SVEUČILIŠTE U ZAGREBU

**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

Bioinformatika

**Projekt**

**Improving Bloom Filter Performance on Sequence Data Using k-merBloom Filters**

Josipa Kelava, 0036478602

Ivan Krpelnik, 0036478618

Tea Pović, 0036476912

Zagreb, siječanj 2018.

Sadržaj

[1. Algoritmi 3](#_Toc503535816)

[1.1 Jednostrani k-mer Bloomov filter 3](#_Toc503535817)

[1.2 dvostrani 3](#_Toc503535818)

[1.3 tri 3](#_Toc503535819)

[1.4 četiri 3](#_Toc503535820)

[1.5 pet 3](#_Toc503535821)

# Algoritmi

Problem oko kojeg je fokusiran rad je spremanje k-merova. K-mer predstavlja kraće podnizove koji su jednake duljine i dobiveni su iz zadane sekvence. Sekvencu želimo podijeliti na više dijelova odnosno k-merova jer algoritmi zasnovani na upotrebi k-merova uvelike poboljšavaju usporedbu sekvenci [1].

Sekvenca zbog svoje veličine može generirati stotine milijuna k-merova. Spremanje takvog broja predstavlja veliki problem te se često za spremanje koristi Bloomov filter. Pozitivna stvar kod korištenja Bloomovih filtera osim smanjena memorije jest ta da se vrlo brzo može provjeriti postojanje određenog K-mera, ali uz mogućnošću pojave lažno pozitivnih rezultata.

U zadanom radu implementiraju se različiti k-mer Bloomovi filteri za koje se testira koliko unaprjeđuju upotrebu običnog Bloomovog filtera. Želi se smanjiti broj lažno pozitivnim k-merova. Zadani rad sadrži 4 implementirana algoritma te jedan algoritam koji nije prikazan u rezultatima rada jer radi nad jednom sekvencom.

Osnovna ideja svih algoritama je da se za trenutno traženi k-mer provjeravaju i susjedi kako bi se smanjio broj lažno pozitivnih.

## Jednostrani k-mer Bloomov filter

U ovom algoritmu stvara se k-mer Bloomov filter koji provjerava postojanje svih mogućih susjeda za trenutni k-mer. Ukoliko pretraga Bloomovog filtera za ni jedan od susjeda ne vrati pozitivnu vrijednost tada se sa sigurnošću može reći da je k-mer bio lažno pozitivan. Ukoliko pretraga susjeda vrati pozitivan rezultat ne možemo tvrditi da je traženi sigurno pozitivan jer je i susjedni k-mer kao i traženi mogao biti lažno pozitivan.

Za svaki k-mer se stvara 8 susjeda, baze A, C, T, G se dodaju s lijeve ili desne strane.

Primjer:

sekvenca: ATGTCGGGTC

traženi k-mer: CGC

generirani susjedi: ACG, TCG, CCG, GCG, GCA, GCT, GCC, GCG

U primjeru se vidi da k-merovi susjeda ne postoje u sekvenci te ukoliko ni jedan od njih ne vrati lažno pozitivan rezultat za traženi k-mer se sa sigurnošću može reći da nije sadržan u sekvenci.

## Dvostrani k-mer Bloomov filter

## 1.3 tri

## 1.4 četiri

## 1.5 pet

fhhf

# Rezultati testiranja

rezultati, tablie, grafovi

# Zaključak

Nešto pametno o radu :D

# Literatura

1. D. Pellow, D. Filippova, C. Kingsford, Improving Bloom Filter Performance on Sequence Data Using k-mer Bloom Filters, Journal of computational biology, Volume 24, Number 6, 2017