SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

HRCM algoritam

Leon Hegedić, Ivan Terzić

Voditelj: Mirjana Domazet-Lošo

Zagreb, svibanj, 2024.

**Sadržaj**

# Uvod

U području bioinformatike čest je rad s dugačkim nizovima znakova. Neka taj dugi niz znakova predstavlja niz baza A, C, G i T u DNA molekuli. Ta sekvenca baza određuje genetsku informaciju koja je cijelom svojom duljinom bitna [1], tj. ne smije se izostaviti niti jedna baza u zapisu baze. Ako se uzme u obzir da se tim nizom može predstaviti i cijeli genom pojedine vrste koji može biti dug od 50 000 pa sve do preko 100 000 000 000 znakova, odnosno baza (za primjer, genom čovjeka ima ~3 000 000 000 baza. [2] Ako se uzme u obzir da se slova u današnjim računalima pohranjuju s jednim bajtom, odnosno 8 bitova, neki niz baza dug 1 000 000 znakova zauzeo bi 1MB, a 100 000 000 gotovo 100MB, a to nije ni približno dovoljno za pohranu genoma na primjer kralježnjaka, jasno je da je potreban efikasan način za pohranu tih niza baza. Ovaj se rad bavi upravo takvim algoritmom: HRCM algoritmom. Rad se bazira na proučavanju algoritma, opisu teoretskih pojmova nužnih za razumijevanje algoritma, opisu samog algoritma te dokumentiraju testiranja algoritma.

# Teorija

Za razumijevanje samog algoritma, potrebno je razumjeti sljedeće pojmove:

* Baza, nukleotid, DNA, gen, genom, sekvenca
* Kompresija, dekompresija
* FASTA format
* k-mer

# Opis algoritma

HRCM algoritam sastoji se od dvije glavne faze: kompresija i dekompresija. U fazi kompresije se sekvence spremaju u

# Primjer izvođenja

# Testiranje

# Zaključak

# Literatura

[1] <https://www.fer.unizg.hr/_download/repository/Bioinformatika_1%20-%20Uvodno_2023_2024.pdf>

[2] https://www.fer.unizg.hr/\_download/repository/Bioinformatika\_%201\_2-3\_predavanje\_Dinami%C4%8Dko\_programiranje.pdf

[3] Haichang Yao, Yimu Ji, Kui Li, Shangdong Liu, Jing He, Ruchuan Wang, "HRCM: An Efficient Hybrid Referential Compression Method for Genomic Big Data", BioMed Research International, vol. 2019, Article ID 3108950, 13 pages, 2019. https://doi.org/10.1155/2019/3108950