SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Analiza i opis alata za analizu metagenoma

Mateo Stjepanović
Voditelj: *Mirjana Domazet-Lošo*

Zagreb, svibanj 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 kvaliteta rada. Pokušat će se pokazati i odrediti koji alat je najbolji za korištenje na prosječnom *kućnom* računalu.
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. Kao podatke pokušat će se naći oni koji bi najbolje predstavljali *real life situation*, npr. močvarna voda, uzorak iz ljudske utrobe itd.
Kvaliteta rada se određivat će se službenom metrikom.Od metrika koristit će se preciznost,osjetljivost (odziv) i F1 score. Preciznost - udio točno klasificiranih primjera u skupu svih pozitivno klasificiranim primjerima. Odziv - udio pozitivno klasificiranih primjera u skupu svih pozitivnih primjera. F1- harmonički prosjek koji se računa između preciznosti i osjetljivosti. Najbolji alat bi imao F1 score jednak 1 a najgori 0.
Relativno velik broj alata je do sada pokušao riješiti probleme između preciznosti i osjetljivosti. Neki od njih staju na stranu osjetljivosti 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

Kraken je ustvari predstavljen bazom podataka koja je srž navedenog alata. Ta baza je napravljena tako da se ne spremaju cijeli uzorci genoma, nego samo njihovi *značajni* dijelovi, tj. dijelovi genoma koji su specifični za određenu skupinu. Takav pristup značajno ubrzava rad, uz ideju da preciznost ostane na, što je više moguće, visokoj razini. Kraken baza podataka se sastoji od k-mera te podataka koji predstavljaju najmanjeg mogućeg pretka (eng. *Lowest common ancestor*), koji će se, radi jednostavnosti, u daljnem tekstu označavati kao *LCA*.
K-mer je dio niza veličina K. U granama kao što su pretraživanje teksta isti se nazivaju n-grami.
Budući da je Kraken alat s potpunim poravnanjem, tj. ne klasificira one genome za koje nema dovoljno dokaza pri pridruživanju podacima u bazi podataka, gubi na svojoj osjetljivosti. Također mu je poprilično teško odrediti niže razine pripadnosti nekoj vrsti. Više razine može klasificirati, jer ima mogućnost iz više k-mera koji se slažu pretražiti više razine i naći točke podudaranja.
Korištenjem specifičnog načina korištenja baze podataka, Kraken je uspio podići brzinu izvođenja na zavidnu razinu. Njegova brzina je 1.5 miliona očitanja po minuta, što je nekoliko redaka veličine brže od prvog sljedećeg alata u vrijeme kreiranja Krakena. Naime u novije vrijeme postoje neki alati koji zadanu akciju mogu izvršiti još brže, npr.Clark(32 miliona očitanja po minuti). Da bi sve funkcioniralo na navedeni način potrebno je cijelu bazu podataka učitati u radnu memoriju. Standardna Kraken baza

podataka je 70 GB i to može predstaviti problem pri korištenju Kraken-a na osobnim računalima. Upravo iz tog razloga su tvorci Kraken-a razvili puno manju bazu podataka imena MiniKraken. Baza podataka veličine 4 GB. Iako je osjetljivost osjetno niža (otprilike 11%), začudo, preciznost je, kroz testiranja tvoraca, veća da sve primjere.
Kroz testove iz originalnog rada se vidi da je Kraken, iznimno loš, pri otkrivanju genoma koji još nisu viđeni. Za usporedbu se vidi da je potrebno naći dodirnu točku između algoritama potpunog i nepotpunog poravnaja, kako bismo dobili najbolji omjer preciznosti i osjetljivosti, točnije kako bismo dobili što viši F1 score.
Postoje alati kao NBC koji imaju jako visoku razinu preciznosti i osjetljivosti, ali nažalost, kao što je navedeno, oni znaju biti jako zahtjevni i spori. Stoga se u posljednje vrijeme radi na alatima koji bi bili i bolji od Krakena (u svoje vrijeme je bio najbolji alat).
Jedan od njih je Clark, koji će biti predstavljen u sljedećem odjeljku.

2.2. Clark

Ideja rada svih alata za analizu metagenoma je,dobivanjem ulaznog podatka, odrediti vrstu, rod itd. dobivenog podatka. Iako se ovaj problem u posljednje vrijeme puno istražuje, upravo zbog toga što na tržište izlaze alati koji mogu u jako kratkom vremenu proizvesti jako puno ulaznih podataka za navedene alate, sama ideja rješavanja brzine i težine izvođenja još nije u potpunosti rješena. Gore navedeni alat, Kraken, je do pojavljivanja Clark-a bio *najbolji* na tržištu, uspoređujući svoju preciznost i osjetljivost s najboljim alatima u tom polju, s jednom iznimkom. Njegova brzina je bila za nekoliko redaka veličine veća od spomenutih.
Kroz svoj rad *CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers* autori su iznjeli tezu i eksperimentalne dokaze da su uspjeli proizvesti alati koji se može mjeriti s, do tada najboljim alatima (NBC i Kraken), te biti nekoliko puta brži od Kraken-a.
Iako je brzina izvođenja veliki problem, ona, nažalost, nije jedini. Količina memorije, bilo da se misli na radnu memoriju ili memoriju uređaja, predstavlja prepreku koju je teško svladati. U tom području Clark također radi razliku, te je, kao i Kraken, razvio verziju koja bi se trebala moći koristiti na većini osobnih računala. Predstavljena verzija se naziva Clark-l i ne zaostaje puno u preciznosti i osjetljivosti, točnije F1 score-u, od *prave* verzije.
Zanimljivost Clark-a i njegov uspijeh se kriju u načinu izvođenja. Naime, on ima dvije faze.

Prva faza nazvana eng.*pre-processing* koja je zadužena za kreiranje specifičnih ili diskriminativnih k-mera. Prvotno se kreira k-spektar, 4^k dimenzionalni vektor koji sadrži svako pojavljivanje određenog k-mera. K-spektar je *labava* reprezentacija k-mera, te se, uzimajući to u obzir, može međusobno uspoređivati. Kreiranjem indeksa k-mera izbacuju se svi oni k-meri koji su zajednički više podataka. Stoga nastaju specifični ili diskriminativni k-meri.
Druga faza je eng.*post-processing* koji vrši samo usporedbu diskriminativnih k-mera kroz bazu podataka i pridaje svakom podatku određenu vrijednost. Ta vrijednost se još naziva i pouzdanost(eng.*confidence score*). Ako je pouzdanost visoka (blizu 1.0) možemo pretpostaviti da je podatak točno klasificiran / nije klasificiran.
Na alatu Clark postoje dva načina rada. Puni način (eng.*full*) i podrazumjevani način (eng.*default*). Puni način prati pouzdanost svakog podatka i upravo zbog toga ima veću preciznost , ali je, isto tako, sporiji od podrazumjevanog načina.Ounit R (2015)
Rezultati izvođenja i usporedba Clark-a se može vidjeti kroz sljedeću sliku.
Statistička analiza se vodila naspram NCB i Kraken-a. Kao točna klasifikikacija se gledao rod, a ako je neki od alata klasificirao neku višu granu, tada se to u njihovom radu smatralo neklasificiranim podatkom. Prvotno se pokušao naći optimalni **k** za oba navedena alata, te tada usporediti Clark sa njima. Radile su se metrike preciznosti i osjetljivosti, a i mjerila se brzina izvođenja.
Postoje neke značajke po kojima su autori izdvojili Clark kao trenutno najperspektivniji alat.

- Može klasificirati *kratka očitanja* s jako velikom točnošću, te se klasificiranje stvarnih uzoraka poklapa s literaturom objavljenom o njima
- Postiže istu ili bolju točnost od *state-of-the-art* alata
- Brži je 5 puta od Krakena, izravnog konkurenta, te bolje iskorištava višedretvenost
- Može kao izlaz davati pouzdanosti, te je, također, pristupačniji korisniku i samodostatan
- Izvršava se s relativno malom potrošnjom radne memorije, te mu nije potreban veliki prostor na disku

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.

7. Sažetak

Sažetak.