definisane OTU-e u *list* datoteci sa dodeljenom taksonomijom u *taxonomy* datoteci, što predstavlja i ulazne parametre funkcije. Izlazne datoteke sadrže podatke o taksonomiji svakog klasterovanog OTU-a, sa brojem sekvenci koju sadrži svaki OTU pojedinačno. Format izlaznih datoteka je *cons.taxonomy*.

2.1.2 Vizualizacija podataka

Sa navedenim dvanaestim korakom završena je analiza hipervarijabilnih regiona sekvence 16S rRNK u mothur alatu. Dalji tok rada predstavlja vizuelizacija dobijenih rezultata u *python* programu, korišćenjem biblioteke *matplotlib.pyplot*. Cilj vizuelizacije je da se prikaže odnos broja sekvenci za koje je dodeljen poznat takson u odnosu na ukupan broj posmatranih sekvenci opisanih u navedenoj studiji [8]. Odnos će biti prikazan u vidu trakastog grafikona. Pored vizuelizacije uspešnosti dodeljivanja taksona, analiza obuhvata prikaz udela sekvenci koje pripadaju nekultivisanim bakterijama [8] u neklasifikovanim, neporavnatim i nepoznatim sekvencama, za svaki hipervarijabilni region.

2.2 Rezultati i diskusija

U ovom radu analizirano je ukupno 2533 sekvenci 16S rRNK, kultivisanih (1216 sekvenci) i nekultivisanih (1317 sekvenci) bakterija koje nastanjuju gastrointestinalni trakt čoveka. Izvršena je taksonomska identifikacija bakterija korišćenjem *mothur* alata i sekvenci pojedinačnih hipervarijabilnih regiona. Iako sekvenca gena 16S rRNK ima devet hipervarijabilnih regiona, u ovom radu analiza je izvršena na sekvencama osam dostupnih regiona.

Prethodno navedenih dvanaest naredbi analize primenjeno je za obradu sekvenci gena koji kodira sintezu 16S rRNK podeljenog u osam hipervarijabilnih regiona (V1-V8). Prve dve naredbe, *unique.seqs()* i *count.seqs()* optimizuju sekvence za kompjutersku analizu i ne zavise od izbora referentne baze. Treća naredba uzima kao jedan od parametara referentnu bazu. Zbog činjenice da je analiza kvalitetna onoliko koliko je kvalitetna korišćena referentna baza, ova naredba predstavlja prekretnicu u toku analize [5].

2.2.1 Taksonomska identifikacija korišćenjem SILVA referentne baze

Korišćenjem naredbe *align.seqs()* vršeno je poravnanje sekvenci koje su ispitivane sa sekvencama iz referentne baze. SILVA referentna baza sadrži 50 000 mogućih pozicija za svaku sekvencu. Imajući u vidu da je dužina 16S rRNA gena oko 1550 nukleotida može se zaključiti da u poravnanjima postoji veliki broj praznih mesta (*eng. gaps*) i poznate su pozicije između kojih se nalaze hipervarijabilni regioni. Poravnanja hipervarijabilnih regiona koja su korišćena kao parametri *screen.seqs()* naredbe, prikazani su u tabeli 2.

Tabela 2: Početna i krajnja pozicija poravnatih sekvenci za svaki hipervarijabilni region

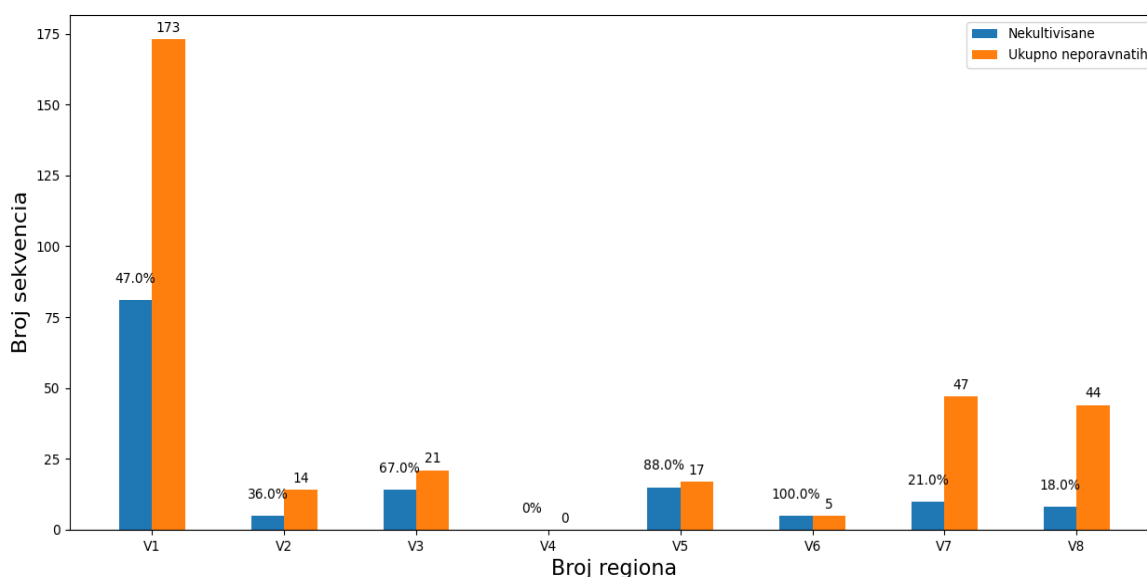
Hipervarijabilni region	Početna pozicija	Krajnja pozicija	Broj uklonjenih sekvenci naredbom <i>screen.seqs()</i> , (<i>n%</i> ukupnog broja sekvenci - 2533)
V1	1143	1793	173, (6,82)
V2	3153	5276	14, (0,55)
V3	9878	10275	21, (0,83)
V4	21930	22538	0, (0,00)
V5	25500	26984	17, (0,67)
V6	31187	33285	5, (0,19)
V7	35461	37688	47, (1,85)
V8	40334	40961	44, (1,74)

U tabeli 2 prikazan je broj odbačenih sekvenci za svaki varijabilni region. Sekvence koje su odbačene su sekvence koje nisu izvršile poravnanje prema pozicijama iz *tabele 2*. Takođe, odbačene su sekvence koje su imale mali broj baza ili su sadržale dvosmislene baze.

Nakon izvršenog tabelarnog prikaza karakteristika poravnanja svakog hipervarijabilnog regiona, iako gubici nisu veliki, iz prikazanog se može zaključiti da je V4 region imao najbolje poravnanje, s obzirom da nije izgubljena nijedna sekvenca, dok V1 region ima najviše gubitaka.

Određen je udeo sekvenci nekultivisanih bakterija u ukupnom broju sekvenci koje su odbačene zbog lošeg poravnanja (*slika 9*). Udeo je računat u *Python* programskom jeziku izvlačenjem imena sekvenci iz odgovarajućih datoteka. Na osnovu dobijenih rezultata može se zaključiti da su odbačene sekvence

i kultivisanih i nekultivisanih bakterija i da udeo nekultivisanih sekvenci među odbačenim sekvencama značajno varira za varijabilne regione. Svih 5 odbačenih sekvenci regiona V6 su poticale od nekultivisanih bakterija, dok je za varijabilni region V8 samo 18 % odbačenih sekvenci poticalo iz nekultivisanih bakterija.



Slika 9: Grafički prikaz udela sekvenci koje su nakon poravnanja odbačene, a pripadaju nekultivisanim bakterijama [8]

Iako u analiziranom setu podataka nisu očekivane himerične sekvence primenjena je naredba za uklanjanje himeričnih sekvenci pozivanjem funkcije *chimera.vsearch()*, *de novo* metodom. U skladu sa očekivanjima ova naredba prilikom analize svakog od hipervarijabilnih regiona nije pronašla himerične sekvence. Provera kvaliteta pretrage *chimera.vsearch()*, izvršena je ponovnom pretragom, korišćenjem naredbe *chimera.uchime()*. Poređenjem sa referentnom bazom **silva.gold.align** [42] naredba *chimera.uchime()* takođe nije pronašla himerične sekvence.

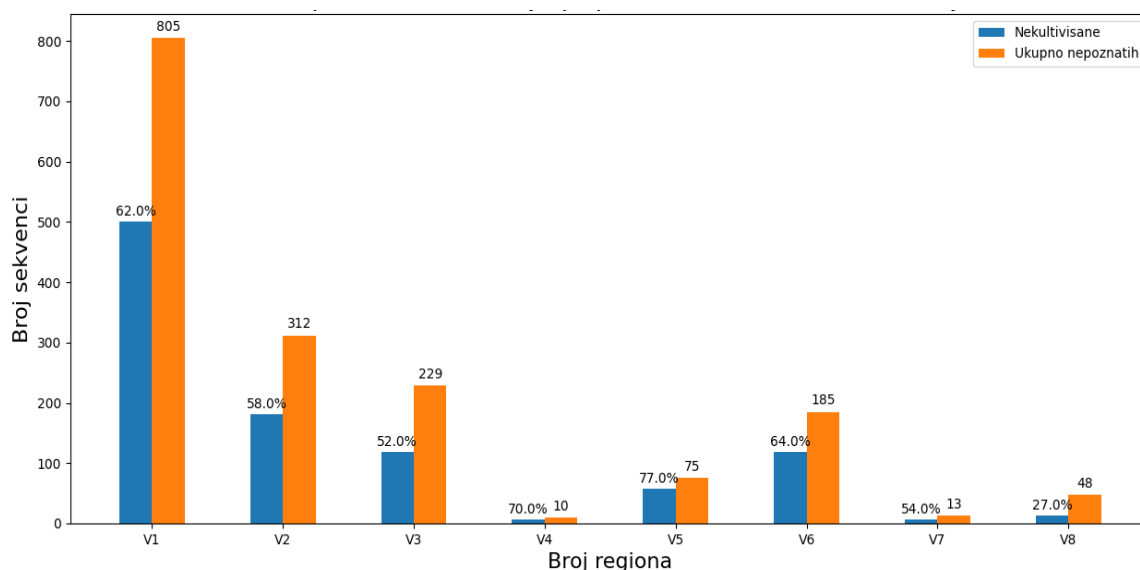
U sledećem koraku ispitivanim sekvencama je dodeljivana taksonomska identifikacija i uklonjene su sekvence koje, na osnovu primenjene analize, ne pripadaju bakterijama ili su označene kao nepoznate pri dodeli taksnomije (tabela 3).

Tabela 3: Udeo nepoznatog taksona u ukupnom broju uklonjenih sekvenci nakon dodeljivanja taksonomije

Hipervarijabilni region	Ukupan broj uklonjenih sekvenci	Broj sekvenci nepoznate takse , (n%)
V1	879	805 , (91,6)
V2	335	312, (93,1)
V3	243	229, (94,2)
V4	29	10, (34,5)
V5	96	75, (78,1)
V6	204	185, (90,1)
V7	28	13, (46,4)
V8	66	48, (72,7)

Iz *tabele 3* može se zaključiti da su sekvence najbolje klasifikovane u regionu V7, jer je uklonjeno svega 28 sekvenci. Međutim, region V4 od ukupnog broja uklonjenih sekvenci, brojčano i procentualno, sadrži najmanje sekvenci nepoznatog taksona. Ovaj podatak može da ukaže na činjenicu da je region V4 jedan od najčešće i najviše ispitivanih regiona. Udeo sekvenci nekultivisanih bakterija izvučenih iz

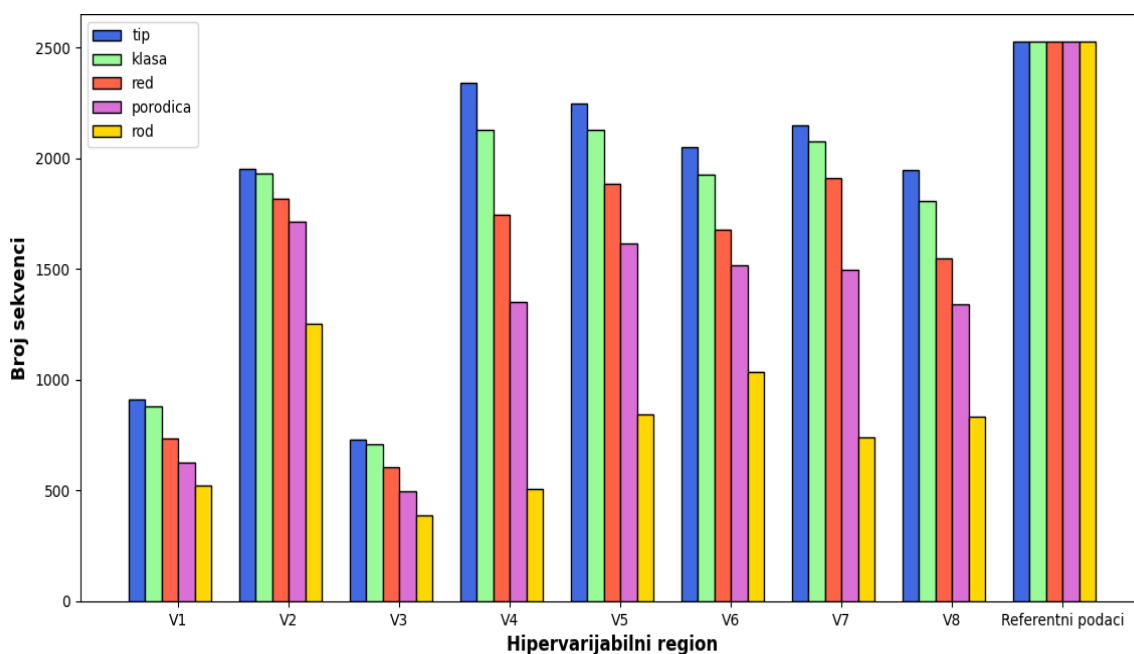
referentnih podataka [8], koje pripadaju dodeljenim nepoznatim taksonima prikazan je u vidu grafika na slici 10.



Slika 10: Grafički prikaz udela sekvenci koje pripadaju nekultivisanim bakterijama [8] i označene su kao nepoznati takson nakon izvršene klasifikacije

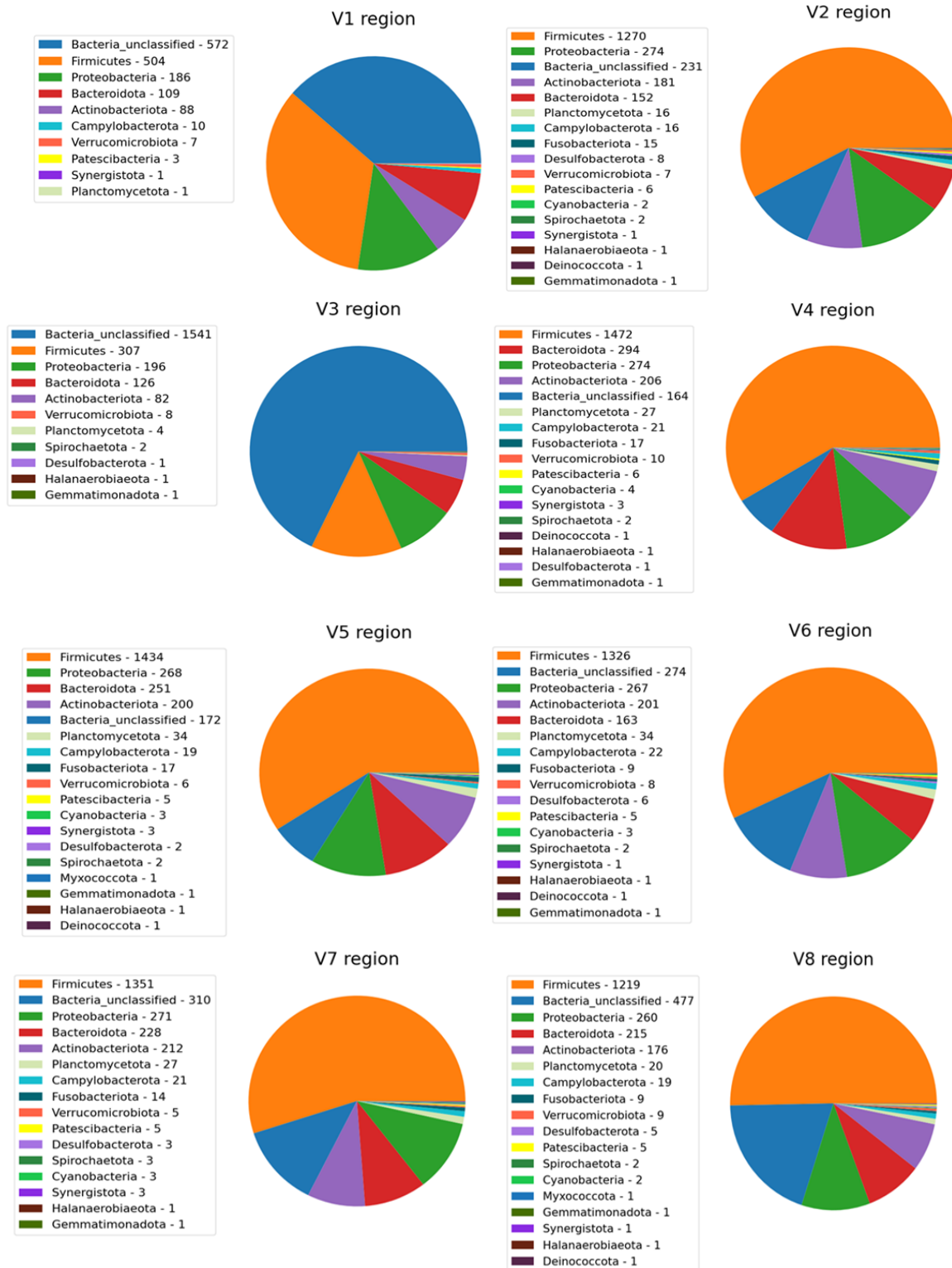
Sa grafičkog prikaza (slika 10), može se zaključiti da više od 50 % sekvenci nepoznatog taksona, u svakom varijabilnom regionu pripada nekultivisanim bakterijama. Regioni V4 i V7 pokazali su se kao region koji imaju najniži broj izgubljenih sekvenci nakon izvršene klasifikacije. Razlog za velike gubitke u regionima V1, V2, V3, V6 je u većem stepenu varijabilnosti ovih regiona, što im pripisuje i višu diskriminativnu moć.

Pozivanjem funkcija *phylotype()* i *classify.otu()* izvršena je raspodela u OTU jedinice. Vizualizacijom podataka dobijenih iz izlaznih datoteka koje su rezultat navedenih funkcija, dobijen je trakasti grafikon (slika 11).



Slika 11: Grafički prikaz oporavljenih sekvenci bioinformatičkom metodom opisanom u ovom radu po taksonomskim nivoima (SILVA referentna baza)

Iz grafičkog prikaza može se videti da regioni V1 i V3 imaju najmanji broj oporavljenih sekvenci, što je potvrđeno na *slikama 12 i 13*, koje pokazuju da ova dva regiona imaju najviše sekvenci kojima nije dodeljen taksonomski nivo tipa. V1 i V3 region imaju najveći broj neklasifikovanih sekvenci već u nivou tipa. V4 hipervarijabilni region ima najviše oporavljenih sekvenci. Međutim pri klasifikaciji do nivoa roda, V4 ima velikih gubitaka u vidu neklasifikovanih sekvenci i sa grafika se može videti da nivo roda regiona V4 ne odstupa mnogo od V1 i V3 regiona. Region V2, pokazao se kao najbolji region u ovoj analizi. U nivou tipa, V2 region ima gubitaka, ali ne prevelikih. Dubljom klasifikacijom do nivoa roda ne dolazi do velikih gubitaka sekvenci.



Slike 12 i 13: Grafički prikaz udela klasifikovanih taksona na nivou tipa, za svih osam hipervarijabilnih regiona

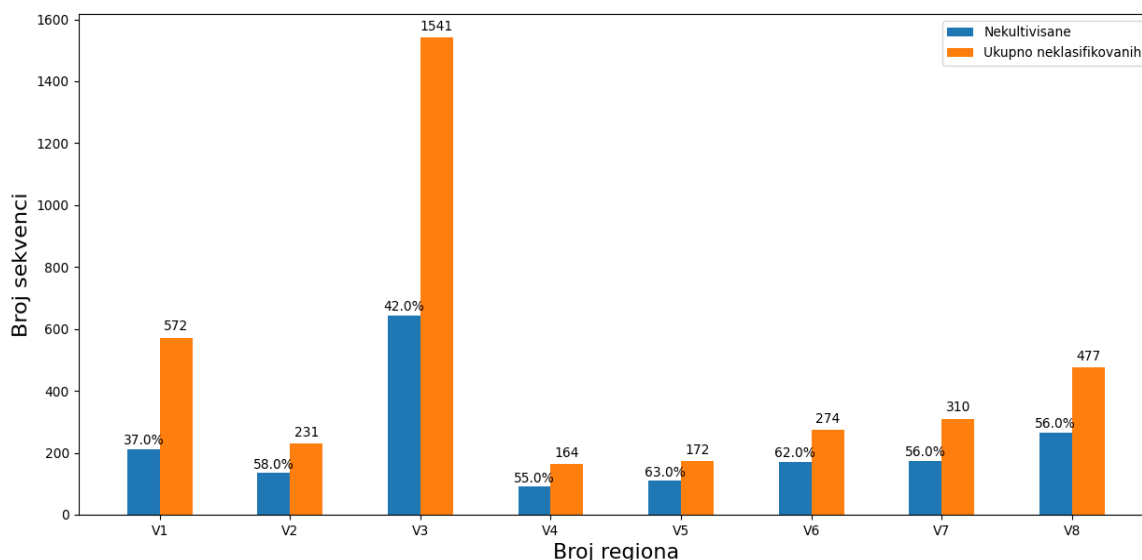
Kružni grafikoni sa *slika 12 i 13*, predstavljaju dodeljene taksoni za nivo tipa. Može se zaključiti da je tip *Firmicutes* najzastupljeniji, kao i da u proseku broji oko 1300 sekvenci. Zatim slede tipovi *Proteobacteria*, *Bacteroidota* i *Actinobacteriota* čiji se broj sekvenci kreće u proseku između 200 i 300 sekvenci (*tabela 4*). Iako su sekvence varijabilnih regiona ekstrahovane iz seta od 2533 sekvenci, rezultati taksonomske analize na osnovu sekvenci varijabilnih regiona su drastično različiti (*slike 12 i 13*), što je sumirano u *tabeli 4*. Ovo je delimično posledica odbacivanja određenog broja sekvenci u prethodnim koracima. Uticaj ima i činjenica da je stepen varijacije u okviru različitih hipervarijabilnih regiona različit zbog čega je i mogućnost precizne identifikacije varijabilna. Treba napomenuti da su pri analizi nepoznatih sekvenci bili primenjeni isti koraci, što ukazuje da izbor hipervarijabilnog regiona izuzetno utiče na rezultate dobijene NGS metodama.

Tabela 4. Sumirani prikaz dodeljene takse na nivou tipa, po hipervarijabilnim regionima [8]

Nivo tipa	REF.	V1	V2	V3	V4	V5	V6	V7	V8
Actinobacteria	216	88	181	82	206	200	201	212	176
Bacteroidetes	310	109	152	126	294	251	163	228	215
Chloroflexi	1								
Crenarchaeota	8								
Cyanobacteria	5		2		4	3	3		2
Deinococcus-Thermus	1		1		1	1	1		1
Euryarchaeota	9								
Firmicutes	1559	504	1270	307	1472	1434	1326	1351	1219
Firmicutes (Halobacteriales)	1								
Fusobacteria	17		15		17	17	9	14	9
Gemmatimonadetes	1		1	1	1	1	1	1	1
Lentisphaerae	3								
Planctomycetes	41	1	16	4	27	34	34	27	20
Proteobacteria	317	186	274	196	274	268	267	271	260
Spirochaetes	3		2	2	2		2	3	2
Synergistetes	4	1	2		3	3	1	3	1
Tenericutes	16								
TM7	7								
Verrucomicrobia	7	7	7	8	10	6	8	5	9
Campylobacterota		10	16		21	19	22	21	19
Patescibacteria		3	6		6	5	5	5	5
Desulfobacteriota			8	1	1	2	6	3	5
Halanaerobiaeota			1	1	1	1	1	1	1
Myxococcota						1			1

Tabela 4 prikazuje sumirano dodeljene taksoni za nivo tipa kao i broj sekvenci koje taksoni sadrže. Tipovi koji su navedeni crnom bojom nisu su nalazili među referentnim podacima. Tipovi navedeni u plavo osenčenim redovima, iako potiču iz referentnih sekvenci koje pripadaju ovim taksonomskim grupama, ovom analizom nisu pronađeni.

Analizom sekvenci koje su označene kao neklasifikovane, za svaki hipervarijabilni region, dobijen je trakasti grafikon (*slika 14*), koji prikazuje broj ukupno neklasifikovanih sekvenci po regionima i udeo koji čine bakterije koje nisu kultivisane [8].



Slika 14: Udeo nekultivisanih bakterija [8] u ukupnom broju neklasifikovanih bakterija

Sa grafika (slika 14), zaključuje se da je najveći procenat bakterija koje nisu klasifikovane ali ne pripadaju grupi nekultivisanih bakterija utvrđen za regione V1 i V3, što je bilo i očekivano s obzirom da imaju najniži broj sekvenci koje su potvrđene analizom.

Razlog za ovako tzv. loše ponašanje regiona V1 i V3, pored kvaliteta referentne baze i rezultata poravnanja, može se naći i u kvalitetu i dužini sirovih sekvenci. Početne, neporavnate sekvence ovih regiona provučene su kroz funkciju `summary.seqs()`, a rezultati funkcije su prikazani na slikama 15 i 16.

```
summary.seqs(count=HITTestV1.count_table)
Using HITTestV1.unique.fasta as input file for the fasta parameter.

Using 4 processors.

      Start  End  NBases  Ambigs  Polymer  NumSeqs
Minimum:    1    0      0      0        1         1
2.5%-tile:   1   22     22      0        2        64
25%-tile:    1   38     38      0        3       634
Median:      1   42     42      0        3      1267
75%-tile:    1   52     52      0        4     1900
97.5%-tile:   1   64     64      0        5     2470
Maximum:     1  176    176      6        7     2533
Mean:    1    43     43      0        3
# of unique seqs:      1849
total # of seqs:      2533

It took 0 secs to summarize 2533 sequences.
```

Slika 15: Prikaz snimljenog ekrana nakon pokretanja naredbe `summary.seqs()` za hipervarijabilni region V1

```

mothur >
summary.seqs(fasta=HITTestV3.fasta)

Using 4 processors.

      Start   End   NBases  Ambigs  Polymer  NumSeqs
Minimum:    1    18     18      0       2        1
2.5%-tile:    1    26     26      0       2       64
25%-tile:    1    26     26      0       3      634
Median:      1    29     29      0       3     1267
75%-tile:    1    49     49      0       4    1900
97.5%-tile:  1    52     52      0       5    2470
Maximum:    1    53     53      5       7    2533
Mean:      1    36     36      0       3
# of Seqs:   2533

It took 0 secs to summarize 2533 sequences.

```

Slika 16: Prikaz snimljenog ekrana nakon pokretanja naredbe `summary.seqs()` za hipervarijabilni region V3

Funkcija `summary.seqs()` pokazala je da su sekvence posmatranih regiona relativno kratke sekvence. Samo 2,5 % sekvenci regiona V1 prelazi dužinu od 60 nukleotidnih baza. Dok je za region V3 kod 2,5 % posmatranih sekvenci dužina 53 nukleotidnih baza, a čak polovina sekvenci ne prelazi dužinu od 30 nukleotida. Ovakvi ulazni podaci sigurno jesu uzročnik loših rezultata analize prikazanih na prethodnim grafikonima.

2.2.2 Taksonomska identifikacija korišćenjem Greengenes referentne baze

Za razliku od SILVA referentne baze, Greengenes referentna baza čuva po sekvenci približno 8000 pozicija. Baza je poslednji put ažurirana 2013 što znači da ne sadrži novootkrivene mikroorganizme. [5]

Regioni V1, V2, V3, V6, V7 i V8 analiziranog seta sekvenci su pokazali izuzetno loše poravnanje sa sekvencama u referentnoj Greengenes bazi. Na *slici 17* prikazan je V1 hipervarijabilni region, nakon poravnanja koji bi trebalo da bude smešten između tačno određenih pozicija.

```

mothur >
summary.seqs(fasta=HITTestV1.unique.align)

Using 4 processors.
[WARNING]: This command can take a namefile and you did not provide one.

      Start   End   NBases  Ambigs  Polymer  NumSeqs
Minimum:    0     0       0       0       1        1
2.5%-tile:  109   117       3       0       2       47
25%-tile:  1614  1713      11       0       2      463
Median:    3979  4064      28       0       3     925
75%-tile:  6442  6456      40       0       4    1387
97.5%-tile: 6811  6818      54       0       5    1803
Maximum:   6845  6852      67       4       7    1849
Mean:    3715  3869      26       0       2
# of Seqs: 1849

It took 1 secs to summarize 1849 sequences.

Output File Names:
HITTestV1.unique.summary

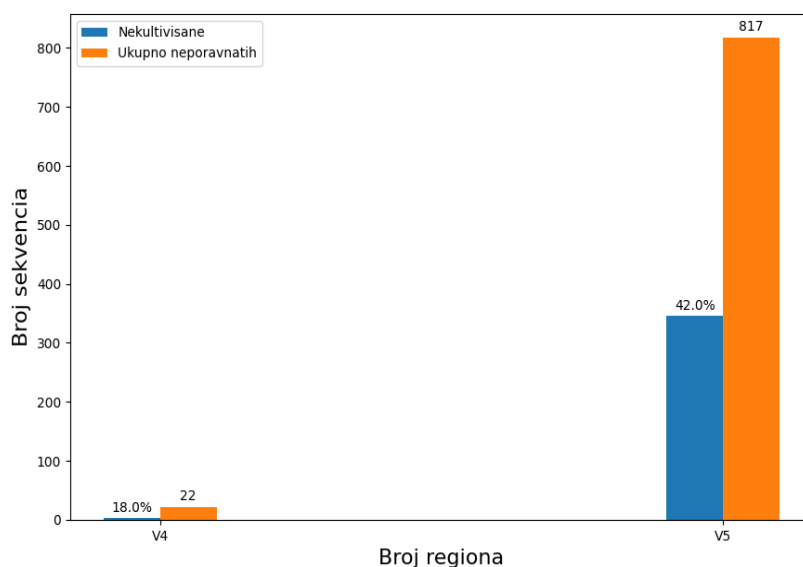
```

Slika 17: Prikazuje deo koda nakon pokrenute naredbe `summary.seqs()` za poravnate sekvence regiona V1

Slika 17 pokazuje da se prvih 2,5 % očitanih sekvenci poravnalo u predelu između 109. i 117. pozicije. 25 % sekvenci poravnalo se između 1614. i 1713. pozicije, što odgovara poziciji V1 hipervarijabilnog regiona. Međutim, ostale sekvence se ravnaју između 3979. - 4064. pozicije i između 6442.- 6456. pozicije. Ukoliko se izvrši uklanjanje sekvenci prema kriterijumima početne i krajnje pozicije, koje su definisane prema pozicijama hipervarijabilnih regiona, za dalju analizu ostaće mali broj sekvenci. Zbog toga su regioni V1, V2, V3, V6, V7 i V8 odbačeni za dalju analizu.

Region V4 nakon poravnavanja prikazuje rezultate pogodne za dalju analizu.. Gotovo kompletan broj analiziranih sekvenci poravnao se između pozicija 3824. i 4000., tako da je filtriranjem sekvenci prema ovom kriterijumu uklonjeno samo 22 sekvence. Od uklonjenih 22 sekvence 4 pripada nekultivisanim bakterijama [8].

Region V5 poravnat je između 4102. i 4469. pozicije. Međutim, nakon primene filtera sa ovim vrednostima uklonjeno je 817 sekvenci. Od 817 uklonjenih sekvenci, 346 pripada nekultivisanim bakterijama.[8] Ovi rezultati prikazani su grafički na slici 18.



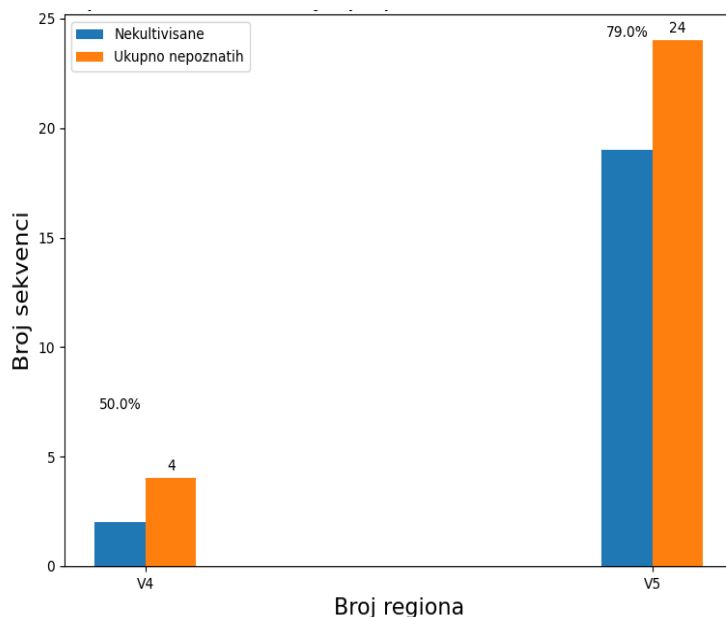
Slika 18: Udeo nekultivisanih sekvenci [8] u sekvencama koje nisu poravnate

Dalji tok analize izvršen je na isti način kao i pri analizi korišćenjem SILVA referentne baze, što je opisano u prethodnom poglavlju. Izvršena je dodatna obrada sekvenci i provera postojanja himeričnih sekvenci i na kraju klasifikacija. U ovom koraku dodeljena je taksonomija koja pripada Greengenes bazi podataka. Uklanjanjem sekvenci sa nepoznatim taksonom u V4 regionu, podaci su umanjeni za 23 sekvence. Za region V5 podaci su umanjeni za 25 sekvenci (tabela 5).

Tabela 5: Paralelni prikaz uklonjenih sekvenci korišćenjem Greengenes i SILVA baze podataka

Greengenes				SILVA		
	Broj sekvenci uklonjenih naredbom <i>screen.secs</i>	Broj sekvenci uklonjenih naredbom <i>remove.lineage</i>	Broj sekvenci nepoznatog taksona, (%)	Broj sekvenci uklonjenih naredbom <i>screen.secs</i>	Broj sekvenci uklonjenih naredbom <i>remove.lineage</i>	Broj sekvenci nepoznatog taksona, (%)
V4	22	23	4, (17,39)	0	29	10, (34,5)
V5	817	25	24, (96,0)	17	96	75, (78,1)

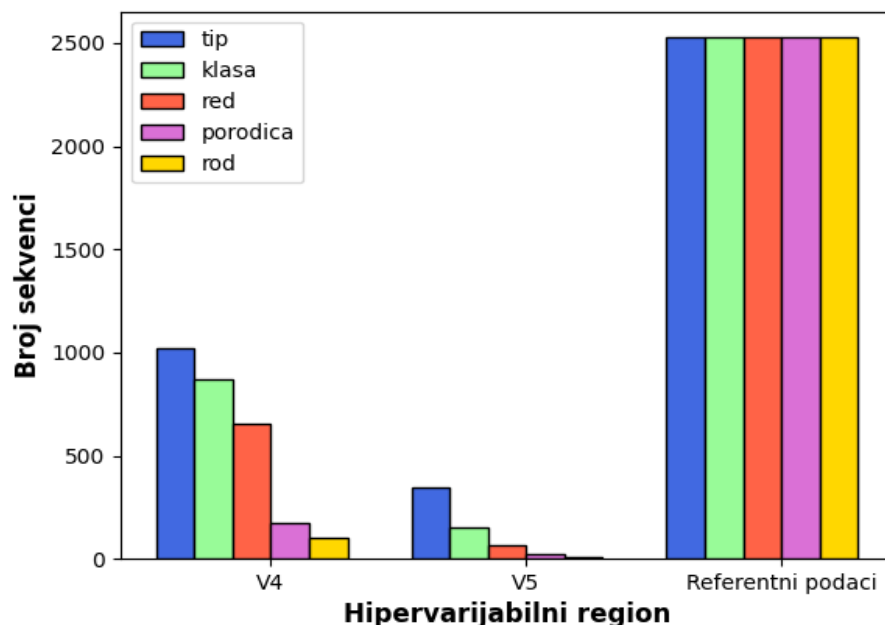
Iz podataka prikazanih u *tabeli 5* može se zaključiti da je za regione V4 i V5, 4 i 24 sekvence, respektivno, označeno kao nepoznato dodeljivanjem Greengenes taksonomije. Udeo nekultivisanih bakterija za ova dva regiona prikazan je na *slici 19*.



Slika 19: Udeo nekultivisanih bakterija [8] u ukupnom broju nepoznate takse

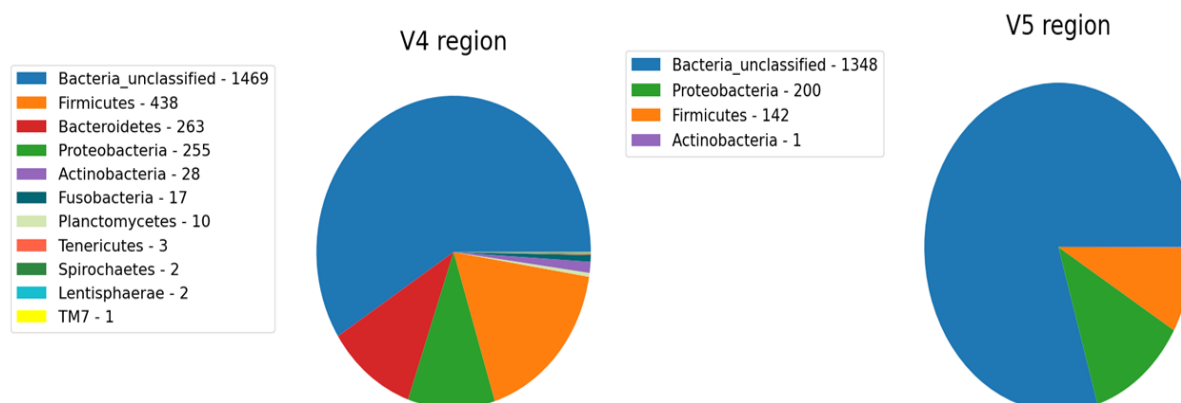
Bitno je napomenuti da, iako je udeo nekultivisanih sekvenci veliki, ipak je reč o manjem broju sekvenci u odnosu na SILVA referentnu bazu.

U daljem toku analize formirani su OTU-i, na osnovu dodeljene taksonomske identifikacije. Rezultati analize prikazani su grafički. (*slika 20*)



Slika 20: Grafički prikaz oporavljenih sekvenci bioinformatičkom metodom opisanom u ovom radu po taksnomskim nivoima (Greengenes referentna baza)

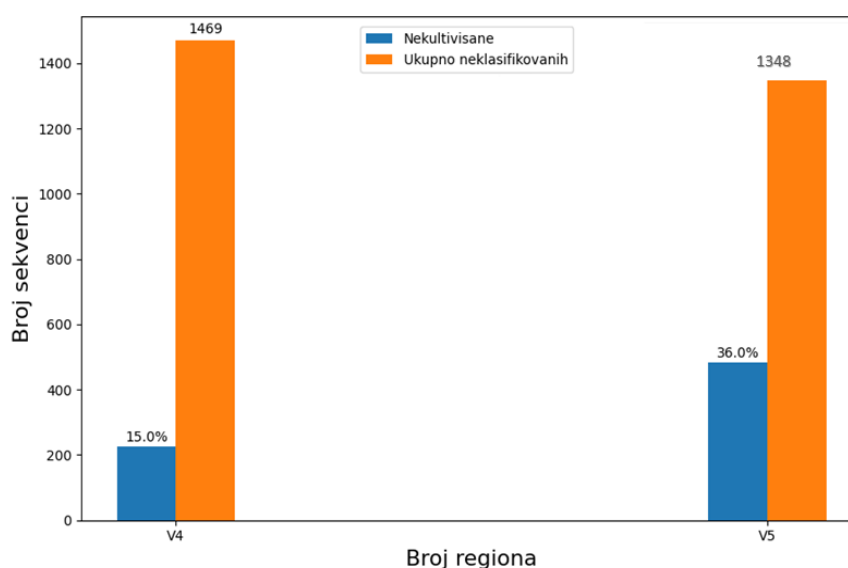
Iz grafičkog prikaza sa slike 20, može se zaključiti da za veliki broj sekvenci, iako je pronađen takson domena bakterija, nije klasifikovan takson tipa. Ovo je potvrđeno analizom nivoa tipa prikazanoj na slici 21, u vidu kružnog grafikona.



Slika 21: Grafički prikaz udela klasifikovanih taksona na nivou tipa za V4 i V5 analizirani region

Iako je veliki broj sekvenci neklasifikovan u regionu V4, korišćenjem Greengenes referentne baze uspešno su klasifikovani tipovi *TM7*, *Tenericutes* i *Lentisphaerae*. Ovi tipovi nisu pronađeni ni u jednom regionu dodeljivanjem taksonomije SILVA referentne baze (redovi označeni plavom bojom, tabela 4).

Udeo sekvenci nekultivisanih bakterija u ukupnom broju neklasifikovanih sekvenci prikazan je na slici 22.



Slika 22: Grafički prikaz udela sekvenci nekultivisanih bakterija [8] u ukupnom broju neklasifikovanih sekvenci

Na osnovu analize prikazane u ovom poglavlju, može se zaključiti da je identifikacijom bakterija korišćenjem Greengenes referentne baze potvrđeno malo manje od 50 % taksona tipa, u odnosu na referentne podatke. U taj procenat ulaze tipovi koje SILVA referentna baza nije uspela da potvrdi. Međutim, broj neklasifikovanih sekvenci je mnogo veći u odnosu na rezultate analize kada je korišćena SILVA referentna baza. Takođe, analiza samo dva od osam hipervarijabilnih regiona je bila moguća.

Zaključak

Na osnovu analize izvršene u ovom radu i prikazanih rezultata moguće je zaključiti sledeće:

1. Bioinformatička analiza primenjena za analizu sekvenci varijabilnih regiona 16S rRNK referentnih bakterija nije omogućila tačnu identifikaciju većine bakterija.
2. Rezultati taksonomske analize korišćenjem različitih varijabilnih regiona su međusobno veoma različiti.
3. Broj sekvenci koje se mogu identifikovati opada kada se klasifikacija izvodi na nižim taksonomskim kategorijama, tj. od tipa prema rodu.
4. Za precizniju klasifikaciju bakterija neophotno je posedovati dovoljno duge sirove sekvence. Veliki broj sekvenci hipervarijabilnih regiona V1 i V3 je bio kratak. Zbog odbacivanja ovih sekvenci kao nepodesnih za dalju analizu, veliki broj sekvenci V1 i V3 regiona nije bilo moguće identifikovati korišćenjem SILVA referentne baze.
5. Veliki broj sekvenci koji je odbačen ili neklasifikovan pripada bakterijama koje nisu kultivisane, ali varira između regiona.
6. Korišćenjem poslednje verzije SILVA referentne baze "pogrešno" je identifikovano pet novih tipova, koji nisu postojali među identifikovanim bakterijama u korišćenoj referentnoj studiji.
7. Zbog izuzetno lošeg poravnanja sa referentnom bazom, većina sekvenci nije mogla biti analizirana korišćenjem Greengenes baze. Samo su sekvence V4 i V5 regiona bile pogodne za analizu u okviru Greengenes baze.
8. Analiza V4 i V5 hipervarijabilnih regiona izvršena korišćenjem Greengenes referentne baze identifikovala je taksone tipa koji nisu potvrđeni korišćenjem SILVA referentne baze.

Literatura

1. Can, T., *Introduction to bioinformatics*. Methods Mol Biol, 2014. **1107**: p. 51-71.
2. Christopher P. Austin, M.D. *Bioinformatics*. 2021; Available from: <https://www.genome.gov/genetics-glossary/Bioinformatics>.
3. Xia, Y., J. Sun, and D.-G. Chen, *Statistical analysis of microbiome data with R*. Vol. 847. 2018: Springer.
4. Woese, C.R., *Bacterial evolution*. Microbiological reviews, 1987. **51**(2): p. 221-271.
5. Park, S.-C. and S. Won, *Evaluation of 16S rRNA Databases for Taxonomic Assignments Using Mock Community*. Genomics & informatics, 2018. **16**(4): p. e24-e24.
6. Bayat, A., *Science, medicine, and the future: Bioinformatics*. BMJ (Clinical research ed.), 2002. **324**(7344): p. 1018-1022.
7. Reller, L.B., M.P. Weinstein, and C.A. Petti, *Detection and Identification of Microorganisms by Gene Amplification and Sequencing*. Clinical Infectious Diseases, 2007. **44**(8): p. 1108-1114.
8. Rajilić-Stojanović, M. and W.M. De Vos, *The first 1000 cultured species of the human gastrointestinal microbiota*. FEMS microbiology reviews, 2014. **38**(5): p. 996-1047.
9. Schloss, P.D., et al., *Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities*. Applied and environmental microbiology, 2009. **75**(23): p. 7537-7541.
10. Sekulić, M.V., *MIKROBIOLOGIJA*, ed. D. Mijin. 2013: Tehnološko - metalurški fakultet, Univerzitet u Beogradu, Beograd, Karnegijeva 4. 211.
11. Meyer, A., et al., *Fast evolving 18S rRNA sequences from Solenogastres (Mollusca) resist standard PCR amplification and give new insights into mollusk substitution rate heterogeneity*. BMC Evolutionary Biology, 2010. **10**(1): p. 70.
12. Mitruka, B.M. and M.J. Bonner, *Methods of detection and identification of bacteria*. 2017: Crc Press.
13. *Different Size, Shape and Arrangement of Bacterial Cells*, in <https://microbiologyinfo.com/different-size-shape-and-arrangement-of-bacterial-cells/>.
14. Woese, C.R., O. Kandler, and M.L. Wheelis, *Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya*. Proc Natl Acad Sci U S A, 1990. **87**(12): p. 4576-9.
15. Franco-Duarte, R., et al., *Advances in Chemical and Biological Methods to Identify Microorganisms-From Past to Present*. Microorganisms, 2019. **7**(5): p. 130.
16. Woese, C.R. and G.E. Fox, *Phylogenetic structure of the prokaryotic domain: the primary kingdoms*. Proceedings of the National Academy of Sciences, 1977. **74**(11): p. 5088-5090.
17. Woese, C.R., *Interpreting the universal phylogenetic tree*. Proceedings of the National Academy of Sciences, 2000. **97**(15): p. 8392-8396.
18. Cech, T.R. and B.L. Bass, *BIOLOGICAL CATALYSIS BY RNA*. Annual Review of Biochemistry, 1986. **55**(1): p. 599-629.
19. Clarridge, J.E., *Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases*. Clinical Microbiology Reviews, 2004. **17**(4): p. 840-862.
20. Yarza, P., et al., *Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences*. Nature Reviews Microbiology, 2014. **12**(9): p. 635-645.
21. Bačić, A., *Analiza uticaja izbora referentne baze sekvenci i metode sekvenciranja 16S rRNK kodirajućeg gena na rezultat analize sastava crevne mikrobiote*, in *Department of Biochemical Engineering and Biotechnology*. 2020, University of Belgrade: Faculty of Technology and Metallurgy. p. 50.
22. Horton, T.R. and T.D. Bruns, *The molecular revolution in ectomycorrhizal ecology: peeking into the black-box*. Molecular ecology, 2001. **10**(8): p. 1855-1871.
23. Tyler, A.D., M.I. Smith, and M.S. Silverberg, *Analyzing the human microbiome: a "how to" guide for physicians*. Official journal of the American College of Gastroenterology| ACG, 2014. **109**(7): p. 983-993.
24. Heather, J.M. and B. Chain, *The sequence of sequencers: The history of sequencing DNA*. Genomics, 2016. **107**(1): p. 1-8.
25. Schloss, P.D., D. Gevers, and S.L. Westcott, *Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies*. PLOS ONE, 2011. **6**(12): p. e27310.
26. Haas, B.J., et al., *Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons*. Genome Res, 2011. **21**(3): p. 494-504.

27. Mysara, M., et al., *CATCH, an ensemble classifier for chimera detection in 16S rRNA sequencing studies*. Applied and environmental microbiology, 2015. **81**(5): p. 1573-1584.
28. EzBiome, I.; Available from: <https://help.ezbiocloud.net/mtp-pipeline/>.
29. Navas-Molina, J.A., et al., *Advancing our understanding of the human microbiome using QIIME*. Methods in enzymology, 2013. **531**: p. 371-444.
30. Wang, Q., et al., *Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy*. Applied and environmental microbiology, 2007. **73**(16): p. 5261-5267.
31. Schloss, P.D. and S.L. Westcott, *Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis*. Applied and environmental microbiology, 2011. **77**(10): p. 3219-3226.
32. Callahan, B.J., P.J. McMurdie, and S.P. Holmes, *Exact sequence variants should replace operational taxonomic units in marker-gene data analysis*. The ISME Journal, 2017. **11**(12): p. 2639-2643.
33. Kuczynski, J., et al., *Using QIIME to analyze 16S rRNA gene sequences from microbial communities*. Current protocols in bioinformatics, 2011. **Chapter 10**: p. Unit10.7-10.7.
34. Caporaso, J.G., et al., *QIIME allows analysis of high-throughput community sequencing data*. Nature Methods, 2010. **7**(5): p. 335-336.
35. Bolyen, E., et al., *Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2*. Nature Biotechnology, 2019. **37**: p. 1.
36. Edgar, R.C., *UPARSE: highly accurate OTU sequences from microbial amplicon reads*. Nat Methods, 2013. **10**(10): p. 996-8.
37. Westcott, S.L. and P.D. Schloss, *OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units*. mSphere, 2017. **2**(2).
38. Huson, D.H., et al., *MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data*. PLOS Computational Biology, 2016. **12**(6): p. e1004957.
39. Huson, D.H., et al., *MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs*. Biology Direct, 2018. **13**(1): p. 6.
40. Callahan, B.J., et al., *DADA2: High-resolution sample inference from Illumina amplicon data*. Nat Methods, 2016. **13**(7): p. 581-3.
41. mothur.org. *Greengenes formatted base*. 2013 [cited 2021; Available from: https://mothur.org/wiki/greengenes-formatted_databases/].
42. mothur.org. *SILVA reference base*. [cited 2021; Available from: https://mothur.org/wiki/silva_reference_files/].
43. Kozich, J.J., et al., *Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform*. Applied and environmental microbiology, 2013. **79**(17): p. 5112-5120.
44. Huse, S.M., et al., *Ironing out the wrinkles in the rare biosphere through improved OTU clustering*. Environmental microbiology, 2010. **12**(7): p. 1889-1898.
45. Rognes, T., et al., *VSEARCH: a versatile open source tool for metagenomics*. PeerJ, 2016. **4**: p. e2584.
46.

<i>Chimeras</i>	<i>file</i>	<i>format</i>	Available from:
https://web.archive.org/web/20150220030441/https://drive5.com/usearch/manual/uchimeout.html .			