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

Razvoj bioinformatike usko je povezan s tehnološkim napretkom računalnih sustava, što se djelomično može pripisati efektu tzv. Mooreovog zakona [1]. Mooreov zakon navodi da će se broj tranzistora u čipu, a 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 [2] (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 [3]. 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.

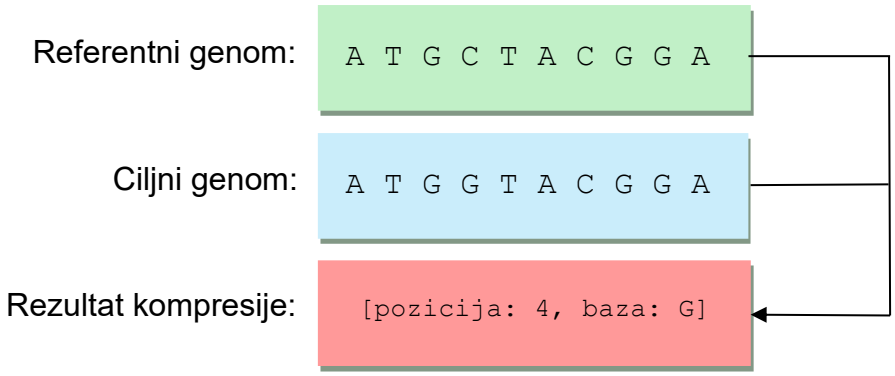
2. Referencijska kompresija i dekompresija genoma

2.1 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 pristupa 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.



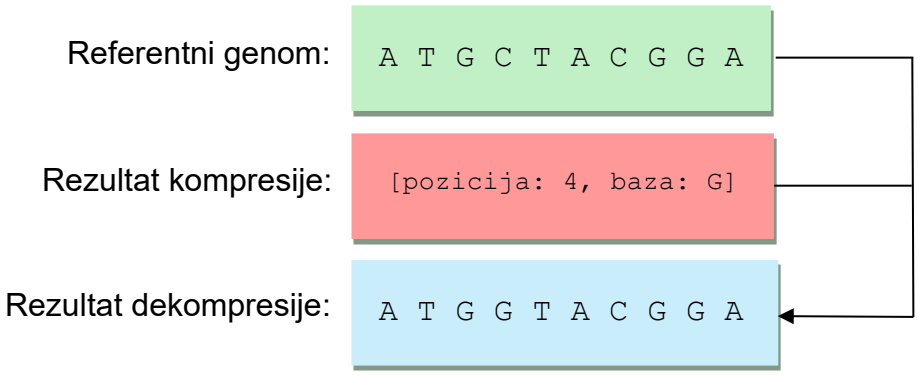
Slika 1: Primjer referencijske kompresije genoma

2.2 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 2: Primjer referencijske dekompresije genoma

2.3 Mapiranje ciljnog genoma na referencijski genom

Najteži dio referencijskog mapiranja je optimizacija procesa mapiranja ciljnog genoma na referencijski genom. Mapiranje je reprezentirano s dvije strukture podataka. U prvu strukturu spremamo parove pozicija 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.

2.4 Algoritam referencijske kompresije genoma

U projektu ćemo koristiti implementaciju algoritma referencijske kompresije genoma. Cilj nam je ostvariti brz i učinkovit algoritam koji mapira ciljni genom na referentni genom.

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

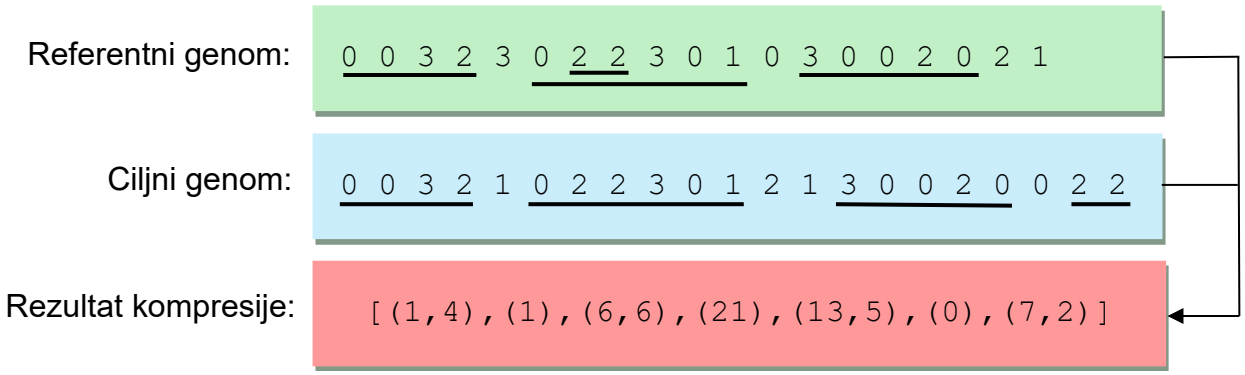
Slika 3: Primjer enkodiranja nukleotida

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 4. Pri računanju vrijednosti funkcije sažetka koristimo se sljedećom formulom:

$$V_i = \sum_{j=0}^{k-1} (g(i+j) \times 4^j)$$

Oznakom *k* označava se duljina podniza, s *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. Duljina podudaranja mora biti najmanje 2 nukleotida. Ostatak sekvence koji preostane neobuhvaćen podudaranjem predstavlja reziduale referentnog i ciljnog genoma. Slika 4 prikazuje jednostavan primjer referencijske kompresije mapiranjem, tablicom sažetka, ciljnog genoma na referentni genom. Dobivene rezultate spremamo u datoteku.

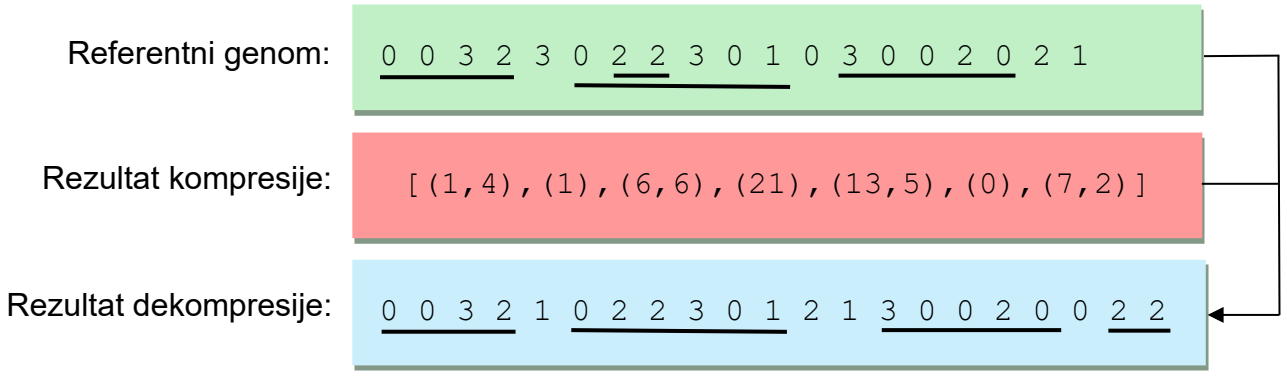


Slika 4: Primjer referencijske kompresije genoma

2.5 Algoritam referencijske dekompresije genoma

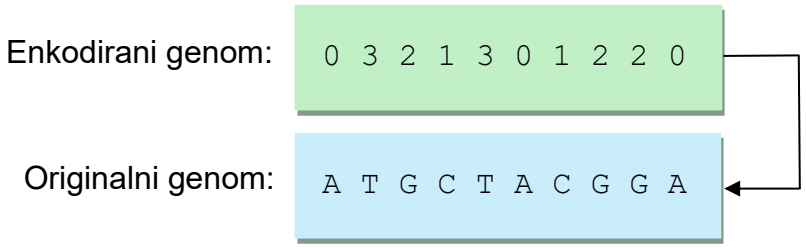
Ulazni podaci u algoritam referencijske dekompresije su referentni genom i rezultat kompresije. Algoritam se sastoji od procesa rekonstrukcije ciljnog genoma te provođenja prilagodbe podataka u standardni oblik.

Rekonstrukcija ciljnog genoma provodi se popunjavanjem ciljnog genoma s blokovima sekvenci podudaranja između kojih se dodaju reziduali ciljnog i referentnog genoma kako bi se očuvala potpuna informacija. Slika 5 prikazuje jednostavan primjer referencijske dekompresije genoma.



Slika 5: Primjer referencijske dekompresije genoma

Dekompresirani zapis sastoji se od dvobitnih numeričke vrijednosti, koje je potrebno dekodirati vrijednostima nukleotida po principu 0 (00) = A, 1 (01) = C, 2 (10) = G, 3 (11) = T. Dobiveni rezultat predstavlja rekonstruirani ciljni genom. Spremamo ga u datoteku. Slika 6 prikazuje primjer dekodiranja nukleotida unutar genoma.



Slika 6: Primjer dekodiranja nukleotida

3. 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) [4].

Korišten je FASTA format datoteka. Prvi red takvog dokumenta počinje znakom > i sadrži opis sekvence ili njen identifikator (ID) koji može biti npr. naziv gena, uzorka, organizma. Za potrebe projekta nije bitan sadržaj prvog reda 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 *E. coli* 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. U projektu koristit ćemo ASM584v2 (*Escherichia coli str. K-12 substr. MG1655*) kao glavni referentni genom, a u specijalnim slučajevima dodatno ćemo testirati na referentnom genomu ASM886v2 (*Escherichia coli 0157:h7 str. Sakai*).

Baza podataka NCBI sadrži 342 303 genoma *E. coli*. Za potrebe projekta odabran je reprezentativan uzorak ciljnih genoma, koji uključuje ASM1038v1, ASM1326v1, ASM21047v1, ASM285371v1 i ASM369716v2 genom. Ciljni genom je genom koji želimo kompresirati, mapirajući razlike na referentni genom.

4. Rezultati

Algoritmi referencijske kompresije i dekompresije genoma implementirani su koristeći programski jezik C++. Uz njih implementirana je funkcionalnost provjere točnosti, koja uzima ciljni genom i uspoređuje ga s dekompresiranim genomom. Za provođenje algoritama korišteno je osobno računalo, stoga su vremena izvođenja ograničena performansama računala i nije ih moguće uspoređivati s drugim objavljenim implementacijama istog algoritma, poput *High-speed and high-ratio referential genome compression* [5] koji je korišten kao temeljni članak.

Početno testiranje napravili smo na jednostavnom primjeru. Tablica 1 prikazuje ulazne i izlazne podatke jednostavnog primjera referencijske kompresije i dekompresije genoma. Referentni i ciljni genom uzeti su različitih duljina kako bi se pokazala robusnost implementiranog algoritma. Rezultat kompresiranog genoma interpretiramo tako da u slučaju podudaranja imamo strukturu podataka (početak podudaranja, duljina podudaranja), dok u slučaju nepodudaranja imamo strukturu podataka (sekvenca nepodudaranja). Na jednostavnom primjeru algoritam postiže odličnu točnost te vremensku i memorijsku učinkovitost.

Referentni genom	AAAAGCTTCG
Ciljni genom	AAAATCTTCGAACAG
Kompresirani genom	[(1,4),(8,2),(7,4),(1,2),(1),(4,2)]
Dekompresirani genom	AAAATCTTCGAACAG
Točnost	100%
Vrijeme izvođenja kompresije	0.003s
Vrijeme izvođenja dekompresije	0.003s
Količina zauzete memorije	4.00 KB

Tablica 1: Primjer referencijske kompresije i dekompresije genoma

Algoritmi su testirani na stvarnim podacima za *E. coli*. Korišten je referentni genom ASM584v2, te ciljni genomi ASM1038v1, ASM1326v1, ASM21047v1, ASM285371v1 i ASM369716v2. U specijalnim slučajevima koristimo referentni genom ASM886v2 kako bi provjerili utjecaj referentnog genoma na prostor potreban za pohranu.

U nastavku prikazat ćemo rezultate algoritma.

Referentni genom	ASM584v2
Ciljni genom	ASM1038v1
Točnost	100%
Vrijeme izvođenja kompresije	3m 33.616s
Vrijeme izvođenja dekompresije	0.115s
Količina zauzete memorije	1.51 MB
Originalno zauzeta memorija	4.71 MB

Tablica 2: Rezultati za genom ASM1038v1

Tablica 2 prikazuje rezultate referencijske kompresije i dekompresije za genom ASM1038v1. Ostvarujemo značajno poboljšanje na području memorijske učinkovitosti, gdje za pohranu kompresiranog genoma trebamo preko tri puta manje prostora, u odnosu na originalni zapis genoma.

Referentni genom	ASM584v2
Ciljni genom	ASM1326v1
Točnost	100%
Vrijeme izvođenja kompresije	6m 59.690s
Vrijeme izvođenja dekompresije	0.162s

Količina zauzete memorije	2.77 MB
Originalno zauzeta memorija	4.89 MB

Tablica 3: Rezultati za genom ASM1326v1

Tablica 3 prikazuje rezultate referencijske kompresije i dekompresije za genom ASM1326v1. Ostvarujemo značajno poboljšanje na području memorijske učinkovitosti, gdje za pohranu kompresiranog genoma trebamo preko 1.75 puta manje prostora, u odnosu na originalni zapis genoma.

Referentni genom	ASM584v2
Ciljni genom	ASM21047v1
Točnost	100%
Vrijeme izvođenja kompresije	2m 52.471s
Vrijeme izvođenja dekompresije	0.117s
Količina zauzete memorije	1,17 MB
Originalno zauzeta memorija	4.97 MB

Tablica 4: Rezultati za genom ASM21047v1

Tablica 4 prikazuje rezultate referencijske kompresije i dekompresije za genom ASM21047v1. Ostvarujemo značajno poboljšanje na području memorijske učinkovitosti, gdje za pohranu kompresiranog genoma trebamo preko četiri puta manje prostora, u odnosu na originalni zapis genoma.

Referentni genom	ASM584v2
Ciljni genom	ASM285371v1

Točnost	100%
Vrijeme izvođenja kompresije	12m 26.356s
Vrijeme izvođenja dekompresije	0.193s
Količina zauzete memorije	5.22 MB
Originalno zauzeta memorija	4.74 MB

Tablica 5: Rezultati za genom ASM285371v1

Tablica 5 prikazuje rezultate referencijske kompresije i dekompresije za genom ASM285371v1. Kompresirani genom koristi više memorijskog prostora od originalnog zapisa genoma, stoga možemo reći da u ovom slučaju algoritam nije memorijski učinkovit. Dodatno ćemo testirati algoritam korištenjem drugog referencijskog genoma.

Referentni genom	ASM886v2
Ciljni genom	ASM285371v1
Točnost	100%
Vrijeme izvođenja kompresije	14m 58.297s
Vrijeme izvođenja dekompresije	0.212s
Količina zauzete memorije	5.14 MB
Originalno zauzeta memorija	4.74 MB

Tablica 6: Rezultati za genom ASM285371v1, uz referentni genom ASM886v2

Tablica 6 prikazuje rezultate referencijske kompresije i dekompresije za genom ASM285371v1, za referentni genom ASM886v2. Kompresirani genom koristi više memorijskog prostora, no u ovom slučaju nešto manje od kompresiranog genoma za referencijski genom ASM584v2.

Referentni genom	ASM584v2
Ciljni genom	ASM369716v2
Točnost	100%
Vrijeme izvođenja kompresije	12m 37.919s
Vrijeme izvođenja dekompresije	0.215s
Količina zauzete memorije	5.22 MB
Originalno zauzeta memorija	4.73 MB

Tablica 7: Rezultati za genom ASM369716v2

Tablica 7 prikazuje rezultate referencijske kompresije i dekompresije za genom ASM369716v2. Kompresirani genom koristi više memorijskog prostora od originalnog zapisa genoma, stoga možemo reći da u ovom slučaju algoritam nije memorijski učinkovit. Dodatno ćemo testirati algoritam korištenjem drugog referencijskog genoma.

Referentni genom	ASM886v2
Ciljni genom	ASM369716v2
Točnost	100%
Vrijeme izvođenja kompresije	15m 3.794s
Vrijeme izvođenja dekompresije	0.212s
Količina zauzete memorije	5.15 MB
Originalno zauzeta memorija	4.73 MB

Tablica 8: Rezultati za genom ASM369716v2, uz referentni genom ASM886v2

Tablica 8 prikazuje rezultate referencijske kompresije i dekompresije za genom ASM369716v2, za referentni genom ASM886v2. Kompresirani genom koristi više memorijskog prostora, no u ovom slučaju nešto manje od kompresiranog genoma za referencijski genom ASM584v2.

Iz prethodnih rezultata možemo vidjeti da algoritam referencijske kompresije i dekompresije genoma radi sa 100% točnosti. Vremensku učinkovitost je zadovoljavajuća. Memorijska učinkovitost varira između različitih ciljnih genoma, no ostvarujemo zadovoljavajuće rezultate, gdje najbolji testni primjer ostvaruje kompresiju genoma za četiri puta.

Implementacija algoritma ima određene izazove. Zbog načina pohranjivanja podataka potrebno je obratiti pažnju na pohranjivanje višeznamenkastih brojeva, kako ne bi došlo do neočekivanog funkcioniranja algoritma na većim primjerima. Algoritam s takvim problemom rezultira padom performansi na točnost do 10%.

5. Zaključak

Algoritam referencijske kompresije i dekompresije genoma pokazao je zadovoljavajuću memorijsku i vremensku učinkovitost kod jednostavnijih, umjetno generiranih primjera. Ipak, prilikom primjene na stvarnim genomskim podacima *E. coli,* kod nekih primjera uočen je pad performansi na području učinkovitosti memorijske pohrane, što ukazuje na nedostatke u postojećoj implementaciji algoritma.

Unatoč navedenim ograničenjima, referencijska kompresija ostaje obećavajući pristup, osobito u kontekstu sve veće količine sekvencijskih podataka u bioinformatičkim istraživanjima. Daljnjim unaprjeđenjem algoritama i prilagodbom stvarnim podacima, moguće je postići primjetne uštede u memorijskoj pohrani i ubrzanje obrade genomskih informacija, čime se otvaraju nove mogućnosti za njihovu primjenu u znanstvenim, istraživačkim i bioinformatičkim disciplinama.

6. Literatura

[1] Max Roser, Hannah Ritchie, Edouard Mathieu (2023), What is Moore's Law?, Out World in Data

[2] Scott Christley, Yiming Lu, Chen Li, Xiaohui Xie (2009), Human genomes as email attachments, Bioinformatics

[3] Sebastian Wandelt, Ulf Leser (2013), FRESCO: Referential compression of highly similar sequences, Bioinformatics

[4] National Center for Biotechnology Information (NCBI) (2013), Escherichia coli genomes, https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=562

[5] Yuansheng Liu, Hui Peng, Limsoon Wong, Jinyan Li (2017), High-speed and high-ratio referential genome compression, Bioinformatics

[6] Bioinformatika 1 (2025), Uvodno predavanje