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

## **Pretprocesiranje**

## **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