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 1](#_Toc159987575)

[2. Seminarski rad 2](#_Toc159987576)

[3. Zaključak 3](#_Toc159987577)

[4. Literatura 4](#_Toc159987578)

[5. Sažetak 5](#_Toc159987579)

# 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 zbirnu 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. Na temelju analize dobiveni su uvidi u strukturu mikrobne zajednice prisutne u konkretnom uzorku. Na kraju su razmotrene prednosti, ograničenja i potencijalne primjene ove metode u budućim metagenomskim istraživanjima.

# Konstrukcija distribucijskih vektora

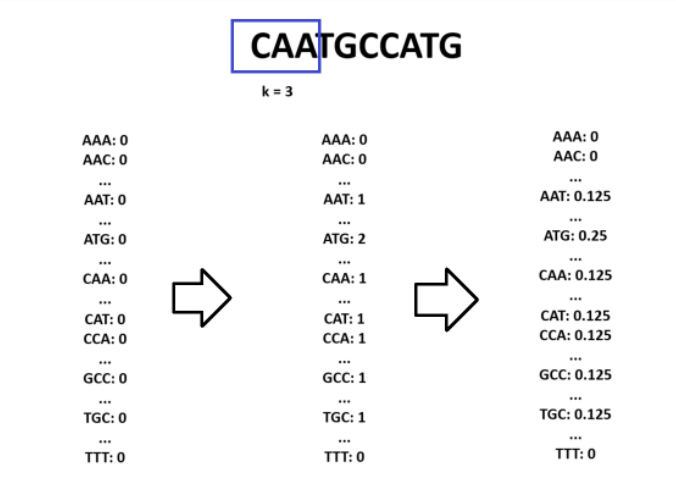
Konstrukcija distribucijskih vektora temelji se na analizi učestalosti kmera, odnosno svih mogućih podnizova određene duljine *k* unutar sekvenci dobivenih iz metagenomskog uzorka. Sekvence, a time i kmeri, sastoje se od slova A, C, G i T. Mogu sadržavati i neka druga, ali njih preskačemo. Svaki mikroorganizam 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.



Za vrijednost *k*, odnosno duljinu kmera, uzeli smo 3, a niz koji želimo analizirati je CAATGCCATG. Na početku se distribucijski vektor sastoji od 4k, 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.

# Kosinusna sličnost

# Zaključak

# Sažetak

# Literatura