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 datoteke sažetog oblika (manjeg memorijskog zauzeća) koje sadržavaju potpunu informaciju o sekvenci, dakle ne gube se informacije ni o malim slovima, ni specijalnim znakovima niti „X” znakovima. U fazi dekompresije se ti pohranjeni podatci vraćaju u svoj izvorni oblik, kako je primljen na ulazu u algoritam.

## Kompresija

Algoritam na svoj ulaz u fazi kompresije prima referentnu datoteku genoma (referentnu sekvencu) i jednu ili više sekvenci koje se trebaju komprimirati, sve datoteke u FASTA formatu. Sekvence za komprimiranje komprimiraju se pomoću referentnog genoma. Sama faza kompresije algoritma dijeli se u 3 glavna koraka: izvlačenje informacija sekvenci, pronalazak podudarnih informacija u nizovima te kodiranje informacija o sekvencama.

### Izvlačenje informacija o sekvencama

U prvom koraku kompresije, na ulazu samog algoritma, nalaze se datoteke sekvenci namijenjene za kompresiju te referentna sekvenca u FASTA formatu. Za svaku se sekvencu pohranjuju sljedeće informacije: identifikator, sekvenca osnovnih baza (A, C, G i T), informacije o pozicijama malih slova, informacije o slovima N, informacije o ostalim znakovima te broj koji označava duljinu retka. Za referentnu sekvencu pohranjuju se samo sekvenca osnovnih baza i informacije o malim slovima, ona ne mora pamtiti informacije o ostalim slovima, itd.

Sam proces obavlja se na sljedeći način: obzirom na FASTA format, prva linija u datoteci je tzv. identifikator, on se izvlači i sprema kao identifikator pojedine datoteke. Nakon toga se, za prvu liniju koja sadrži niz baza, gleda njena duljina te pohranjuje kao duljina retka. Nakon toga se u nizu baza redom bilježe pozicije malih slova te potom i N slova i posebnih slova. Obzirom da je očekivano da će algoritam raditi s izuzetno dugim nizovima znakova, pohrana informacija o malim i N slovima se radi pomoću pohrane relativnog indeksa (udaljenosti) od prošle pojave malog ili N slova (naravno, prošla pojava malog slova se gleda od prošle pojave malog slova, itd.) i pohrane duljine tog niza znakova. Tako se izbjegava pohrana velikih brojeva kao pozicija, što bi bilo očekivano za slučaj kad bi se koristili apsolutni indeksi u nizu znakova. Nakon pohrane tih informacija, znakovi se spremaju u jedan niz znakova, brišu se posebni znakovi i N znakovi te se dobije čist niz znakova A, C, G ili T, odnosno baza. Za referentu sekvencu postupak je analogan, osim što se spremaju samo podatci potrebni za taj niz, dakle niz baza i podaci o malim slovima. Iako se informacije ne spremaju, također se brišu N znakovi i specijalni znakovi. Ovako zapisani nizovi znakova ulaze u drugu fazu kompresije: pronalazak podudarnih informacija.

### Pronalazak podudarnih informacija

U drugoj fazi kompresije, pronalasku podudarnih informacija, kao ulaz primamo nizove znakova zapisane kako je opisano u prošlom odlomku, točno 1 referentan niz znakova i barem jednu sekvencu koja je namijenjena za kompresiju. U ovoj fazi cilj je pronaći podudarnosti između referentne sekvence i sekvenci koje se komprimiraju kako bi se što više smanjilo memorijsko zauzeće zapisa. Ova se faza sastoji od pronalaska podudaranja u nizu baza te u malim slovima. Ukoliko postoji više od jedne sekvence koje se trebaju komprimirati, algoritam koristi i tzv. podudaranje drugog reda (engl. *second-level matching*).

Prvo se u ovoj fazi radi pronalazak podudaranja u sekvencama

### Kodiranje

## Dekompresija

# 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