SVEUČILIŠTE U ZAGREBU

**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

**Bioinformatika 1**

**High-speed and high-ratio referential genome compression algorithm**

*Katarina Mišura, Marko Marfat*

Voditelj: *Mirjana Domazet-Lošo*

Zagreb, svibanj, 2022.

**Sadržaj**

[Uvod 1](#_Toc103122008)

[1. Koraci algoritma 1](#_Toc103122009)

[1.1. Pretprocesiranje 1](#_Toc103122010)

[1.2. Pohlepno podudaranje pomoću *hash* tablice 1](#_Toc103122011)

[1.3. Postprocesiranje 1](#_Toc103122012)

[1.4. Dekompresija 1](#_Toc103122013)

[2. Rezultati testiranja 1](#_Toc103122014)

[Zaključak 1](#_Toc103122015)

[Literatura 1](#_Toc103122016)

[Sažetak 2](#_Toc103122017)

# Uvod

U ovome će se projektu, iz predmeta Bioinformatika 1, implementirati algoritam za kompresiju genoma. Puni naziv algoritma je „High-speed and high-ratio referential genome compression“ ili „HiRGC“, a cilj mu je uvesti poboljšane performanse u odnosu na već dostupne algoritme za kompresiju genoma [1].

Motivacija za uvođenje ovakvog algoritma je sve veći broj sekvenciranih genoma koji zahtijevaju velike količine memorijskog prostora za pohranu. Neke od tradicionalnih metoda kompresije poput „gzip-a“ (<http://www.gzip.org/>) i „bzip2-a“ (<http://www.bzip.org>) nisu dovoljno učinkovite u kompresiji genomskih podataka. Razlog tomu je što ovakve metode ne iskorištavaju intrinzična svojstva DNA sekvenci poput male abecede, čestih ponavljanja i palindroma. Preferirani pristup je kompresija bez gubitaka temeljena na pohranjivanju razlike između ciljnog genoma i referentnog genoma. Upravo to je pristup na kojemu se zasniva ovaj algoritam.

# Koraci algoritma

Pri pokretanju program kao argumente prima imena referentnog i ciljnog FASTA dokumenta. Nakon provjere točnosti unesenih argumenata može se započeti daljni proces. Cijeli algoritam možemo podijeliti u četiri cjeline: pretprocesiranje, pohlepno podudaranje pomoću *hash* tasblice, posprocesiranje i dekompresija. Svaka cjelina je opisana u nastavku.

## **Pretprocesiranje**

Pretprocesiranje je proces pripreme sekvenci referentnog i ciljnog FASTA dokumenta za daljnje uspoređivanje i kompresiju. Cilj preprocesiranja je izolirati sve znakove iz sekvence ciljnog FASTA dokumenta koji nisu unutar skupa znakova Fi koji sadrži znakove A, C, G, T. Prilikom uklanjanja tih znakova potrebno je sačuvati informacije o pozicijama i duljini intervala tih znakova u ciljnoj sekvenci. Za referentu sekvencu potrebno je ukloniti sve znakove koji nisu unutar zadanog skupa znakova Fi. Očuvane pomoćne informacije o znakovima koji se ne nalaze u skupu znakova Fi biti će potrebne kasnije pri dekompresiji i ponovnom sastavljanju cjelokupne ciljne sekvence.

Pretprocesiranje možemo podijeliti u tri dijela. Prvi dio je procesiranje ciljne sekvence, drugi dio je procesiranje referentne sekvence, a treći dio je spremanje pomoćnih infomracija o ciljnoj sekvenci u dokument.

Funkcija *target\_preprocess()* kao argument prima niz znakova *file\_name* koji predstavlja ime FASTA dokumenta iz kojeg učitavamo ciljnu sekvencu. Funkcija započinje čitanjem redaka iz dokumenta i izdvajanjem prvog redka u kojem se nalaze informacije o sekvenci. Ostali redci se čitaju jedan po jedan i spremaju se u niz znakova, a duljina svakog retka se sprema u vektor brojeva. Nakon učitavanja svih redaka iz dokumenta u niz znakova zatvara se pristup dokumentu.

//nastaviti opis dalje – stala na liniji 73 – opisati do 255.

## **Pohlepno podudaranje pomoću *hash* tablice**

## **Postprocesiranje**

## **Dekompresija**

# Rezultati testiranja

# Zaključak

# Literatura

[1] Liu, Y., Peng, H., Wong, L., Li, J. (2017). *High-speed and high-ratio referential genome compression*. Bioinformatics Vol. 33 Iss 21 pp. 3364-3372. Poveznica: <https://doi.org/10.1093/bioinformatics/btx412>; pristupljeno 11. svibnja 2022.

# Sažetak