SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

PROJEKT IZ BIOINFORMATIKE 1

Određivanje sastava metagenomskog uzorka

Roko Barišić, Domagoj Marić

Voditelj: *doc. dr. sc. Krešimir Križanović*

Zagreb, lipanj 2025.

Sadržaj

1. Uvod

Metagenomika je interdisciplinarno područje koje spaja molekularnu biologiju, bioinformatiku i ekologiju s ciljem proučavanja mikrobioloških zajednica izravno iz okolišnih uzoraka, bez potrebe za izolacijom i uzgojem pojedinačnih organizama. Takav pristup omogućuje analizu ukupnog genetskog materijala (DNK) svih mikroorganizama prisutnih u određenom uzorku, čime se zaobilaze ograničenja klasičnih mikrobioloških metoda koje se temelje na kultivaciji. Metagenomske analize danas imaju ključnu ulogu u brojnim znanstvenim i primijenjenim područjima – od medicine i ekologije do industrije i poljoprivrede.

Metagenomski uzorak predstavlja DNK izoliranu iz mješovitog mikrobiološkog okoliša, poput tla, morske vode ili biološkog materijala iz nekog organizma. U takvom uzorku nalaze se fragmenti genoma brojnih organizama – bakterija, arheja, virusa i eukariotskih mikroba – što analizu čini izazovnom, ali i iznimno informativnom. Analizom metagenomskih uzoraka moguće je odrediti sastav mikrobne zajednice, procijeniti njezinu biološku raznolikost i funkcionalni potencijal te pratiti promjene koje nastaju uslijed utjecaja iz okoliša, bolesti ili ljudske aktivnosti.

Jedan od ključnih zadataka u analizi metagenomskih podataka je određivanje sastava uzorka, odnosno identifikacija i kvantifikacija organizama prisutnih u uzorku. U tu se svrhu koriste različite računalno učinkovite metode - poput analize distribucije kmera.

Kmeri su kratki uzastopni nizovi DNK duljine *k* koji se izlučuju iz većih sekvenci. Svaki mikroorganizam ima karakterističan obrazac distribucije kmera koji se može iskoristiti kao svojevrsni „genomski otisak prsta“ za taksonomsku klasifikaciju. Analiza distribucije kmera omogućuje brzu procjenu sastava metagenomskih uzoraka, osobito kada su sekvence fragmentirane ili kada nedostaje kvalitetna referentna baza. Zbog toga se ova metoda sve češće koristi u velikim metagenomskim projektima koji uključuju milijune sekvenci.

Metagenomika ima široku primjenu u različitim znanstvenim i primijenjenim područjima. U medicini se koristi za analizu ljudskog mikrobioma i identifikaciju patogena iz kliničkih uzoraka bez potrebe za uzgojem mikroorganizama. U ekologiji omogućuje proučavanje biološke raznolikosti u okolišu te praćenje utjecaja zagađenja i klimatskih promjena. U poljoprivredi i stočarstvu pomaže u unaprjeđenju zdravlja tla i životinja analizom mikrobnih zajednica. U industriji i biotehnologiji metagenomski podaci koriste se za otkrivanje korisnih gena i enzima s primjenom u proizvodnji, preradi otpada i razvoju lijekova. Također, važna je i za javno zdravstvo, gdje se primjenjuje za praćenje antimikrobne rezistencije i biološke sigurnosti hrane i vode.

U ovom je radu prikazana metoda određivanja sastava metagenomskog uzorka korištenjem analize distribucije kmera. Opisani su korišteni algoritmi i alati, obrada podataka, način generiranja i interpretacije distribucija kmera, te evaluacija rezultata.

2. Konstrukcija distribucijskih vektora

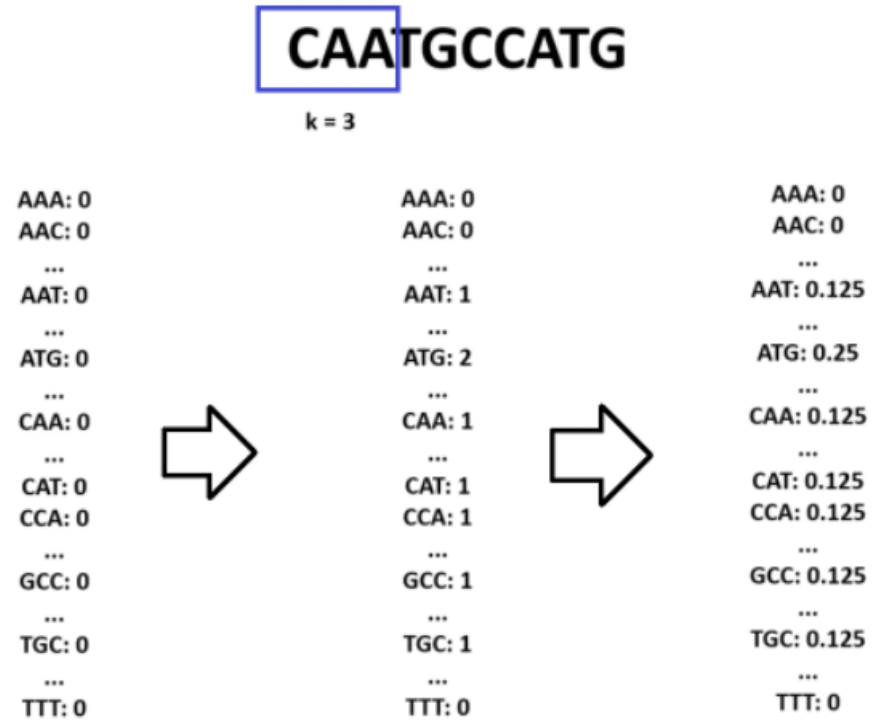
Konstrukcija distribucijskih vektora temelji se na analizi učestalosti kmera, odnosno svih mogućih podnizova određene duljine *k* unutar sekvenci dobivenih iz metagenomskog uzorka. Sekvence, a time i kmeri, sastoje se od slova A, C, G i T. Mogu sadržavati i neka druga, ali njih preskačemo. Svaki mikroorganizam ima jedinstven profil kmera koji odražava specifičnosti njegova genoma, što omogućuje da se ti profili iskoriste kao taksonomski prepoznatljivi obrasci.

Prvi korak u izradi distribucijskog vektora podrazumijeva odabir vrijednosti *k*. Nakon odabira vrijednosti, iz svih dostupnih sekvenci iz uzorka izdvajaju se svi kmeri duljine *k*, pri čemu se prozor pomiče po sekvenci jedno po jedno slovo, čime se dobiva potpuni skup kmera koji se pojavljuju u uzorku. Svaki kmer broji se i bilježi njegova učestalost, čime se formira vektor koji odražava distribuciju kmera u cijelom uzorku.

Dobivena distribucija zatim se normalizira, s obzirom na ukupan broj kmera, kako bi se podaci mogli uspoređivati među uzorcima različite veličine. Tako formirani distribucijski vektor predstavlja kvantitativni prikaz sadržaja uzorka i može se koristiti za usporedbu s referentnim vektorima poznatih organizama. Za tu usporedbu koristimo kosinusnu sličnost.

Tako je moguće identificirati organizme čiji vektori distribucije najviše nalikuju onima iz analiziranog uzorka, što omogućuje procjenu taksonomskog sastava. Ovakav pristup, temeljen na frekvencijama kmera, posebno je pogodan za analizu velikih količina podataka jer izbjegava zahtjevna poravnanja sekvenci i omogućuje brzu i skalabilnu obradu metagenomskih uzoraka.

Objasnimo ovaj postupak na primjeru sa slike 2.1.



Slika 2.1. Konstrukcija distribucijskog vektora

Za vrijednost *k*, odnosno duljinu kmera, uzeli smo 3, a niz koji želimo analizirati je CAATGCCATG. Na početku se distribucijski vektor sastoji od 4^k, odnosno 64 para ključ:vrijednost, pri čemu je ključ određeni kmer (od AAA do TTT), a vrijednost za sve parove 0. Kao prvi kmer uzimamo prva tri slova niza, CAA, i povećavamo vrijednost para s ključem CAA za 1. Potom pomičemo prozor jedno mjesto udesno i povećavamo vrijednost para s ključem AAT. Kada tako prođemo cijeli niz, dobit ćemo distribucijski vektor gdje će vrijednost uz svaki ključ biti broj pojavljivanja tog ključa u početnom nizu. Završni je korak normalizirati sve vrijednosti u vektoru tako da ih podijelimo s ukupnim brojem kmera u nizu.

3. Kosinusna sličnost

Kosinusna sličnost je mjera koja se koristi za kvantitativnu procjenu sličnosti između dvaju vektora u višedimenzionalnom prostoru. Temelji se na kutu između vektora, a ne na njihovoj veličini. Vrijednost kosinusne sličnosti kreće se od 0 do 1, gdje 1 označava identične smjerove vektora (potpunu sličnost), a 0 označava međusobnu ortogonalnost (potpunu različitost).

Matematički, kosinusna sličnost između dvaju vektora **A** i **B,** s kutom θ između njih, definira se izrazom:

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

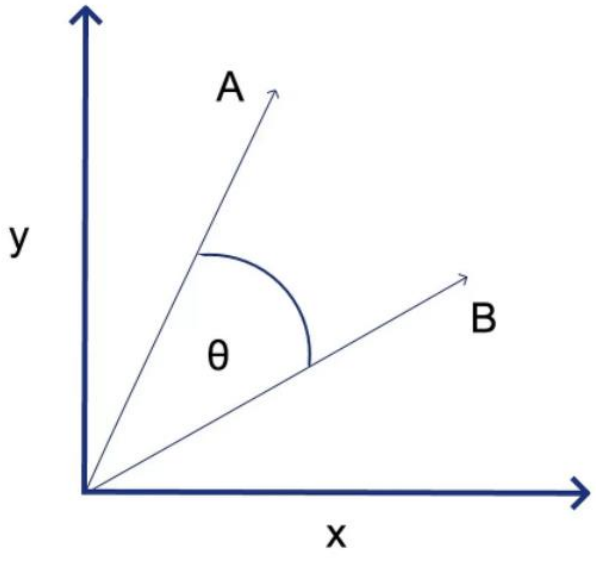
Pritom je u brojniku skalarni umnožak vektora **A** i **B**, a u nazivniku umnožak euklidskih normi tih vektora. To sve računamo ovako:

$$A \cdot B = \sum_{i=1}^{n} a_i \cdot b_i \qquad \|A\| = \sqrt{\sum_{i=1}^{n} a_i^2} \qquad \|B\| = \sqrt{\sum_{i=1}^{n} b_i^2}$$

Pod *n* mislimo na duljinu vektora - u našem je slučaju to 4^k, ako je *k* duljina kmera.

Primjenom kosinusne sličnosti moguće je kvantificirati koliko su distribucije iz uzorka i iz baze referentnih genoma međusobno slične.

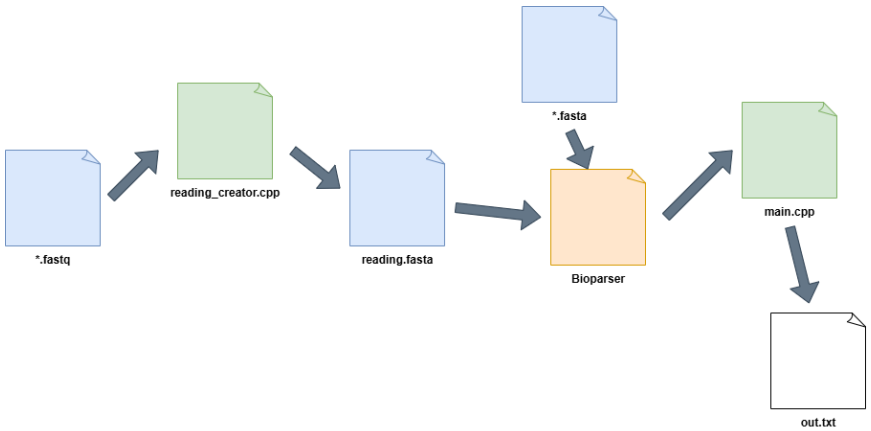
U analizi metagenomskih podataka, ovaj pristup omogućuje identifikaciju organizama čiji kmer-profil najviše odgovara onome pronađenom u uzorku. Budući da se kosinusna sličnost temelji na relativnoj raspodjeli vrijednosti unutar vektora, ona je otporna na razlike u veličini uzoraka i koristi se za klasifikaciju čak i u slučajevima kad je količina podataka neujednačena.



Slika 3.1. Kosinusna sličnost

4. Programska implementacija

Programska implementacija ovoga projekta dostupna je na GitHubu: https://github.com/rokobarisic/odredivanje-sastava-metagenomskog-uzorka-2024-25/. Na slici 4.1. prikazana je struktura implementacije.



Slika 4.1. Struktura programske implementacije

Datoteka *reading_creator.cpp* kao ulaz prima očitanja, odnosno datoteke s nastavkom .fastq, i iz svakog bira određeni broj sekvenci koje kombinira u datoteku reading.fasta. Ta datoteka, zajedno s datotekama referentnih genoma, s nastavkom .fasta, ide u *main.cpp* – taj kod koristi biblioteku Bioparser, dostupnu na https://github.com/rvaser/bioparser, za brzo i efikasno učitavanje sekvenci. Potom *main.cpp* odrađuje glavnu logiku projekta, analizira koliko je očitanja pronađeno za svaki referentni genom i izvještaj zapisuje u datoteku *out.txt*.

4.1. Generiranje datoteke *reading.fasta*

Datoteke s očitanjima imaju nastavak .fastq i sastoje se od „četvorki“ – svaka sekvenca unutar datoteke zapisana je tako da je u prvom retku naziv očitanja, u drugom sama sekvenca, u trećem znak +, i u četvrtom niz znakova koji označava kvalitetu očitanja. Pritom je važno naglasiti da je prvi znak u prvom retku znak @.

Datoteka *reading_creator.cpp* otvara sve datoteke koje se nalaze u mapi *Data/Readings/* i iz svake uzima nasumičan broj sekvenci iz intervala, koji je lako moguće promijeniti u kodu zbog parametrizacije. Za svaku od tih sekvenci uzima prva dva retka, dakle naziv očitanja i samu sekvencu, postavlja prvi znak prvog retka u > i zapisuje ta dva retka u datoteku *reading.fasta*.

4.2. Bioparser

Bioparser je biblioteka namijenjena brzom i učinkovitom parsiranju bioloških sekvencijskih datoteka, poput FASTA i FASTQ formata. Implementirana je u jeziku C++ i često se koristi u bioinformatičkim alatima gdje je važna obrada velikih količina sekvencijskih podataka uz minimalno vrijeme i memorijsku potrošnju. Bioparser omogućuje čitanje i iteraciju kroz zapise bez potrebe za učitavanjem cijele datoteke u

memoriju, što ga čini pogodnim za visokoprotočne metagenomske analize. Zahvaljujući svojoj modularnosti i kompatibilnosti s modernim C++ alatima, često se koristi kao dio većih bioinformatičkih sustava za obradu sirovih podataka iz sekvenciranja. Mi ju koristimo kako bismo efikasno učitali datoteke s referentnim genomima, ali i datoteku s očitanjima koju smo prethodno stvorili.

4.3. Glavna logika programa

Na početku parsiramo datoteke s referentnim genomima. Za svaku od učitanih sekvenci moramo kreirati distribucijski vektor prije objašnjenim postupkom. Prilikom kreiranja distribucijskog vektora, nećemo zapisivati ključ, odnosno kmer, kao skup slova („AACGT“), nego ta slova kodiramo radi uštede memorije i zapisujemo to brojčano, s dva bita za svako slovo („00123“, odnosno „0000011011“). Kada smo kreirali distribucijski vektor referentnog genoma, zapisujemo ga u rječnik distribucijskih vektora referentnih genoma.

Nakon kreiranja distribucijskih vektora svih referentnih genoma, parsiramo datoteku *reading.fasta*. Za svako očitanje u toj datoteci kreiramo njegov distribucijski vektor. Potom računamo kosinusne sličnosti distribucijskog vektora tog očitanja i distribucijskih vektora svih referentnih genoma. Pronalazimo referentni genom za koji je ta vrijednost najveća i zaključujemo da očitanje koje analiziramo pripada tom referentnom genomu.

Kada smo analizirali sva očitanja, izvještaj zapisujemo u datoteku *out.txt*.

4.4. Evaluacija dobivenih podataka

Za evaluaciju rezultata dobivenih klasifikacijom temeljenom na distribuciji kmera korišten je alat Minimap2, koji omogućuje brzo poravnanje očitanja na referentne genomske sekvence.

Potrebno je spojiti sve referentne genome u jednu datoteku: pozicioniramo se u References mapu i koristimo „*cat *.fasta > references.fasta*“. Zatim u istu mapu stavimo datoteku *reading.fasta* i *references.fasta* i pozicioniramo se u nju. Pomoću naredbe „*minimap2 -ax map-pb -t 8 references.fasta reading.fasta > izlaz.sam*“ generiramo datoteku *izlaz.sam* koja će sadržavati podatke koji nas zanimaju.

Ako koristimo „*grep -v "^@" izlaz.sam | awk '$3 != "*" && and($2, 4) == 0 {count[$3]++} END {for (r in count) print r ": " count[r]}'*“, dobit ćemo podatke o broju očitanja po referentnom genomu.

Ova obrada omogućila je kvantitativnu analizu zastupljenosti referentnih organizama u uzorku na temelju poravnanja, čime se mogla usporediti stvarna distribucija očitanja s rezultatima dobivenim klasifikacijom putem vektora kmera. Time je osigurana pouzdana evaluacija učinkovitosti primijenjene metode.

Nakon evaluacije otkrivamo da naša metoda klasifikacije temeljena na distribuciji kmera nije savršena, postoje neslaganja između izlaza dobivenoga našim programom i izlaza koji je generirao Minimap2. Pritom trebamo uzeti u obzir da Minimap2 neka očitanja (čak i do dvadesetak posto) neće klasificirati.

5. Analiza rezultata na realnom primjeru

Kako bismo proučili točnost ovoga programskog rješenja, poslužit ćemo se primjerom. Primjer koristi deset bakterija, a za svaku od njih uzeli smo datoteku s referentnim genomom i jednu datoteku s očitanjima.

Kada smo sve datoteke stavili u potrebne direktorije i pokrenuli *reading_creator.cpp*, nastala je datoteka *reading.fasta* koja sadrži 243071 očitanje. Za svaku od 10 datoteka nasumično je odabrano između 10000 i 100000 očitanja, a naš je cilj vidjeti koliko je precizan *main.cpp* program u otkrivanju broja očitanja po referentnom genomu. Pokrenut ćemo program koristeći različite duljine kmera, a potom ćemo to usporediti s izlazom iz Minimapa te točnim rezultatima.

#	bakterija	reference	broj očitanja po našem programu (k=5)	broj očitanja po našem programu (k=8)	broj očitanja po Minimapu	broj očitanja zapravo
1	Escherichia coli	NC_000913.3	13987	26119	26943	23954
2	Salmonella enterica	NC_003198.1, NC_003384.1, NC_003385.1	43997	30211	34239	34491
3	Helicobacter pylori	NC_000915.1	25063	24404	27084	27092
4	Legionella pneumophila	NC_002942.5	30235	43340	26441	26177
5	Haemophilus influenzae	NC_000907.1	23188	20599	25902	27921
6	Dermacoccus nishinomiyaensis	NZ_CP008889.1, NZ_CP008890.1	19829	15379	15479	15557
7	Citrobacter amalonaticus	NZ_CP011132.1, NZ_CP011133.1	23379	19740	16635	23305
8	Pseudomonas fluorescens	NC_016830.1	19230	22326	5679	21874
9	Proteus vulgaris	NZ_KN150745.1, NZ_KN150746.1	25153	23404	27376	28753
10	Mycoplasma synoviae	NZ_CP011096.1	18858	17444	15201	13947

Slika 5.1. Usporedba rezultata

Razjasnimo prvo zašto neka bakterija u ćeliji s referencama (odnosno ID-jevima) ima više zapisa. Naime, određena datoteka s referentnim genomom može sadržavati, osim tog genoma, i genome plazmida. Tako na primjer datoteka s referentnim genomom bakterije Salmonella enterica sadrži i genome dvaju plazmida. I naš program i Minimap2 pridruživat će određena očitanja genomima plazmida, a radi jednostavnije procjene točnosti pribrojit ćemo ta očitanja odgovarajućim bakterijama.

Vidljivo je da ni u kojem slučaju točnost nije savršena. Što se tiče našeg programa, možemo vidjeti da je točnost ipak nešto veća kada koristimo veću duljinu kmera. Tome u prilog ide i činjenica da su rezultati Minimapa daleko najtočniji jer su korišteni kmeri duljine 19 znakova. Kod Minimapa vidimo samo jedno veliko odstupanje – kod bakterije Pseudomonas fluorescens. Uzrok tome vjerojatno leži u činjenici da Minimap2 neka očitanja ne pridruži nijednoj referenci ako ih ne možemo dovoljno sigurno pridružiti.

Dakle, veći *k* => veća točnost.

6. Zaključak

U ovom je radu prikazana metoda određivanja sastava metagenomskog uzorka korištenjem distribucije kmera i usporedbe distribucijskih vektora pomoću kosinusne sličnosti. Korištenjem frekvencijskog pristupa omogućena je efikasna i skalabilna analiza bez potrebe za poravnanjem sekvenci, što je osobito korisno pri obradi velikih količina metagenomskih podataka. Razvijeni programski sustav pokazao se funkcionalnim i praktičnim za određivanje taksonomskog sastava, a njegova primjena demonstrirana je kroz konkretan primjer i evaluaciju pomoću alata Minimap2.

Iako su uočene određene razlike u rezultatima dobivenima našim pristupom i korištenjem alata Minimap2, metodologija se pokazala pouzdanom, a moguća odstupanja djelomično se mogu pripisati ograničenjima samog alata za poravnanje. Ovakav pristup ima velik potencijal za primjenu u budućim bioinformatičkim analizama i projektima u područjima ekologije, medicine, industrije i javnog zdravstva.

7. Sažetak

Rad se bavi analizom metagenomskih uzoraka korištenjem frekvencijske distribucije kmera i kosinusne sličnosti za određivanje taksonomskog sastava. Opisana je konstrukcija distribucijskih vektora, računanje kosinusne sličnosti, implementacija klasifikacijskog algoritma u programskom jeziku C++ i evaluacija rezultata pomoću alata Minimap2. Predložena metoda omogućuje brzo i učinkovito određivanje sastava uzorka bez potrebe za poravnanjem, što je čini prikladnom za analizu velikih količina podataka. Rezultati su pokazali zadovoljavajuću točnost, uz mogućnosti daljnjeg unaprjeđenja i optimizacije za primjenu u znanstvenim i primijenjenim bioinformatičkim istraživanjima.

8. Literatura

[1] D. Sviličić, „Analiza metagenomskog uzorka na temelju distribucije k-mera“. 2024. https://repozitorij.fer.unizg.hr/islandora/object/fer:12751
[2] R. Grbelja, „Usporedba sekvenci pomoću distribucije k-mera“
[3] Bioparser. https://github.com/rvaser/bioparser
[4] Minimap2. https://github.com/lh3/minimap2