SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

BIOINFORMATIKA 1

Brza i visoko učinkovita referencijska kompresija genoma

Karla Budimir i Mateo Jakšić

Voditelj: Mirjana Domazet-Lošo

Zagreb, lipanj, 2025.

**Sadržaj**

[1. Uvod 1](#_Toc199892966)

[2. Referencijska kompresija i dekompresija genoma 2](#_Toc199892967)

[2.1 Kompresija genoma 2](#_Toc199892968)

[2.2 Dekompresija genoma 2](#_Toc199892969)

[2.3 Mapiranje ciljnog genoma na referencijski genom 3](#_Toc199892970)

[2.4 Algoritam referencijske kompresije genoma 4](#_Toc199892971)

[2.5 Algoritam referencijske dekompresije genoma 5](#_Toc199892972)

[3. Skup podataka 6](#_Toc199892973)

[4. Rezultati 7](#_Toc199892974)

[5. Zaključak 8](#_Toc199892975)

[6. Literatura 9](#_Toc199892976)

# Uvod

Razvoj bioinformatike usko je povezan s tehnološkim napretkom računalnih sustava, što se djelomično može pripisati efektu tzv. Mooreovog zakona. Mooreov zakon navodi da će se broj tranzistora u čipu, time i računalna snaga, udvostručavati otprilike svake dvije godine, dok će se cijena po tranzistoru smanjivati. Sličan trend uočen je i u sekvenciranju genoma, gdje se platformama za sekvenciranje visokog protoka i algoritmima za sastavljanje genoma iz fragmenata, omogućio brzi porast broja dostupnih genoma. Glavni razlog tome je značajno smanjenje cijene sekvenciranja genoma, s prvotnih 100.000.000 na manje od 1.000 američkih dolara po genomu.

Tradicionalni algoritmi kompresije ne mogu zadovoljiti zahtjeve za visokom potražnjom kompresije zbog intrinzičnih izazova DNA strukture, poput male veličine abecede, učestalih ponavljanja i palindroma. Veliki iskorak u području kompresije genoma bila je objava prvog referencijskog algoritma kompresije genoma DNAZip (Christley *et al.*, 2009). Referencijski algoritmi kompresije genoma temelje se na pohranjivanju samo razlike između dva slična genoma, čime postižu značajan napredak na području pohrane, prijenosa i kompresije genoma. Motivacija za algoritam je činjenica da ljudski genomi imaju više od 99% sličnosti. U ovom projektu proučavat ćemo implementaciju algoritma brze i visoko učinkovite referencijske kompresije i dekompresije genoma.

# Referencijska kompresija i dekompresija genoma

## Kompresija genoma

Kompresija genoma, također poznata i kao sažimanje genoma, postupak je smanjenja količine podataka potrebnih za reprezentaciju genomske informacije (npr. sekvence DNA) bez gubitka bitnih informacija. Omogućuje učinkovitu pohranu, brži prijenos i lakšu analizu podataka. S obzirom na način očuvanja informacija, razlikujemo dvije vrste kompresije:

1. kompresija bez gubitka (engl. *Lossless compression*)
2. kompresija s gubitcima (engl. *Lossy compression*)

Primjer metode kompresije bez gubitka je referencijska kompresija genoma. Kod tog ristupa razlikujemo dvije vrste genoma, referencijski i ciljni genom. Referencijski genom predstavlja poznatu i standardiziranu sekvencu, dok ciljni genom predstavlja genom koji želimo kompresirati. Mapirajući ciljni genom na referencijski genom, pohranjujemo samo razlike između genoma. Time značajno smanjujemo količinu podataka i povećavamo brzinu izvođenja. Slika 1 prikazuje jednostavan primjer referencijske kompresije.

A T G C T A C G G A

A T G G T A C G G A

[pozicija: 4, baza: G]

Referentni genom:

Ciljni genom:

Rezultat kompresije:

Slika : Primjer referencijske kompresije genoma

## Dekompresija genoma

Dekompresija genoma postupak je rekonstrukcije izvorne genomske informacije (npr. sekvence DNA) iz prethodno kompresiranih podataka. Moguće je dekompresirati podatke dobivene kompresijom bez gubitka i kompresijom s gubitkom. Kod dekompresiranja podataka dobivenih kompresijom bez gubitka moguća je potpuno točna dekompresija genoma, dok kod kompresije s gubitkom ne dobivamo potpuno identične informacije o genomu.

Referencijska dekompresija koristi se referencijskim genomom i rezultatom kompresije, odnosno spremljenim razlikama između referencijskog i ciljnog genoma, za dobivanje potpuno točne rekonstruirane genomske informacije. Slika 2 prikazuje jednostavan primjer referencijske dekompresije genoma.

Slika : Primjer referencijske dekompresije genoma

A T G C T A C G G A

[pozicija: 4, baza: G]

A T G G T A C G G A

Referentni genom:

Rezultat kompresije:

Rezultat dekompresije:

## Mapiranje ciljnog genoma na referencijski genom

Najteži dio referencijskog mapiranje je optimizacija procesa mapiranje ciljnog genoma na referencijski genom. Mapiranje je reprezentirano s dvije strukture podataka. U prvu strukturu spremamo parove pozicije i duljine podudaranja, a u drugu strukturu spremamo nepodudarne segmente ciljnog genoma, koji se ne mogu rekonstruirati iz referentnog genoma.

Preduvjet za mapiranje je identifikacija podudarnosti između ciljnog i referencijskog genoma. Taj proces može biti spor i memorijski skup. Najčešće korištene metode identifikacije podudarnosti su primjena sufiksne liste i sufiksnog stabla, tablice sažetka (engl. *hash table*) te modificiranih algoritama i adaptivnih pristupa.

Metoda tablica sažetka dijeli referencijski genom na manje podnizove čije pozicije uzima kao ključeve u tablici sažetka. Kao vrijednost u tablicu sažetka pohranjujemo sažetu vrijednost. Za mapiranje ciljnog genoma moramo klizećim putem prolaziti kroz ciljni genom i tražiti podudaranje među ključevima tablice sažetka referencijskog genoma. Nakon identificiranja svih podudaranja moramo identificirati postoje li duži nizovi podudaranja od opažanih. Taj proces možemo napraviti pohlepnim algoritmom (engl. *Greedy algorithm*). Cilj je otkriti najduže nizove kod kojih nema preklapanja.

## Algoritam referencijske kompresije genoma

U projektu koristit ćemo se implementaciju algoritma referencijske kompresije genoma. Cilj nam je ostvariti brzi i učinkoviti algoritam koji mapira ciljni genom na referentni genom.

Ulazni podaci u model su datoteke ciljnog i referentnog genoma. Prvi korak algoritma je priprema podataka tijekom koje uklanjamo identifikatore sekvence. S obzirom da radimo s provjerenim podaci pretpostavit ćemo da su sva slova u sekvenci velika slova. Zbog važnosti memorije i brzine izvođenja, enkodiramo nukleotide u dvobitne integer vrijednosti, po principu A = 00 = 0, C = 01 = 1, G = 10 = 2, T = 11 = 3. Slika 3 prikazuje primjer enkodiranja nukleotida unutar genoma.

A T G C T A C G G A

0 3 2 1 3 0 1 2 2 0

Originalni genom:

Enkodirani genom:

Slika : Primjer enkodiranja nukleotida

Nakon enkodiranja ciljnog i referentnog genoma potrebno je mapirati ciljni genom na referentni. Algoritam korišten u projektu koristi metodu tablice sažetka. Za računanje ključeva tablice sažetka definiramo duljinu podnizova 20. Pri računanju vrijednosti funkcije sažetka koristimo se sljedećom formulom:

Oznakom *k* označava se duljina podniza, *g* označavamo genom, i označava redni broj podniza, dok j označava redni broj elementa u podnizu. Prvo izračunamo tablicu sažetka za referentni genom. Nakon toga za svaki podniz ciljnog genoma računamo vrijednost funkcije sažetka i određujemo maksimalno podudaranje. Koristimo se pohlepnim algoritmom da pronađemo najduže podudaranje. Ostatak sekvence koji preostane neobuhvaćen podudaranjem predstavlja reziduale referentnog i ciljnog genoma. Slika 4 prikazuje jednostavan primjer mapiranja, tablicom sažetka, ciljnog genoma na referentni genom.

0 0 3 2 3 0 2 2 3 0 1 0 3 0 0 2 0 2 1

0 0 3 2 1 0 2 2 3 0 1 2 1 3 0 0 2 0 0 2 2

[(1,4),(1),(6,6),(21),(13,5),(0),(7,2)]

Referentni genom:

Ciljni genom:

Rezultat mapiranja:

Slika : Primjer mapiranja tablicom sažetka

Nakon mapiranja preostaje provesti postupak delta enkodiranja. To je postupak koji sprema samo reziduale referentnog i ciljnog genoma, i njihove pozicije. Slika 5 prikazuje jednostavan primjer delta enkodiranja koje se nastavlja na proces mapiranja sa Slika 4.

[(1,4),(1),(6,6),(21),(13,5),(0),(7,2)]

[(5,1),(12,21),(19,0)]

Rezultat mapiranja:

Delta enkodiranje:

Slika : Primjer delta enkodiranja

## Algoritam referencijske dekompresije genoma

# Skup podataka

Testiranje algoritma referencijske kompresije genoma napravljeno je nad skupovima podataka *E. coli* (*Estherichia coli*). Podaci su preuzeti s internetske stranice Nacionalne knjižnice medicine (engl. *National Library of Medicine*, NLM), Nacionalnog centra za biotehnološke informacije (engl. *National Center for Biotechnology Information*, NCBI).

Korišten je FASTA format datoteka. Prvi red takvog dokumenta počinje znakom > i sadrži opis sekvence ili njen identifikator (ID) koji može biti biti npr. naziv gena, uzorka, organizma. Za potrebe projekta nije bitan sadržaj prvog red pa ga uklanjamo tijekom pripreme podataka. Sljedeći redovi sadrže sekvencu, zapisanu kao niz slova koja predstavljaju nukleotide. Za DNA to su slova A, T, G i C.

Kao referentni genomi koriste se genomi znanstvenih naziva Escherichia coli str. K-12 substr. MG1655 (EcoCyc Project, SRI International, 2013) i Escherichia coli O157:H7 str. Sakai (GIRC, 2018). Status referentnih dobili su jer predstavljaju najpouzdaniju i najkvalitetniju sekvencu koja se može koristiti kao standard za cijelu vrstu. Kada postoji mnogo različitih varijanti vrste moguće je imati više od jednog referentnog genoma, kao što je slučaj s *E. coli* za koju postoje dva referentna genoma po bazi NCBI-a.

Baza podataka NCBI sadrži 342 303 genoma *E. coli*. Za potrebe projekta odabran je reprezentativan uzorak ciljnih genoma različitih duljina nizova. Najkraći je duljine X, dok je najduži duljine X. Ciljni genom je genom koji želimo kompresirat, mapirajući razlike na referentni genom.

# Rezultati

# Zaključak

# Literatura

1. Yuansheng Liu, Hui Peng, Limsoon Wong, Jinyan Li (2017), High-speed and high-ratio referential genome compression, Bioinformatics
2. Max Roser, Hannah Ritchie, Edouard Mathieu (2023), “What is Moore's Law?“, Our World in Dana
3. Scott Christley, Yiming Lu, Chen Li, Xiaohui Xie (2009), Human genomes as email attachments, Bioinformatics
4. Bioinformatika 1 (2025), Uvodno predavanje
5. National Center for Biotechnology Information (2025), Escherichia coli genomes, NCBI (<https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=562>)
6. Sebastian Wandelt, Ulf Leser (2013), FRESCO: Referential compression of highly similar sequences, Bioinformatics