SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Analiza i opis alata za analizu metagenoma

Mateo Stjepanović
Voditelj: *Mirjana Domazet-Lošo*

Zagreb, travanj 2018.

SADRŽAJ

1. Uvod

Bioinformatika kao interdisciplinarna znanost raste kroz posljednje desetljeće. Najviše se ističe proučavanje i analiza metagenoma. Iako brzorastuća znanost još uvijek nailazi na mnoge probleme koje pokušava riješiti. Neka od tih rješenja će biti predstavljena kroz ovaj rad, te će se pokušati usporediti iz više pogleda. Usporedit će se zauzeće memorije, vrijeme izvedbe te kvalitetu rada.

Zauzeće memorije se ispituje na dva načina. Zauzeće radne memorije (RAM), i memorija koja je potrebna za pohranu, s idejom da će zauzeće radne memorije za dimenziju strožije od memorije za pohranu. Želimo imati proizvod koji će funkcionirati s što boljom preciznošću i osjetljivošću, ali i da se može istovremeno koristiti na osobnim računalim (bez potrebe za *high-end* računalima).

Vrijeme izvedbe će se također ispitivati na dva načina. Brzinu podešavanja sustava da bude spreman za korištenje, te samo vrijeme koje je potrebno da bi se određeni skup podataka klasificirao. Pod podešavanje sustava svrstavamo kreiranje baze podataka koja se koristi, te samu instalaciju sustava.

Podaci koji će se ispitivati su izvučeni s NCBI stranice, te smješteni u lokalni *file system*, a baze podataka će biti referente baze podataka koje su predstavljene na službenim stranicama alata, te će biti posebno naglašene kroz daljni tekst.

Kvaliteta rada se odnosi na preciznos i osjetljivost (odziv). Preciznost - udio točno klasificiranih primjera u skupu svih pozitivno klasificiranim primjerima. Odziv - udio pozitivno klasificiranih primjera u skupu svih pozitivnih primjera.

Relativno velik broj alata je do sada pokušao riješiti probleme između preciznosti i odziva. Neki od njih staju na stranu odziva gdje žele klasificirati što više podataka, iako bi to moglo značiti da je veliki broj također pogrešno klasificiran. U tom tonu se razvijaju dvije vrste alata. Alati s potpunim i nepotpunim poravnanjem. Alati s potpunim poravnanjem rade na način da da klasificiraju samo one podatke koje su uspjeli naći u bazi podataka sa stopostotnim podudaranjem. Podaci koji to ne uspiju ostaju neklasificirani. Alati s nepotpunim poravnanjem rade na način da uzimaju male dijelova podataka te traže potpunu točnost s nekim dijelom podataka unutar baze. Tada

šire testni podatak te traže ono podudaranje s kojim se može testni podatak najviše proširiti. Upravo tim postupkom podižemo razinu odziva, a iz navedenog načina rada možemo vidjeti zašto je preciznost upitna.
Kroz ovaj rad upoznat ćemo se i s jednim alatom koji pokušava spojiti oba načina te tvrdi da ima najbolji omjer preciznosti i odziva. Bit će obrađen u svom posebnom naslovu.
Iako će se testirati u *laboratorijskom okruženju* rezultati koji bi imali najviše težine su oni iz prirodnog okruženja. Budući da za potrebe ovog rada navedni podaci nisu mogući te informacije i tablice preuzeti iz službenih izvora.

2. Alati s potpunim poravnanjem

Alati s potpunim poravnanjem rade na način da se ulazni podatak, u nekom od formata priklanim za obradu, pretvori u niz k-mera te se traže točna poravnanja s nekim od podataka unutar baze podataka. Kada se nadje takav klasificira se onaj koji ima najveći *score*. U nastavku ovog poglavlja predstavljena su dva referentna alata s potpunim poravnanjem, te jedan koji je mješavina potpunog i nepotpunog poravnanja.

2.1. Kraken

2.2. Clark

2.3. Kaiju - prijelaz između potpunog i nepotpunog poravnanja

3. Alati s nepotpunim poravnanjem

3.1. Centrifuge

4. Statistička obrada

5. Zaključak

Zaključak.

6. Literatura

Florian P. Breitwieser Daehwan Kim, Li Song i et al. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, 26:1721 – 1729, 2016.

Derrick E. Wood i Steven L.Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology 15:R46*, 2014.

S. Lindgreen i et al. An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports*, 6:19233 –, 2016.

P. Menzel i et al. Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature Communications*, 7:11257–, 2016.

Rachid Ounit i Stefano Lonardi. Higher classification sensitivity of short metagenomic reads with clark-s. *Bioinformatics*, 32:3823 – 3825, 2016.

Close TJ Lonardi S. Ounit R, Wanamaker S. Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, 16, 2015.

Ounit R (2015) Ounit i Lonardi (2016) Daehwan Kim i et al. (2016) Menzel i et al. (2016) E. Wood i L.Salzberg (2014) Lindgreen i et al. (2016)and

7. Sažetak

Sažetak.