SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 6161

Ocjena alata za identifikaciju vrsta u metagenomskim uzorcima

Josipa Lipovac

Zagreb, lipanj 2019.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA ZAVRŠNI RAD MODULA

Zagreb, 13. ožujka 2019.

ZAVRŠNI ZADATAK br. 6161

Pristupnik: **Josipa Lipovac (0036499172)**
Studij: Računarstvo
Modul: Računarska znanost

Zadatak: **Ocjena alata za identifikaciju vrsta u metagenomskim uzorcima**

Opis zadatka:

Identifikacija vrsta u metagenomskim uzorcima, tj. skupovima DNA očitanja dobivenih sekvenciranjem uzoraka iz okoliša, koji sadrže genetski materijal više različitih organizama, može biti korisna za dijagnozu bolesti, kontrolu kvalitete hrane i utvrđivanje štetnih nametnika na biljkama. Danas postoji više alata koji implementiraju različite metode za identifikaciju vrsta u metagenomskim uzorcima i na različite načine ocjenjuju svoju uspješnost. Većina tih alata razvijena je za rad sa kratkim očitanjima dobivenim sekvencerima druge generacije.
U ovom radu potrebno je osmisliti i implementirati sustav procjene kvalitete postupaka za identifikaciju vrsta u metagenomskim uzorcima te pomoću njega usporediti kvalitetu nekoliko najpopularnijih postojećih alata. Potrebno je definirati mjere kvalitete, pripremiti testne skupove podataka te implementirati alat za automatsku evaluaciju. Za evaluaciju potrebno je koristiti i sintetske skupove podataka, ali i javno dostupne stvarne skupove podataka. Skupovi za testiranje trebaju sadržavati očitanja dobivena sekvencerima druge generacije, kao i očitanja dobivena sekvencerima treće generacije.
Programski kod je potrebno komentirati i pri pisanju pratiti neki od standardnih stilova. Kompletnu aplikaciju postaviti na repozitorij Github.
U svezi dobivanja detaljnih informacija obratiti se Josipu Mariću, mag. ing.

Zadatak uručen pristupniku: 15. ožujka 2019.
Rok za predaju rada: 14. lipnja 2019.

Mentor:
Mile Šikić
Prof. dr. sc. Mile Šikić

Djelovođa:
Izv. prof. dr. sc. Tomislav Hrkać

Predsjednik odbora za završni rad modula:
Marko Čupić
Doc. dr. sc. Marko Čupić

Hvala mentoru Mili na silnoj podršci i inspiraciji te hvala Josipu Mariću na strpljenju i korisnim savjetima.

SADRŽAJ

1. UVOD

Razvitak metagenomike, odnosno znanosti u bilo kojem obliku, od velikog je značaja za čovjeka i njegovo zdravlje. Jedan od načina razvitka metagenomike, grane znanosti koja se bavi proučavanjem genetskih uzoraka iz okoliša, jest identifikacija novih vrsta u metagenomskim uzorcima. Ta identifikacija značajna je ponajprije u medicini pri dijagnozama bolesti koje moraju biti brze i točne. Na ljudsko zdravlje velik utjecaj ima i prehrana. Identifikacija vrsta svoju primjenu nalazi i u određivanju kvalitete hrane. Značajna je i za utvrđivanje štetnih nametnika na biljkama što dodatno doprinosi poboljšanju hrane koju čovjek svakodnevno konzumira. Kao što se već moglo i zaključiti, važan cilj metagenomike svakako je otkriće novih organizama, a ispravna klasifikacija uzoraka preuzetih iz same prirode izazov je svakog alata klasifikatora. Danas postoji nekoliko takvih alata koji se bave identifikacijom vrsta u uzorcima te na različit način klasificiraju genome te određuju sam uzorak.

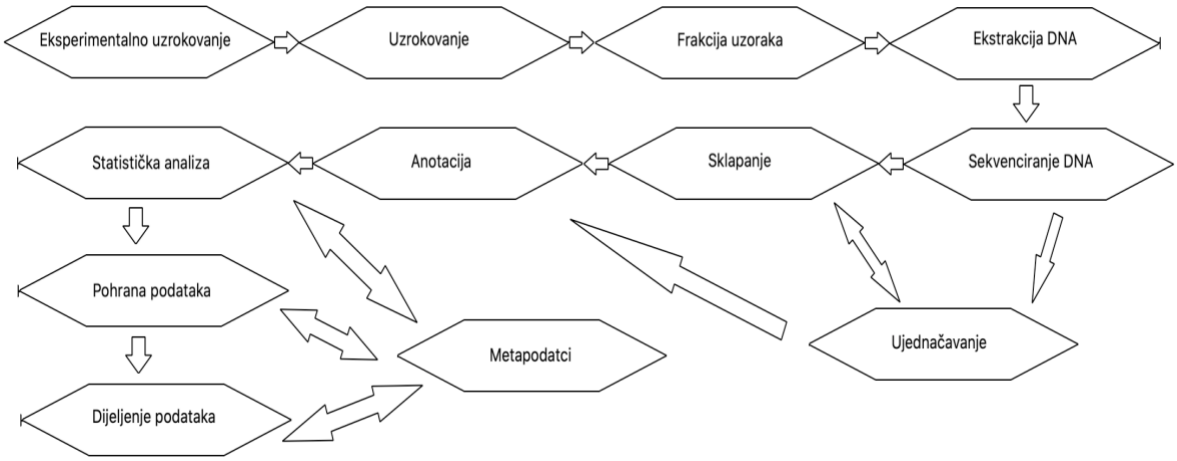
U ovom radu obradit će se funkcionalnost tri klasifikacijska alata, Kraken, MetaPhlAn2 te CLARK. Ocijenit će se njihov rad na temelju dva simulirana skupa podataka, jednog stvorenog od očitanja dobivenih sekvencerima druge generacije te drugog stvorenog od očitanja dobivenih sekvencerima treće generacije.

2. METAGENOMIKA – PREGLED PODRUČJA

Metagenomika je grana bioinformatike koja se bavi proučavanjem metagenoma, tj. genetskog uzorka dobivenog direktno iz okoliša. Radi se o području znanosti koje je postalo iznimno popularno u zadnjem desetljeću. Ona omogućuje pristup funkcionalnom sastavu gena mikrobnih zajednica te nam daje genetske informacije o mikrobnoj raznolikosti i ekologiji specifičnog okruženja [1].

Začetak metagenomike obilježen je početkom korištenja metode sekvenciranja pod nazivom *„shotgun"* ili neciljano sekvenciranje. Ta metoda zahtjeva velike računalne resurse, no u mogućnosti je odjednom obraditi više od jednog terabajta podataka, omogućuje sekvenciranje cijelog okolišnog uzorka te stvaranje cjelovitih sekvenci gena [2]. Dijagram toka *„shotgun"* metode prikazan je na slici (**Slika 2.1.**).

Glavna prednost *„shotgun"* metagenomike nad ostalim metodama jest ta što *„shotgun"* ima mogućnost analize uzoraka direktno prikupljenih iz okoliša [2]. Točnije, te uzorke nije potrebno uzgajati u laboratorijima.



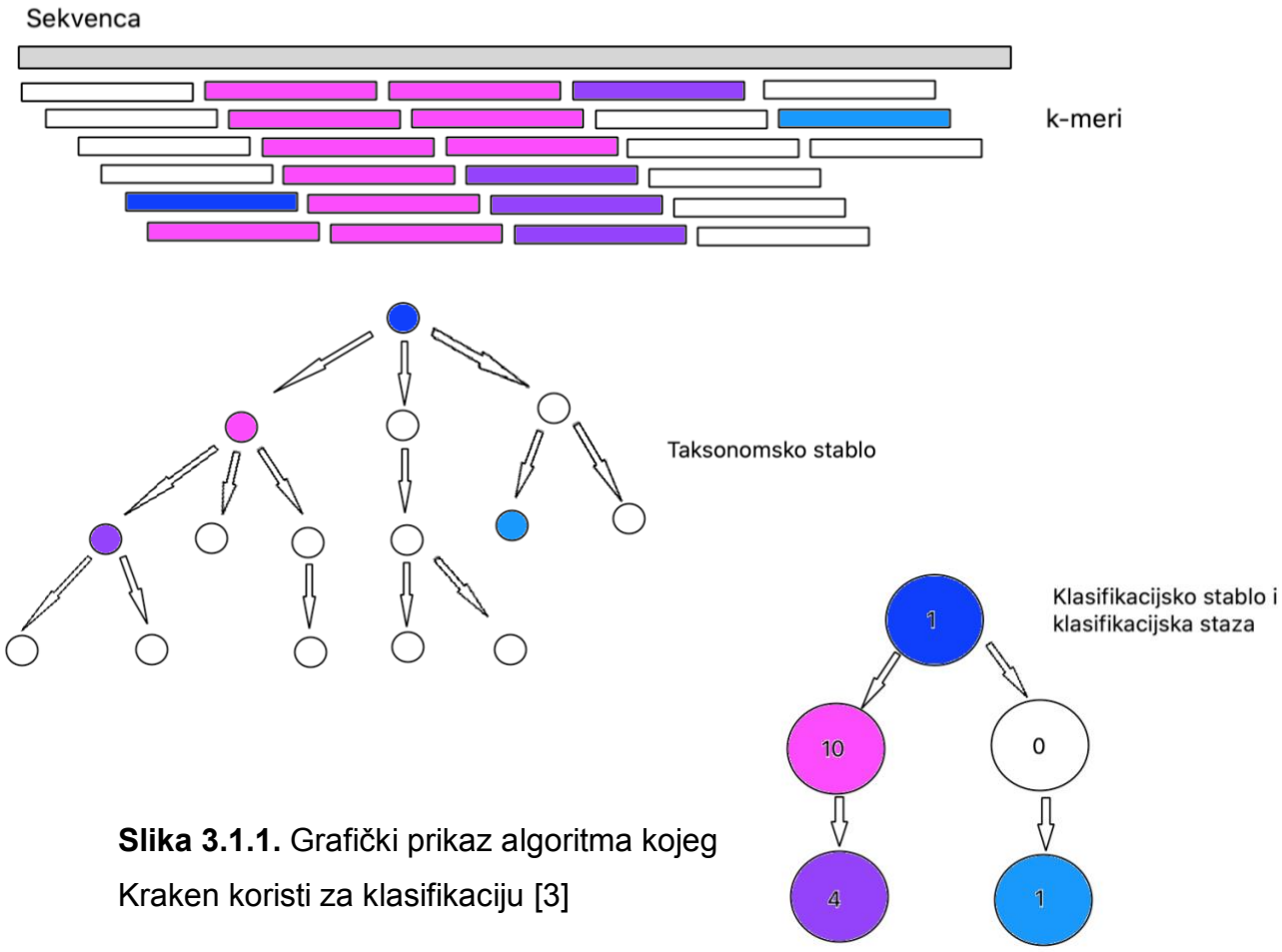
Slika 2.1. Dijagram toka „shotgun" metode [1]

Poboljšanje tehnike sekvenciranja DNA dovodi i do poboljšanja metagenomskog istraživanja. Upravo to dovodi do razvitka alata za analizu taksonomskih baza podataka te identificiranje vrsta u metagenomskim uzorcima što može biti od velike koristi za određivanje dijagnoze bolesti ili kvalitete hrane, otkrivanje novih vrsta mikroorganizama, utvrđivanje štetnih nametnika na biljkama te mnogo toga ostaloga.

3. ALATI ZA KLASIFIKACIJU

3.1. Kraken

Kraken je program za dodjeljivanje taksonomskih oznaka metagenomskim slijedovima DNA. Kraken klasificira očitanja tako da ih dijeli na preklapajuće k-mere. Svaki k-mer u nizu mapira se na najniži zajednički predak genoma koji sadrži taj k-mer u bazi podataka [3]. Za samu klasifikaciju koristi se klasifikacijsko stablo (**slika 3.1.1.**). Njega tvore svojte povezane s k-merima sekvence.

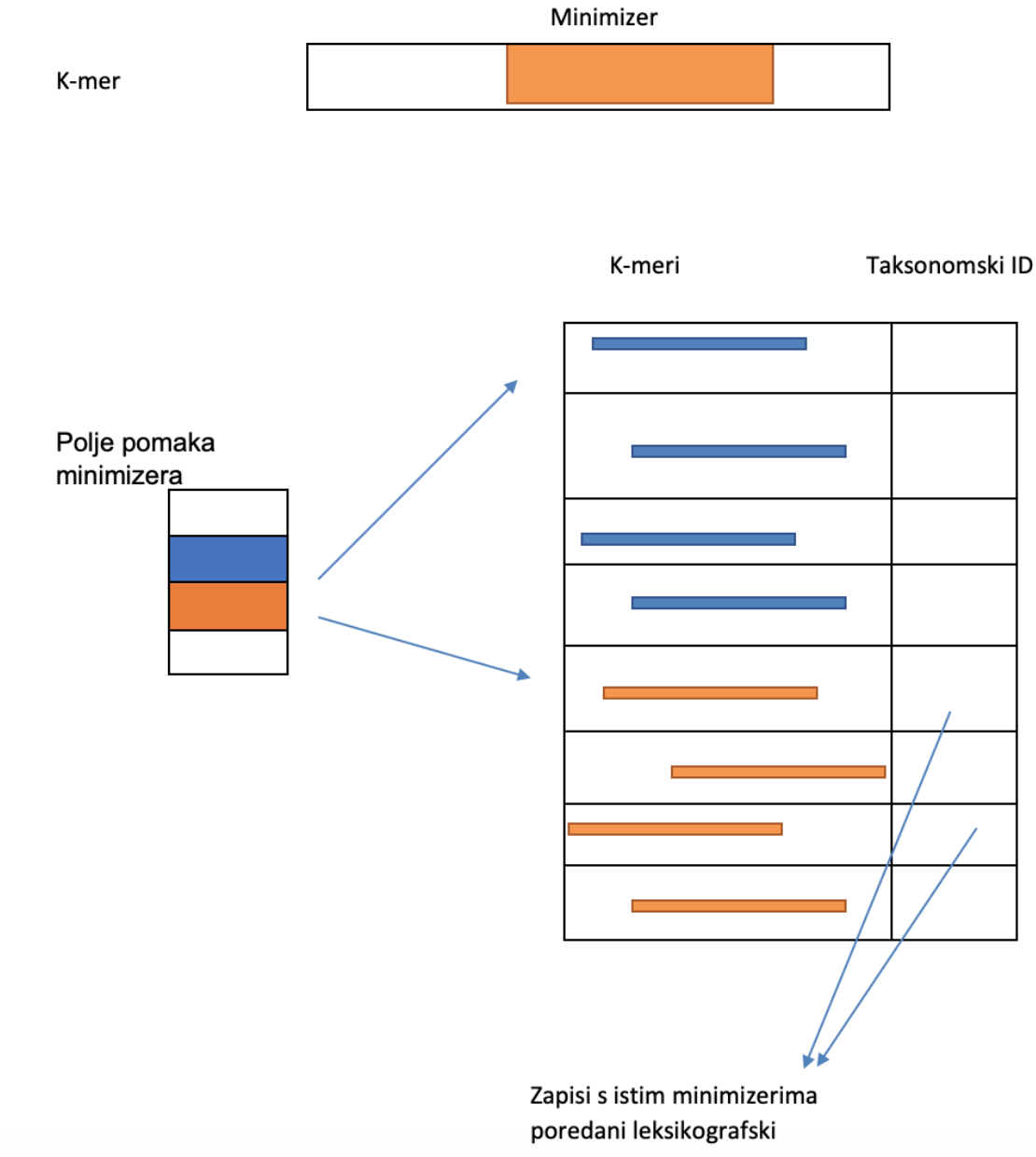


Slika 3.1.1. Grafički prikaz algoritma kojeg Kraken koristi za klasifikaciju [3]

Kraken se oslanja na bazu podataka (**slika 3.1.2.**) koja sadrži zapise koji se sastoje od k-mera i analize životnog ciklusa svih organizama čiji genomi sadrže baš taj k-mer [3]. Knjižnica genoma uz pomoć koje je baza podataka izgrađena, omogućava brzo traženje najspecifičnijeg čvora u taksonomskom stablu koje je povezano s danim k-merom. Sekvence se klasificiraju na način da šalju upite bazi podataka za svaki k-mer u nizu, a zatim koriste analizu životnog ciklusa taksona. Taksoni povezani s k-merima sekvence zajedno s pretcima taksona tvore podrezano stablo taksonomije [3]. Svaki čvor tog stabla ima težinu jednaku broju k-mera u slijedu povezanom s taksonom čvora. One sekvence koje se ne nalaze u bazi podataka Kraken smatra kao neklasificirane jer se očekuje da selektivni klasifikator ima veću preciznost uz neke troškove. Kraken gradi bazu podataka s veličinom k-mera k = 31, no ta se vrijednost može modificirati od strane korisnika. Svakoj putanji od korijena do lista u klasifikacijskom stablu daje se ocjena tako da se dodaju sve težine putanje. Putanja s maksimalnom ocjenom naziva se klasifikacijski put.

Od ključne važnosti u samoj klasifikaciji je brzina klasifikacije. Da bi postigao maksimalnu brzinu Kraken se mora izvoditi na računalu s dovoljno RAM-a za skladištenje cijele baze podataka. Cijela baza podataka zahtijeva 70 GB RAM-a [3]. Postoji i umanjena baza podataka koja je razvijena na način da su k-meri uklonjeni iz baze. Ta verzija baze podataka naziva se MiniKraken te joj je za nju osobito to što teže prepoznaje kratka očitanja. Uz MiniKraken postoje još i verzije Kraken-Q i MiniKraken-Q te Kraken-GB [3].

Cjelovita Krakenova baza kreira se u više koraka, počevši s izborom knjižnice genomskih sekvenci. Knjižnica se temelji na kompletnim mikrobnim genomima u RefSeq bazi Nacionalnog Centra za Biotehnološke Informacije (NCBI) [4]. Po potrebi knjižnica se može prilagoditi pojedinim korisnicima. Sljedeći korak jest korištenje višestrukog Jellyfish [5] k-mer brojača koji stvori bazu podataka sačinjenu od različitih 31-mera. Nakon toga pohranjuju se taksonomske identifikacijske oznake analize životnog ciklusa k-mera. Tada se svaka sekvenca povezuje s njoj vezanim taksonom. Informacije o taksonu dobivene su iz NBCI taksonomske baze podataka.



Slika 3.1.2. Prikaz organizacije baze podataka kod Kraken-a [3]

Za grupiranje sličnih k-mera koristi se minimizer koncept iz razloga što Kraken vrlo često za upite bazi podataka koristi k-mer te jer susjedni k-meri dijele značajnu količinu sekvence [3]. U Krakenovoj bazi podataka, svi k-meri s istim minimizerom pohranjuju se uzastopno i razvrstavaju se leksikografskim poretkom svojih kanonskih prikaza.

U SRR8115247.1.2 0 300 0:270

C SRR8115247.5.1 549 300 53335:58 549:3 53335:11 549:82 53335:5 549:2 91347:2 53335:5 549:22 53335:3 91347:5 0:31 549:8 0:33

C SRR8115247.5.2 549 300 0:46 549:4 0:220

C SRR8115247.6.1 549 301 549:271

C SRR8115247.6.2 549 301 549:177 0:31 549:20 0:36 549:6 0:1

C SRR8115247.7.1 549 301 549:25 53335:11 549:31 53335:5 549:31 0:43 549:116 53335:8 549:1

Dodatak 3.1.1 Primjer izlazne datoteke Kraken-a

Izlazna datoteka koju Kraken generira sadrži podatke klasifikacije. U gonjem ispisu vidimo primjer dijela Krakenovog ispisa. Ispis se sastoji od pet podataka odijeljenih tabulator graničnikom. Prvi podatak U/C označava je li sekvenca klasificirana ili ne. Drugi podatak jest identifikacijska oznaka sekvence, treći je identifikacijska oznaka taksona, a četvrti duljina sekvence u bp (*base pair*). Peti dio zapravo je skup podataka koji označava mapiranje svakog k-mera u nizu. Na primjeru klasificirane sekvence s identifikacijskom oznakom SRR8115247.5.1 (**Dodatak 3.1.1**) možemo vidjeti da je identifikacijska oznaka njezinog taksona 549, duljina sekvence 300bp, a da je prvih 58 k-mera mapirano na identifikacijsku oznaku taksona 53335, slijedeća 3 k-mera poravnata su na oznaku 549, nakon toga 11 ih je ponovno poravnato na 53335 i na kraju 82 k-mera je poravnato na taksonomsku identifikacijsku oznaku 549 [6].

Moguće je generirati još neke od izlaznih datoteka, npr. Kraken prevoditelj koji ispisuje puni taksonomski naziv pridružen svakom ulaznom nizu. Uz prevoditelj postoji i Kraken izvještaj pomoću kojeg je lakše vizualizirati rezultate na cijelom uzorku.

3.2. MetaPhlAn2

MetaPhlan2 (*metagenomic philogenetic analysis*) je sljedeći alat za klasifikaciju metagenomskih uzoraka koji klasifikaciju izvodi na totalno drugačiji način od prije spomenutog Krakena. Riječ je o alatu za profiliranje sastava mikrobnih zajednica (bakterija, arheja, eukariota i virusa) iz metagenomskih podataka dobivenih metodom *"shotgun"* [7]. Ima mogućnost identifikacije specifičnih sojeva te praćenja sojeva u uzorcima svih vrsta.

Marker geni su geni ili segmenti DNA koji imaju poznatu lokaciju na kromosomu. Koriste se za identifikaciju jedinki, populacija ili vrsta. MetaPhlAn2 se oslanja na približno 1 milijun jedinstvenih marker gena specifičnih za određene grupe organizama identificirane iz oko 17 000 referentnih genoma. Marker geni specifični za određene grupe organizama jesu kodirajuće sekvence koje se snažno čuvaju unutar genoma grupe i nemaju značajnu lokalnu sličnost s bilo kojim slijedom izvan spomenute grupe. Katalog gena eliminira gene koji se ne mogu nedvosmisleno povezati s pripadajućom grupom [7].

MetaPhlAn2 normalizira ukupan broj očitanja u svakoj klasi po duljini nukleotida svojih markera. Time osigurava relativnu brojnost svake taksonomske jedinice, uzimajući u obzir sve oznake specifične za podgrupu organizama [9].

Sekvence specifičnih markera računaju se na način da ih odabiru kodirajuće sekvence koje nedvosmisleno identificiraju mikrobne klase na vrsti ili višoj taksonomskoj razini. MetaPhlAn2 procjenjuje relativnu zastupljenost mikrobnih stanica mapiranjem očitanja na reduciran skupu tih sekvenci, tj. specifičnih markera. Alat uspoređuje svako metagenomsko očitavanje iz uzorka s katalogom marker gena kako bi identificirao podudarnost. Prethodna obrada metagenomske DNA nije potrebna jer je mala vjerojatnost da će lažna očitanja imati značajne podudarnosti s markerskim slijedom. Mikrobiološki podatci koji pripadaju određenim grupama organizama bez dostupnih sekvencijskih genoma prikazani su kao neklasificirana podgrupa najbližeg pretka s dostupnim podatcima o sekvenci [7].

Kod ovog alata omogućeni su nedvosmisleni taksonomski zadatci, točne procjene relativne brojnosti organizama te rezolucija na razini vrsta za bakterije, arheje, eukariote te viruse. Također, moguća je identifikacija i praćenje soja. S druge

strane postojeće taksonomske metode profiliranja neučinkovite su za veće skupove podataka [9].

MetaPhlAn2 pri svojoj klasifikaciji koristi alat Bowtie2. Riječ je o alatu za poravnavanje očitanja sekvenci na duge referentne sekvence [10].

SRR8115247.19.1_19_length=301	gi|372137338|ref|NZ_ADWZ01000003.1|:c21145-20258
SRR8115247.177.1_177_length=301	gi|372137336|ref|NZ_ADWZ01000001.1|:647366-647731
SRR8115247.237.2_237_length=300	gi|389822529|ref|NZ_BAEF01000001.1|:504752-505138
SRR8115247.629.1_629_length=301	gi|372137337|ref|NZ_ADWZ01000002.1|:c180940-180593
SRR8115247.629.2_629_length=300	gi|372137337|ref|NZ_ADWZ01000002.1|:c180940-180593
SRR8115247.739.1_739_length=301	gi|372137357|ref|NZ_ADWZ01000022.1|:c1446-511
SRR8115247.739.2_739_length=301	gi|372137357|ref|NZ_ADWZ01000022.1|:c1446-511

Dodatak 3.2.1. Primjer Bowitie2 izlazne datoteke

Metaphlan2 generira dvije izlazne datoteke. Prva izlazna datoteka (**Dodatak 3.2.1.**) primjer je Bowitie2 izlazne datoteke te sadrži rezultate međusobnog mapiranja u jedinstvene marker gene sekvenci [8].

```
#SampleID	Metaphlan2_Analysis
k__Bacteria	100.0
k__Bacteria|p__Firmicutes	75.77506
k__Bacteria|p__Bacteroidetes	24.22494
k__Bacteria|p__Firmicutes|c__Negativicutes	53.19418
k__Bacteria|p__Bacteroidetes|c__Bacteroidia	24.22494
k__Bacteria|p__Firmicutes|c__Bacilli	22.58087
k__Bacteria|p__Firmicutes|c__Negativicutes|o__Selenomonadales	53.19418
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales	24.22494
k__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacillales	22.58087
k__Bacteria|p__Firmicutes|c__Negativicutes|o__Selenomonadales|f__Veillonellaceae	53.19418
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Prevotellaceae	24.22494
k__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacillales|f__Streptococcaceae	22.58087
k__Bacteria|p__Firmicutes|c__Negativicutes|o__Selenomonadales|f__Veillonellaceae|g__Veillonella	53.19418
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Prevotellaceae|g__Prevotella	24.22494
k__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacillales|f__Streptococcaceae|g__Streptococcus	22.58087
k__Bacteria|p__Firmicutes|c__Negativicutes|o__Selenomonadales|f__Veillonellaceae|g__Veillonella|s__Veillonella_unclassified	35.84554
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Prevotellaceae|g__Prevotella|s__Prevotella_histicola	21.32804
k__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacillales|f__Streptococcaceae|g__Streptococcus|s__Streptococcus_parasanguinis	19.37632
k__Bacteria|p__Firmicutes|c__Negativicutes|o__Selenomonadales|f__Veillonellaceae|g__Veillonella|s__Veillonella_atypica	17.34864
k__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacillales|f__Streptococcaceae|g__Streptococcus|s__Streptococcus_salivarius	3.20456
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Prevotellaceae|g__Prevotella|s__Prevotella_melaninogenica	2.8969
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Prevotellaceae|g__Prevotella|s__Prevotella_histicola|t__GCF_000234055	21.32804
k__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacillales|f__Streptococcaceae|g__Streptococcus|s__Streptococcus_parasanguinis|t__Streptococcus_parasanguinis_unclassified	19.37632
k__Bacteria|p__Firmicutes|c__Negativicutes|o__Selenomonadales|f__Veillonellaceae|g__Veillonella|s__Veillonella_atypica|t__Veillonella_atypica_unclassified	17.34864
k__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacillales|f__Streptococcaceae|g__Streptococcus|s__Streptococcus_salivarius|t__Streptococcus_salivarius_unclassified	3.20456
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Prevotellaceae|g__Prevotella|s__Prevotella_melaninogenica|t__Prevotella_melaninogenica_unclassified	2.8969
```

Dodatak 3.2.2. Primjer MetaPhlAn2 izlazne datoteke

Druga izlazna datoteka (**Dodatak 3.2.2.**) sadrži konačne izračunate brojnosti organizama. Prvi stupac prikazuje grupe organizama, od taksonomskog kraljevstva

(bakterije, arheje, eukarioti, itd.) do vrste. Zbroj svake razine iznosi 100%. Svaka taksonomska razina određene grupe ima svoj predznak:

k_ - kraljevstvo

p_ - koljeno

c_ - klasa

o_ - red

f_ - porodica

s_ - vrsta [8][9].

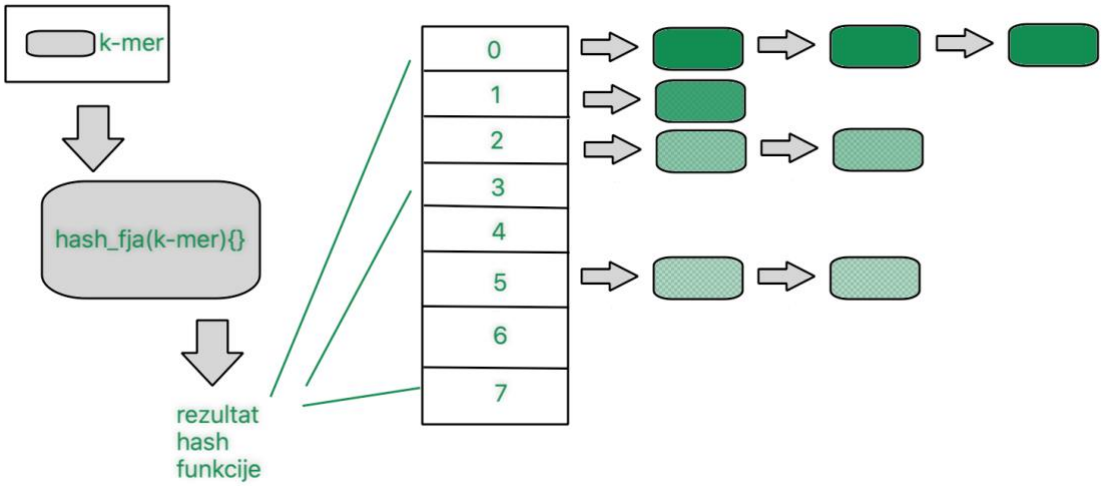
3.3. CLARK

Posljednji alat za klasificiranje metagenomskih očitanja koji će biti obrađen u ovom radu naziva se CLARK (CLAssifier based on Reduced K-mers). CLARK, kao i Kraken, klasificira objekte na temelju reduciranih skupova k-mera.

Prvi korak kod pokretanja samog alata je priprema. U tom koraku CLARK gradi veliki indeks koji sadrži k-spektre svih ciljnih sekvenci, odnosno sekvenci koje koristimo za klasifikaciju. K-spektar nekog niza jest vektor veličine 4k koji prikuplja broj ponavljanja svih mogućih k-mera u tom nizu. Radi se o sažetom prikazu niza koji omogućuje usporedbu sekvenci. Kad su svi k-spektri ciljnih sekvenci sakupljeni u indeksu, CLARK uklanja sve zajedničke k-mere između ciljeva. Preostali k-meri su ili ciljno specifični ili diskriminativni. Oni predstavljaju genomske regije koje jedinstveno karakteriziraju cilj. Cilj se dodijeli objektu s kojim se dijeli najveći broj k-mera [12].

Pri samoj klasifikaciji potrebno je odabrati optimalan k. Kada je k iz raspona [19,22] maksimizirani su preciznost, osjetljivost, pouzdanost te brzina. Povećanjem k povećavamo preciznost i pouzdanost, no smanjuju se osjetljivost te brzina dodjele.

CLARK osim uobičajenih FASTA i FASTQ ulaza, prihvaća i tekstualnu datoteku koja sadrži k-mer distribuciju, tj. svaki redak te datoteke sastoji se od k-mera te njegovog broja ponavljanja. Indeks iz ciljnih sekvenci gradi se u slučaju da već ne postoji za tu ulaznu datoteku [13]. Korisnik ima mogućnost klasificirati podatke na razini bilo kojeg taksonomskog ranga, no tada se od njega očekuje da generira ciljeve grupiranjem genoma istog roda ili ciljeve koji posjeduju istu taksonomsku oznaku [13].



Slika 3.3.1. Grafički prikaz tablice raspršenog adresiranja

Indeks se pohranjuje u obliku tablice raspršenog adresiranja ili „hash-tablice“ (**slika 3.3.1.**). U „hash-tablicu“ pohranjene su identifikacijske oznake za cilj koji sadrži određeni k-mer, broj različitih ciljeva koji sadrže taj k-mer te broj pojavljivanja tog k-mera u svim ciljevima. Za rješavanje kolizije „hash-tablica“ koristi zasebno ulančavanje, točnije metodu kojoj je ideja da svaka ćelija „hash-tablice“ pokaže na povezani popis zapisa koji imaju istu vrijednost „hash“ funkcije. CLARK zatim uklanja bilo koji k-mer koji se pojavljuje u više od jednog cilja [12]. Korisnik ima mogućnost navesti minimalan broj pojavljivanja k-mera pa na temelju toga k-meri u indeksu također mogu biti uklonjeni. To je prednost kad se je riječ o klasificiranju skupova niske kvalitete. Uklanjanjem zajedničkih k-mera smanjuje se šum u procesu klasifikacije. To rezultira time da CLARK ne stvara stablo taksonomije za klasificiranje objekata (kao npr. KRAKEN), već koristi jednu "ravnu" razinu [12]. Rezultirajući skupovi k-mera specifičnih za cilj pohranjuju se u disk za sljedeću fazu.

Prethodno opisani postupak koristi prošireni način rada CLARK-a. Uz njega postoji i zadani način rada. U tom načinu rada CLARK čim postoji barem jedan cilj koji prikuplja najmanje polovicu ukupnih mogućih pogodaka zaustavlja upite. U glavnu memoriju učitava se polovica ciljnih k-mera. CLARK to postiže tako da preskaće k-mere specifične za cilj na temelju njihovih pozicija indeksa [13].

Za korisnike s ograničenom količinom RAM-a postoji i verzija CLARK-l (*„light"*) koja zahtjeva daleko manje RAM-a, no učinkovitost je po pretpostavci

podjednaka. Prednost ove verzije CLARK-a jeste da vrlo brzo gradi „hash-tablicu“ [12].

S druge strane ako je korisniku brzina od velike važnosti postoji verzija CLARK-E. Budući da koristimo k-mer skupove specifične za cilj, CLARK-E temelji se na činjenici da ako objekt potječe iz jednog od ciljeva onda će biti pogođen ili k-mer iz tog objekta ili neće biti pogotka uopće [13]. Na taj način smanjuje se broj upita u bazu podataka i povećava brzina izvođenja.

Object_ID, Length, Assignment
SRR6982805.2.1,10000,198620
SRR6982805.9.1,10000,198620
SRR6982805.10.1,10000,198620
SRR6982805.11.1,10000,198620
SRR6982805.12.1,10000,198620
SRR6982805.13.1,10000,198620
SRR6982805.14.1,10000,NA
SRR6982805.15.1,10000,198620
SRR6982805.16.1,10000,198620
SRR6982805.17.1,10000,198620
SRR6982805.18.1,10000,287

Dodatka 3.3.1. Primjer CLARK izlazne datoteke

Izlazna datoteka za zadani način rada sastoji se od 3 stupca (**Dodatak 3.3.1.**). U prvom stupcu nalazi se identifikacijska oznaka objekta, u drugom je duljina objekta, a u trećem se nalazi taksonomska identifikacijska oznaka koja je dodijeljena određenom objektu. Oznaka "NA" označava neklasificirani objekt.

4. SIMULATOR

4.1. Stvaranje testnog skupa podataka

S ciljem ispitivanja mogućnosti gore navedenih alata za klasifikaciju metagenomskih uzoraka za prvi korak bilo je potrebno stvoriti skup podataka koji će predstavljati upravo jedan od uzoraka. Upravo nad tim uzorkom provedena je klasifikacija korištenjem navedenih alata.

Klasifikacija se trebala provesti na dvije vrste testnih skupova podataka. Prvi skup je bio sastavljen od sekvenci dobivenih sekvenciranjem druge generacije. U tom skupu korištene su sekvence dobivene Illumina sekvenciranjem. Sekvenciranje druge generacije daje kratka očitanja, dok sekvenceri treće generacije generiraju sekvence dugih očitanja s većom pogreškom. Drugi skup sastojao se od sekvenci dobivenih upravo sekvenciranjem treće generacije te su tu korišteni rezultati dobiveni Oxford Nanopore sekvenciranjem.

	Pantonea agglomerans	**Pseudomonas koreensis**	**Klebsiella pneumoniae**
Taksonomski ID	549	198620	573

Tablica 4.1.1. Bakterije korištene za izradu testnog skupa podataka te njihove pripadne taksonomske identifikacijske oznake

Oba testna skupa sastojala su se od sekvenci tri različite bakterije: Pantonea agglomerans [15], Pseudomonas koreensis [16] te Klebsiella pneumoniae [17]. S ciljem ne zauzimanja velike količine memorije, za klasifikatore je odlučeno da će koristiti samo baze podataka koje se sastoje od bakterija. Iz tog razloga u cijeli uzorak dodan je i dio sekvenci čovjekovog kromosoma. Taj uzorak trebao bi predstavljati šum, odnosno dio sekvenci koje očekujemo da klasifikatori neće prepoznati te da će ostati neklasificirani.

4.2. Programska implementacija simulatora

Programsko rješenje za simuliranje testnog skupa podataka izvedeno je u obliku Python 3.6.5. skripte te je dostupno na GitHub poveznici: https://github.com/lbcb-edu/BSc-thesis-18-19/tree/jlipovac . Skripta kao argumente prima nazive datoteka u FASTA formatu triju bakterija te ljudskog kromosoma, spomenutih u prijašnjem poglavlju. Uz nazive datoteka, skripta prima i postotak koliko sekvenci želimo uzeti iz koje datoteke te željeni prefiks naziva izlaznih datoteka.

Broj sekvenci koje će imati izlazna datoteka određuje postotak koji se zadaje kao ulazni argument. Skripta iz svake ulazne datoteke uzima taj postotak sekvenci od ukupnog broja sekvenci unutar te datoteke.

Čitanje datoteka te izdvajanje sekvenci provedeno je korištenjem Biopythona, točnije alata za biološko računanje [14].

Skripta svakoj odabranoj sekvenci iz primljenih datoteka pridružuje valjanu taksonomsku identifikacijsku oznaku (**Tablica 4.1.1.**). Sekvence ljudskog kromosoma označene su taksonomskom oznakom -10 kako bi se naglasilo da se radi o šumu, odnosno sekvencama za koje se ne očekuje klasificiranje.

```
>ERR1008684.1.2 1 length=125
TATGTTCGAGTGGAGTCCGCCGTGTCACTTTCGCTTTGGCAGCAGTGTCTTGCCCGATTG
CAGGATGAGTTACCAGCCACAGAATTCAGCATGTGGATCCGCCCATTGCAGGCGGAACTG
AGCGA
>ERR1008684.5.1 5 length=125
TCGAGTGGAGTCCGCCGTGTCACTTTCGCTTTGGCAGCAGTGTCTTGCCCGATTGCAGGA
TGAGTTACCAGCCACAGAATTCAGCATGTGGATCCGCCCATTGCAGGCGGAACTGAGCGA
TAACA
```

Dodatak 4.2.1. Primjer FASTA formata iz generirane izlazne datoteke

Skripta generira tri izlazne datoteke. Prva je u FASTA obliku (**Dodatak 4.2.1.**). Riječ je o generiranom testnom skupu podataka nad kojim se pokreću klasifikatori. Sljedeća generirana datoteka sadržava popis taksonomskih

identifikacijskih oznaka u testnom skupu podataka. Svakoj taksonomskoj oznaci dodijeljen je postotak sekvenci koje se nalaze u testnom skupu podataka te imaju tu taksonomsku oznaku. Uz te dvije datoteke skripta simulatora generira i datoteku u kojoj se nalazi popis svih identifikacijskih oznaka sekvenci te njima pridruženih taksonomskih oznaka. Ta datoteka koristit će se za točnu usporedbu s izlaznim datotekama klasifikatora.

Organizam	Pantonea agglomerans	Pseudomonas koreensis	Klebsiella pneumoniae	Ljudski kromosom
Taksonomski ID	549	19862	573	-10
Udio u izlaznoj datoteci	24.32%	41.84%	29.69%	4.15%

Tablica 4.2.1. Udio pojedine bakterije u testnom skupu stvorenom od kratkih očitanja – sekvenciranje druge generacije

Kod stvaranja testnog skupa podataka nastalih sekvenciranjem druge generacije postotak koji je poslan kao argument skripti simulatora iznosio je 15%. Kao što je već prethodno navedeno iz svake ulazne datoteke od ukupnog broja sekvenci 15% nasumičnih je prihvaćeno za izlaznu datoteku. Na takav način dobiveni su podatci navedeni u tablici (**Tablica 4.2.1.**).

Organizam	Pantonea agglomerans	Pseudomonas koreensis	Klebsiella pneumoniae	Ljudski kromosom
Taksonomski ID	549	19862	573	-10
Udio u izlaznoj datoteci	26.49%	24.55%	19.43%	29.53%

Tablica 4.2.2. Udio pojedine bakterije u testnom skupu stvorenom od dugih očitanja – sekvenciranje treće generacije

Pri stvaranju testnog skupa podataka nastalih sekvenciranjem treće generacije koristio se isti postupak, no izabrani postotak iznosio je 80%. Daleko veći postotak odabran je s ciljem da se u izlaznoj datoteci generira podjednaki broj sekvenci kao u izlaznoj datoteci kratkih očitanja. Naime, ulazne datoteke s dugim očitanjima imale su manji broj sekvenci od ulaznih datoteka kratkih očitanja. Podjednak broj sekvenci bio je potreban kako bi se mogla donijeti krajnja usporedba klasifikatora na temelju klasificiranja dugih i kratkih očitanja. Podjela pojedinačnih udjela navedena je u tablici. (**Tablica 4.2.2.**)

5. PROVOĐENJE KLASIFIKACIJE

Nad testnim skupovima podataka koje je simulator generirao zatim je provedena klasifikacija koristeći alate Kraken i CLARK. Klasifikacija je provedena korištenjem i programa MetaPhlAn2. Naime, MetaPhlAn2 u svojoj izlaznoj datoteci generira popis marker gena koje je potrebno povezati s bazom podataka NBCI s ciljem dobivanja taksonomske identifikacijske oznake. Povezivanje s bazom zahtijevalo je lokalno pohranjivanje cijele baze podataka što je u ovom trenutku stvaralo prepreku te je evaluacija MetaPhlAn-a u okvirima ovog rada na kraju ipak izostavljena.

Budući da je evaluacija MetaPhlAn-a izostavljena, prije same obrade evaluacije te evaluiranih podataka potrebno je svakako dati generalan komentar na izlazne podatke tog alata. Naime, osim već spomenutog problema povezivanja baze podataka s taksonomskom oznakom bakterije, klasifikacija MetaPhlAn-a generirala je izlazne datoteke u kojima je veoma malen broj sekvenci klasificiran. U slučaju kratkih očitanja radi se o nešto većem broju klasificiranih sekvenci nego u slučaju testnog skupa sastavljenog od dugih očitanja. U prvom slučaju, odnosno u slučaju kratkih očitanja klasificirano je nešto više od 400 sekvenci, dok u drugom kada su korištena duga očitanja svega njih desetak.

6. EVALUATOR

6.1. Programska implementacija evaluacije

Kako bi se provela evaluacija alata bilo je potrebno osmisliti skriptu koja bi obrađivala podatke izlaznih datoteka simulatora te podatke iz njih uspoređivala s podatcima dobivenim obradom izlaznih datoteka klasifikacije. Programsko rješenje izvedeno je koristeći Python 3.6.5 te je također dostupno na GitHub poveznici: https://github.com/lbcb-edu/BSc-thesis-18-19/tree/jlipovac .

Skripta evaluatora kao ulazne argumente primala je izlaznu datoteku Kraken-a te izlaznu datoteku CLARK-a. Uz te dvije datoteke primala je i ostale datoteke koje je prethodno generirao simulator. Posljednji argument je željeni naziv izlazne datoteke evaluatora.

Prvotno je trebalo obraditi podatke izlaznih datoteka simulatora. Podatci su pročitani iz datoteka te pohranjeni u određene strukture. Sljedeći korak bio je učitati izlazne datoteke samih klasifikatora. Datoteke su generirane na način da ih se lako moglo pročitati korištenjem csv čitača. Pri čitanju datoteka bilo je potrebno pohranjivati sljedeće podatke:

- broj očitanih sekvenci,
- broj klasificiranih sekvenci,
- identifikacijsku oznaku sekvence povezanu s njezinom taksonomskom identifikacijskom oznakom,
- broj sekvenci za svaku određenu bakteriju.

Nakon toga, svaka se identifikacijska oznaka sekvence iz izlazne datoteke simulatora, gdje je generiran popis identifikacijskih oznaka sekvence povezanih s njihovim pripadnim taksonomskim oznakama, uspoređuje s popisom

identifikacijskih oznaka sekvenci dobivenih čitanjem generiranih Kraken i CLARK datoteka. Ako kod te provjere uočimo i poklapanje taksonomskih oznaka pridijeljenih tim sekvencama te ako se radi o sekvenci koja je u generiranim datotekama Kraken-a i CLARK-a označena kao klasificirana povećavamo broj točno klasificiranih sekvenci za taj alat.

Nakon gore navedenih prikupljenih podataka provedena je statistička obrada kako bi se postigla krajnja evaluacija alata. Statistička obrada ispisivala se u izlaznoj datoteci evaluatora odvojeno za Kraken te odvojeno za CLARK.

Izlazna datoteka generira se u obliku:

1. redak: ukupan broj sekvenci u testnom skupu podataka

2. redak: broj klasificiranih sekvenci u izlaznoj datoteci alata

3. redak: broj neklasificiranih sekvenci:

$$broj\ očitanih\ sekvenci\ u\ izlaznoj\ datoteci\ alata - broj\ klasificiranih\ sekvenci$$

4. redak: broj točno klasificiranih sekvenci

5. redak: broj netočno klasificiranih sekvenci:

$$broj\ klasificiranih\ sekvenci - broj\ točno\ klasificiranih\ sekvenci$$

6. redak: $\frac{broj\ klasificiranih\ sekvenci}{broj\ sekvenci\ u\ testnom\ skupu\ podataka}$ %

7. redak: $\frac{broj\ točno\ klasificiranih\ sekvenci}{broj\ sekvenci\ u\ testnom\ skupu\ podataka}$ %

8. redak: $\frac{broj\ netočno\ klasificiranih\ sekvenci}{broj\ sekvenci\ u\ testnom\ skupu\ podataka}$ %

9. redak: $\frac{broj\ točno\ klasificiranih\ sekvenci}{broj\ klasificiranih\ sekvenci}$ %

10. redak: $\frac{broj\ netočno\ klasificiranih\ sekvenci}{broj\ klasificiranih\ sekvenci}$ %

Za retke 12., 14. i 16. koristi se sljedeća formula:

$$greška(y) = \frac{|x - broj\ sekvenci\ bakterije\ y\ u\ testnom\ skupu\ podataka|}{broj\ sekvenci\ bakterije\ y\ u\ testnom\ skupu\ podataka}$$

11. redak: broj sekvenci bakterije Pseudomonas koreensis u izlaznoj datoteci alata (x)

12. redak: $greška(\text{Pseudomonas koreensis})$

13. redak: broj sekvenci bakterije Pantonea agglomerans u izlaznoj datoteci alata (x)

14. redak: $greška(\text{Pantonea agglomerans})$

15. redak: broj sekvenci bakterije Klebsiella pneumoniae u izlaznoj datoteci alata (x)

16. redak: $greška(\text{Klebsiella pneumoniae})$

6.2. Analiza rezultata

Pokretanjem evaluatora s gore navedenim argumentima za testni skup podataka sa sekvencama dobivenim drugom generacijom sekvenciranja dobiveni su sljedeći rezultati:

Rb.	Mjere	Kraken	CLARK
1.	**Ukupan broj sekvenci u testnom skupu**	141093	141093
2.	**Broj klasificiranih sekvenci u izlaznoj datoteci**	133567	88593
3.	**Broj neklasificiranih sekvenci u izlaznoj datoteci**	7526	52500
4.	**Broj točno klasificiranih**	91117	87823
5.	**Broj netočno klasificiranih**	42450	770
6.	**Postotak klasificiranih u testnom skupu**	94.67 %	62.79 %
7.	**Postotak točno klasificiranih u testnom skupu**	64.58 %	62.24 %
8.	**Postotak netočno klasificiranih u testnom skupu**	30.09 %	0.55 %
9.	**Postotak točno klasificiranih u skupu klasificiranih**	68.22 %	99.13 %
10.	**Postotak netočno klasificiranih u skupu klasificiranih**	31.78 %	0.87 %
11.	**Broj sekvenci bakterije Pseudomonas koreensis u izlaznoj datoteci alata**	48985	48208
12.	**Pogreška Pseudomonas koreensis**	0.1701	0.1833
13.	**Broj sekvenci bakterije Pantonea agglomerans u izlaznoj datoteci alata**	8971	6504
14.	**Pogreška Pantonea agglomerans**	0.7859	0.8448
15.	**Broj sekvenci bakterije Klebsiella pneumoniae u izlaznoj datoteci alata**	33161	33112
16.	**Pogreška Klebsiella pneumoniae**	0.0336	0.0363

Tablica 6.2.1. Evaluirani rezultati za testni skup podataka sastavljen od kratkih očitanja

	Pantona agglomerans	Klebsiella Pneumoniae	Pseudomonas koreensis	Kromosom	Ostalo
Pantona agglomerans	**33161**	0	0	0	718
Klebsiella Pneumoniae	0	**8971**	0	0	32848
Pseudomonas koreensis	0	0	**48985**	0	8255
Kromosom	0	0	0	**0**	0
Ostalo	0	0	0	0	**0**

Tablica 6.2.2. Odnos klasificiranih sekvenci za Kraken – kratka očitanja

	Pantona agglomerans	Klebsiella Pneumoniae	Pseudomonas koreensis	Kromosom	Ostalo
Pantona agglomerans	**33112**	4	1	0	122
Klebsiella Pneumoniae	1	**6504**	0	0	137
Pseudomonas koreensis	0	0	**48207**	0	173
Kromosom	0	0	0	**0**	0
Ostalo	0	0	0	0	**0**

Tablica 6.2.3. Odnos klasificiranih sekvenci za CLARK – kratka očitanja

Za testni skup podataka sa sekvencama dobivenim sekvenciranjem treće generacije dobiveni su sljedeći rezultati evaluacije:

Rb.	Mjere	Kraken	CLARK
1.	**Ukupan broj sekvenci u testnom skupu**	105794	105794
2.	**Broj klasificiranih sekvenci u izlaznoj datoteci**	72362	63204
3.	**Broj neklasificiranih sekvenci u izlaznoj datoteci**	33432	42590
4.	**Broj točno klasificiranih**	57934	58376
5.	**Broj netočno klasificiranih**	14428	4828
6.	**Postotak klasificiranih u testnom skupu**	68.40 %	59.74 %
7.	**Postotak točno klasificiranih u testnom skupu**	54.76 %	55.18 %
8.	**Postotak netočno klasificiranih u testnom skupu**	13.64 %	4.56 %
9.	**Postotak točno klasificiranih u skupu klasificiranih**	80.06 %	92.36 %
10.	**Postotak netočno klasificiranih u skupu klasificiranih**	19.94 %	7.64 %
11.	**Broj sekvenci bakterije Pseudomonas koreensis u izlaznoj datoteci alata**	20462	20360
12.	**Pogreška Pseudomonas koreensis**	0.2122	0.2161
13.	**Broj sekvenci bakterije Pantonea agglomerans u izlaznoj datoteci alata**	9978	10571
14.	**Pogreška Pantonea agglomerans**	0.5145	0.4856
15.	**Broj sekvenci bakterije Klebsiella pneumoniae u izlaznoj datoteci alata**	27494	27445
16.	**Pogreška Klebsiella pneumoniae**	0.0190	0.0212

Tablica 6.2.4. Evaluirani rezultati za testni skup podataka sastavljen od dugih očitanja

	Pantona agglomerans	Klebsiella Pneumoniae	Pseudomonas koreensis	Kromosom	Ostalo
Pantona agglomerans	**27494**	0	0	0	251
Klebsiella Pneumoniae	0	**9978**	0	0	8866
Pseudomonas koreensis	0	0	**20462**	0	2153
Kromosom	0	0	0	**0**	0
Ostalo	0	0	0	0	**0**

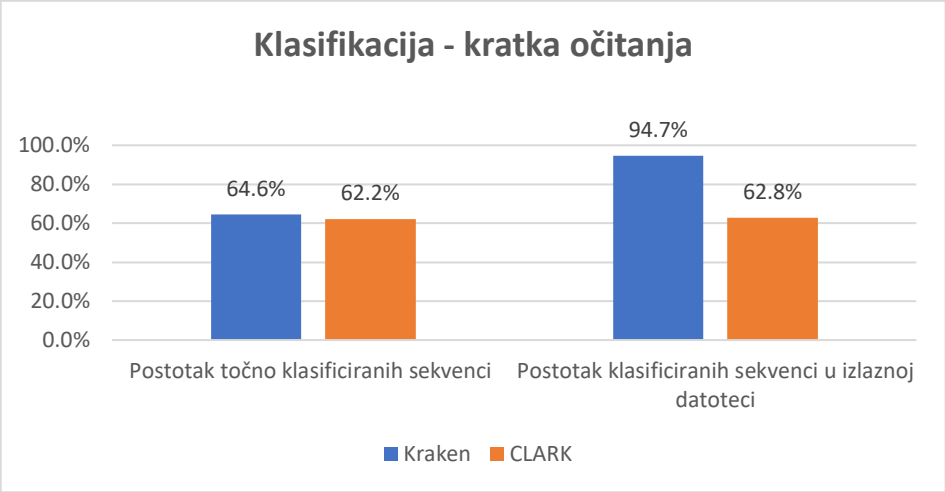
Tablica 6.2.5. Odnos klasificiranih sekvenci za Kraken – duga očitanja

	Pantona agglomerans	Klebsiella Pneumoniae	Pseudomonas koreensis	Kromosom	Ostalo
Pantona agglomerans	**27445**	5	0	0	107
Klebsiella Pneumoniae	3	**10571**	0	0	2181
Pseudomonas koreensis	2	0	**20360**	0	891
Kromosom	0	0	0	**0**	0
Ostalo	0	0	0	0	**0**

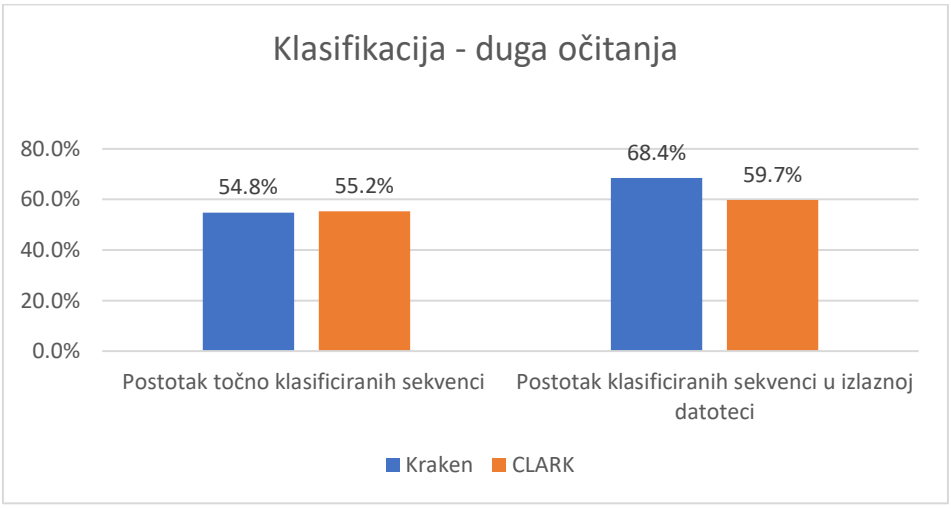
Tablica 6.2.6. Odnos klasificiranih sekvenci za CLARK – duga očitanja

Detaljnije analizirajući dobivene podatke prvu stvar koju možemo učiti je kako Kraken u slučaju klasifikacije skupa podataka kratkih očitanja (**Tablica 6.2.1.**) ima velik postotak klasificiranih sekvenci. U slučaju dugih očitanja (**Tablica 6.2.4.**), taj postotak znatno je manji, točnije manji je za nešto više od 26%. CLARK pak s druge strane u slučaju kratkih očitanja ima dosta nizak postotak klasificiranih sekvenci. Pogledamo li rezultate dobivene klasifikacijom dugih očitanja, možemo uočiti da je postotak klasificiranih sekvenci kod CLARK-a podjednak onom kod klasifikacije kratkih očitanja.

Nastavljajući analizu, nailazimo na sljedeći podatak koji je od velike važnosti pri ocjeni samih alata. Riječ je o broju točno klasificiranih sekvenci. Iako smo kod Kraken-a u slučaju kratkih očitanja uočili visok postotak klasificiranih sekvenci, lako je za uvidjeti da je postotak točno klasificiranih sekvenci znatno manji. U cijelom skupu klasificiranih sekvenci, samo 64.58% je točno klasificiranih. Njihov raspored možemo vidjeti u tablici (**Tablica 6.2.2.**). Jedan redak u tablici predstavlja jednu od vrsta ulaznih sekvenci. Na dijagonali tablice vidimo podatke za točno klasificirane sekvence, a ostali podatci pokazuju oblik klasifikacije krivo klasificiranih sekvenci. Može se uočiti da niti jedna krivo klasificirana sekvenca nije klasificirana na bakterije koje se nalaze u našem skupu. Kod dugih očitanja, situacija je malo drugačija. Postotak klasificiranih jest znatno manji, no udio točno klasificiranih sekvenci u skupu klasificiranih sekvenci iznosi 80.06%. Odnos klasificiranih sekvenci za duga očitanja također je vidljiv u tablici (**Tablica 6.2.5.**). Kod CLARK-a nailazimo na skroz drugačiju situaciju. Iako smo prethodno uvidjeli da je postotak klasificiranih sekvenci kod CLARK-a znatno manji nego kod Kraken-a, ovdje je udio točno klasificiranih sekvenci u skupu klasificiranih sekvenci izrazito velik. Za kratka očitanja postotak točno klasificiranih sekvenci kod CLARKA iznosi čak 99.13%, dok je za duga nešto niži, a iznosi 92.36%. Iz tablice (**Tablica 6.2.3.**) vidimo da je CLARK, za slučaj kratkih očitanja, četiri sekvence koje su u izlaznoj datoteci simulatora očitane kao sekvence Pantonae agglomerans klasificirao kao sekvence Klebsiella pneumoniae. U tablici (**Tablica 6.2.6.**) vidimo raspored klasificiranih sekvenci za duga očitanja. Ovi podatci prikazani su i grafički s ciljem bolje predodžbe ove mjere (**Graf 6.2.1.** i **Graf 6.2.2.**).



Graf 6.2.1. Usporedba točno klasificiranih i ukupno klasificiranih sekvenci za kratka očitanja



Graf 6.2.2. Usporedba točno klasificiranih i ukupno klasificiranih sekvenci za duga očitanja

Promatrajući rezultate ostalih navedenih mjera, možemo zamijetiti kako u velikoj većini rezultata kod kratkih očitanja Kraken ima pozitivnije rezultate nego CLARK. Npr. ako promotrimo pogreške za svaku od bakterija vidimo da razlike u pogreškama Kraken-a i CLARK-a nisu pretjerano velike, no u svim pogreškama CLARK ipak prednjači. Kod dugih očitanja te razlike su još i manje, a u slučaju bakterije Pantonea agglomerans vidimo kako CLARK ima manju pogrešku nego Kraken.

Općenito promatrajući sve dobivene rezultate zajedno, možemo uvidjeti kako Kraken ima znatno bolje rezultate kod klasifikacije testnog skupa stvorenog od kratkih očitanja, dok CLARK u oba slučaja generira podjednake rezultate.

7. ZAKLJUČAK

Provodeći ovakav način evaluacije rada alata za identifikaciju vrsta u metagenomskim uzorcima možemo doći do sljedećeg zaključka. Naime, rezultati su pokazali kako Kraken daje dosta veći broj klasificiranih očitanja te omogućuje veću razinu identifikacije. Međutim, taj veći broj klasifikacija ima svoju cijenu, a to je da se upravo na takav način povećava i broj krivo klasificiranih sekvenci koje nas mogu dovesti do krajnje krivih zaključaka, a već prije je navedeno kako je identifikacija jedan od ključnih koraka za medicinu i prehranu, točnije za čovjekovo zdravlje. Upravo zato s druge strane CLARK nudi manji broj klasificiranih sekvenci, no njegova točnost klasifikacije ipak je na visokom nivou.

Preostaje nam još klasifikacijski alat MetaPhlAn2 čija funkcionalnost ipak nije do kraja istražena. Naime, rezultati klasifikacija bili su niske značajnosti. Jedan od razloga možda je i pogrešan način korištenja samog alata ili pak oblik sekvenci koje su se nalazile u testnom skupu podataka. Provjera tih pretpostavki ostavljena je za daljnje istraživanje.

Pri donošenju odluke koji program odabrati pri identifikaciji vrsta u metagenomskim uzorcima definitivno treba razmisliti koja nam je ciljna svrha, veći broj klasificiranih sekvenci ili veća točnost pri klasifikaciji.

8. LITERATURA

[1]. Thomas T, Gilbert J, Meyer F. Metagenomics - **A guide from sampling to data analysis**. *Microb Inform Exp*. 2012;2(1):3. 9. 2. 2012. doi:10.1186/2042-5783-2-3

[2]. Kunin V., Copeland A., Lapidus A., Mavromatis K., Hugenholtz P**. A Bioinformatician's Guide to Metagenomics**. *American Society for Microbiology Journals*. 3. 12. 2008. https://mmbr.asm.org/content/72/4/557 Datum pristupa: 5.6.2019.

[3]. Wood D. E., Salzberg S. L. **Kraken: ultrafast metagenomic sequence classification using exact alignments**. *Genome Biology.* 15. 3. 2014. 15:R46

[4]. The National Center for Biotechnology Information: https://www.ncbi.nlm.nih.gov. Datum pristupanja: 11.6.2019.

[5]. JELLYFISH: https://www.cbcb.umd.edu/software/jellyfish Datum pristupanja: 11.6.2019.

[6]. Kraken Manual. https://ccb.jhu.edu/software/kraken/MANUAL.html Datum pristupa: 5.6..2019.

[7]. Segata N., Waldron L., Ballarini A., Narasimhan V., Jousson O., Huttenhower C. **Metagenomic microbial community profiling using unique clade-specific marker genes**. *Nature Methods.* 2012. 811 EP -

[8]. MetaPhlAn: Metagenomic Phylogenetic Analysis. http://huttenhower.sph.harvard.edu/metaphlan. Datum pristupa: 5.6.2029.

[9]. Truong, D.T. i Segata, N. **MetaphlAn2 – Metagenomic Phylogenetic Analysis**. https://bitbucket.org/biobakery/metaphlan2/src/default/ . Datum pristupa: 5.6.2019.

[10]. Bowtie2: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml . Datum pristupa: 11.6.2019.

[11]. Truong D. T., Franzosa E. A., Tickle T. L. Scholz M., Weingart G., Pasolli E., Tett A. Huttenhower C., Segata N. **MetaPhlAn2 for enhanced metagenomic taxonomic profiling.** *Nature America.* 2015. 902:903

[12]. Ounit R., Wanamaker S., Close T. J., Lonardi S. **CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative *k*-mers,** *BMC Genomics* 25.3.2015 **16**:236

[13]. CLARK. http://clark.cs.ucr.edu Datum pristupa: 5.6.2019.

[14]. Biopython: https://biopython.org . Datum pristupa: 5.6.2019.

[15]. Humphrey J., Seitz T., Haan T., Ducluzeau A., Drown D. M. **Complete Genome Sequence of *Pantoea agglomerans* TH81, Isolated from a Permafrost Thaw Gradient**, *American Society for Microbiology Journals,* 3.1.2019. e01486-18

[16]. Schmid M., Frei D., Patrignani A., Schlapbach R., Frey J. E., Remus-Emsermann M. N. P., Ahrens C. H. **Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats,** *Nucleic Acids Research*, Volume 46, 28. 9. 2018, 8953–8965

[17]. Wick R. R., Judd L. M., Gorrie C. L., Holt K. E. **Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads,** *PLoS Comput Biol,* 8.6.2017., e1005595

9. DODATCI

Ocjena alata za identifikaciju vrsta u metagenomskim uzorcima

Sažetak

Metagenomika jest relativno novo područje bioinformatike koje se bavi proučavanjem uzoraka preuzetih direktno iz prirode. Veliku primjenu ima u analizi i održavanju ljudskog zdravlja. U toj primjeni svoju ulogu igra sama identifikacija vrsta u metagenomskim uzorcima za koju se koriste alati kao što su Kraken, MetaPhlAn2 te CLARK. Za ispitivanje njihove funkcionalnosti osmišljena su dva različita testna skupa podataka, skup kratkih i skup dugih očitanja. Nakon simulacije tih testnih skupova, provela se klasifikacija uzoraka. Kraken se pokazao kao alat koji klasificira velik broj sekvenci, no točnost te klasifikacije mu je bila lošija. CLARK je klasificirao manji broj sekvenci, no točnost te klasifikacije bila je na visokoj razini. Kod alata MetaPhlAn2 klasificiralo se neznatno malo sekvenci, a daljnja evaluacija za taj alat je izostavljena.

Ključne riječi: metagenomika, Kraken, CLARK, MetaPhlAn2, ocjena alata

A Benchmark of Tools for Metagenomic Species Identification

Abstract

Metagenomics is a relatively new area of bioinformatics that deals with the study of samples taken directly from nature. It has great application in analyzing and maintaining human health. Here identification of metagenomic species has great role using tools like Kraken, MetaPhlAn2 and CLARK. To test their functionality, two different sets of data were created, a set of short and long readings. After the simulation of these test sets, the sample classification was performed. Kraken proved to be a tool that classifies a large number of sequences, but the accuracy of that classification gave bad results. CLARK has classified a lower number of sequences, but the accuracy of this classification was high. MetaPhlAn2 classified little number of sequences, and further evaluation of that tool was omitted.

Keywords: metagenomics, Kraken, CLARK, MetaPhlAn2, benchmark