

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 5133

Algoritam staničenja velike baze podataka genoma

Mateo Stjepanović

Zagreb, lipanj 2017.

Umjesto ove stranice umetnite izvornik Vašeg rada.
Da bi ste uklonili ovu stranicu obrišite naredbu \izvornik.

Zahvaljujem mentorici doc. dr. sc. Mirjana Domazet-Lošo na podršci.

SADRŽAJ

1. Uvod	1
2. Definiranje problema	3
3. Strukture podataka i algoritmi u Kraken-u	4
3.0.1. Baza podataka	4
3.0.2. Algoritam klasifikacije podataka	7
3.0.3. LCA(<i>Lower Common Ancestor</i>)	8
4. Algoritmi indeksiranja i pretrage	9
4.0.1. Minimizer - algoritam za reduciranje podataka	9
4.0.2. Binarno pretraživanje	9
5. Pseudokod i razrada algoritma	10
6. Analiza učinkovitosti rješenja	11
7. Zaključak	12
Literatura	13

1. Uvod

Bioinformatika je interdisciplinarna znanost koja se bavi razvojem programa i metoda za interpretiranje bioloških podataka. Spaja matematiku, statistiku, računalnu znanost te druge prirodno matematičke znanosti. U posljednjem desetljeću bioinformatika kao znanost doživljava veliki porast, te kao takva uspjeva mapirati genome mnogih živih bića.

U navedenom porastu najviše se ističu metode analiziranja genoma, te, spajanjem statističkih analiza i algoritama iz područja računalne znanosti, metode određivanja položaja istih u taksonomskom stablu. Upravo u tom području postoje mnogi alati koji su razvijeni upravo s ciljem poboljšanja točnosti određivanja taksonomskog stabla novoj, nepoznatoj, jedinki.

Za većinu jedinki u uzorku su vrsta, rod i više razine stabla nepoznati. Ako je jedinka potpuno nepoznata tada će se klasificirati kao nova nepoznata vrsta te se neće dalje klasificirati. U velikom broju slučajeva će postojati neke sličnosti sa nekim već određenim vrstama, te je potrebno pronaći te sličnosti kako bi se sjedinka uspješno klasificirala. Za to služe razni algoritmi poravnanja. Jedan od tih je BLAST. Postoje druge metode i programi koji pospješuju učinkovitost BLAST-a uvođenjem metoda strojnog učenja. Iako bolje preciznosti ti programi imaju jednu veliku manu. Vrijeme potrebno za klasificiranje podatka je jako veliko, stoga su više manje neiskoristivi.

Tu na scenu nastupa alat za klasifikaciju metagenoma nazvan Kraken (Derrick E. Wood i Steven L.).

*Kraken is ultrafast and high accurate program for assigning taxonomic labels to metagenomic DNA sequences.*¹

Za razliku od alata koji su pokušali poboljšati preciznost BLAST algoritma, te time izgubili na brzini, Kraken je jedan od rijetkih alata koji postiže točnost koja premašuje onu u BLAST algoritmu, s time da ne gubi na brzini izvođenja. Rad Krakena se sastoji od toga da se ulazni podatak "razbija" na k-mere, te se minimizer algoritmom traže oni

¹Wood and Salzberg: **Kraken: ultrafast metagenomic sequence classification using exact alignments.** *Genome Biology* 2014 15:R46

k-meri koji imaju istu vrijednost minimzer-a. Tada se taj podatak uspoređuje s podacima u Kraken-ovoj bazi podataka, s tim da je baza sortirana na način da su podaci sa istim minimizerom smješteni jedan pored drugog, tako da se pretraživanje jako pospješuje i ubrzava. Autori alata su shvatili da Kraken može naići na problem prilikom izvođenja na računalu s ograničenim resursima, točnije na računalu s RAM-om ispod 70GB. Iz tog razloga je razvijena MiniKraken baza podataka koja je sa prvotnih 70GB podataka smanjena na 4GB.

Classifier	HiSeq		MiSeq		simBA-5	
	Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity
Megablast	99.03	79.00	92.44	75.76	96.93	93.67
NBC	82.33	82.33	77.78	77.78	97.64	97.64
PhymmBL	79.14	79.14	76.21	76.21	96.11	96.11
PhymmBL65	99.13	73.95	92.47	73.03	99.08	95.45
Kraken	99.20	77.15	94.71	73.46	99.90	91.25
Kraken-Q	99.12	76.31	94.69	70.41	99.92	89.54
MiniKraken	99.44	66.12	97.41	67.95	99.95	65.87
MiniKraken-Q	99.36	65.67	97.32	65.84	99.98	65.31
Kraken-GB	99.51	93.75	98.48	86.23	99.48	91.13

Tablica 1.1: Klasifikacija roda za tri metagenoma

2. Definiranje problema

Uz već spomenutost rješenja problema korištenja velike količine resursa, nailazi se na problem koji će se pokušati riješiti tijekom ovog rada.

Classifier	HiSeq		MiSeq		simBA-5	
	Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity
Kraken	99.20	77.15	94.71	73.46	99.90	91.25
MiniKraken	99.44	66.12	97.41	67.95	99.95	65.87

Tablica 2.1: Isječak iz Tablice 1.1

Iako se iz dane tablice vidi da se preciznost korištenjem MiniKraken baze podataka neznatno povećala u odnosu na Kraken bazu podataka. S druge strane osjetljivost jako opada. Osjetljivost predstavlja mjeru koja određuje postotak uočenih lažno negativnih podataka, tj. predstavlja postotak točno kvalificiranih podataka koji su uistinu tako i predstavljeni. S druge strane preciznost predstavlja izbjegavanje lažno pozitivnih podataka, tj. koliko podataka je uočeno da ne pripadaju nekoj skupini, s tim da oni uistinu ne pripadaju toj skupini.

Ideja ovog rada je da se algoritmima sličnim algoritmima straničenja osposobi računalo s ograničenim resursima za rad s bazom podataka čija veličina uvelike nadilazi količinu RAM-a raspoloživog na računalu. Prvotno je potrebno podijeliti bazu podataka na više manjih koje će se moći uspješno učitati u memoriju te pretražiti. Ideja je iskoristiti "kontejnere" k-mera koji su određeni svaki jedinstvenom minimizer vrijednosti, te pomoću binarnog pretraživanja indeksa učitavati samo onu bazu podataka pomoću koje je moguće klasificirati podatak u taksonomsko stablo. Želeći se približiti straničenju, predstaviti će se samo iteriranje po indeksima te učitavati slijedno baze podataka i pretraživati svaku.

3. Strukture podataka i algoritmi u Kraken-u

3.0.1. Baza podataka

Kreiranje baze podataka

Nastanak baze podataka se događa u nekoliko koraka. Za početak se sa NCBI-a preuzimaju biblioteke koje sadrže trenutno poznate vrste. Budući da pretraživanje baze podataka u Kraken-u osniva na k-merima potrebno je proizvesti k-mere za svaki podatak u bazi podataka. Za to služi alat naziva Jellyfish(Guillaume Marçais i Carl Kingsford) koji računa k-mere zadane duljine 31 (korisnik također može zadati duljinu k-mera).

Jellyfish je alat za brzo i memorijski povoljno prebrojavanje podnizova u danom nizu. Za potrebe prilagodbe svim resursima za Jellyfish postoje opcije za brzo izračunavanje koje je memorijski zahtjevnije, te ono sporije ali memorijski ne toliko zahtjevno. U kontekstu bioinformatike Jellyfish se koristi za prebrojavanje k-mera u zadanom metagenomu. Sam rad se bazira na hash tablici koja je osposobljena za paralelan rad preko više CPU-a. Hash tablica se sastoji od para (ključ,vrijednost), gdje je ključ zadani k-mer a vrijednost je broj njegovog ponavljanja u metagenomu.

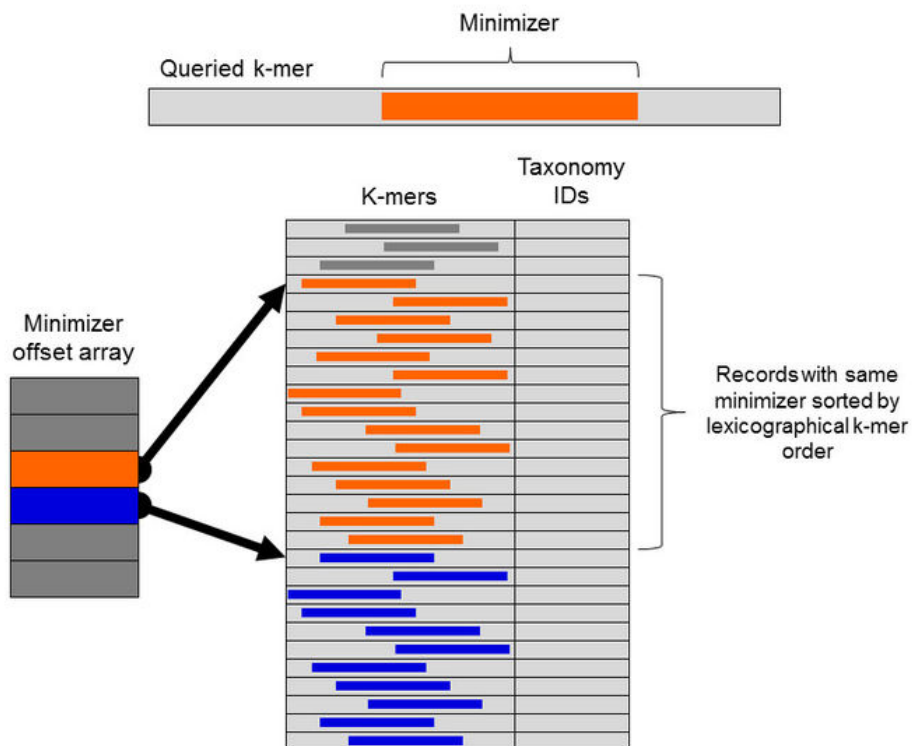
Tablica 3.1: Prikaz niza te njegovih k-mera ($k = 4$)

Ulaz:	AGATCGAGTG
4-mer:	AGAT GATC ATCG TCGA CGAG GAGT AGTG

Nakon rada Jellyfish-a u bazu se spremaju 4 bajtne vrijednosti k-mera te se kreira identifikacijska oznaka za svaki podatak. Tada program kreiranja baze podataka ponovno prolazi kroz istu, te za svaki identifikator podatka izračunava njegov LCA(*Najmanji zajednički predal*), te ga sprema u bazu podataka zajedno s identifikatorom i njegovim k-merima.

Strukture baze podataka

Gore navedeni proces klasificiranja ulaznog genoma pomoću baze podataka se, već navedeno, bazira na pretraživanju k-mera. Baza podataka se sastoji od identifikacijskih oznaka svakog podatka u njoj, njihovog izračunatog najmanjeg zajedničkog pretka, te k-mera koje on tvori.



Slika 3.1: Prikaze strukture baze podataka

S ciljem poboljšanja brzine pretraživanja, uvodi se algoritam indeksiranja i raspoređivanja poznat pod imenom Minimizer (Roberts M, Hayes W, Hunt B, Mount s, te Yorke J.).

```

ptr <- pokazivač_na_k-mer;
k_ct <- brojčak_kontejnera;
vrijednosti <- 1ull « (nt * 2); while brojač je različit od broja kontejnera do
    | brojčak++;
    | dohvati_k-mer(ptr);
    | ključ <- izračunaj_ključ();
    | b_brojac[ključ]++;
end
b_offset_offset[vrijednosti + 1];
for i < vrijednosti do
    | b_offset = b_offset[i-1] + b_brojac[i-1];
end
indeks(b_offset);

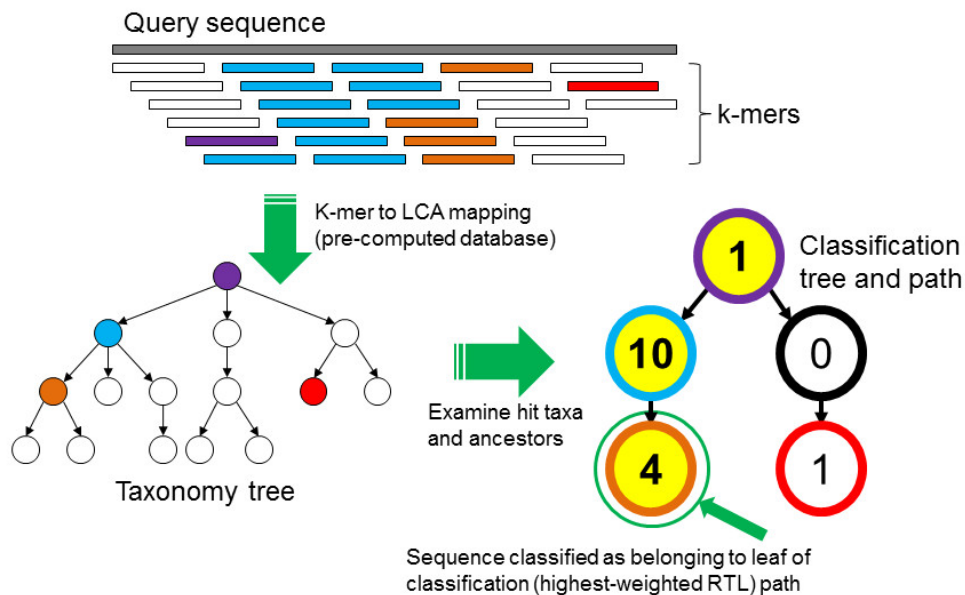
```

Algoritam 1: Računanje indeksa kontejnera

Prilikom sortiranja i indeksiranja se stvaraju svojevrсни kontejneri koji sadržavaju one podatke čiji k-meri imaju iste minimizer vrijednosti. Tom idejom se poboljšava pretraživanje na način da se za svaki ulazni podataka računaju njegovi k-meri te njegove minimizer vrijednosti, te se pretragom liste indeksa dohvaća samo onaj kontejner koji sadrži istu minimizer vrijednost.

3.0.2. Algoritam klasifikacije podataka

Za klasifikaciju ulaznog podatka prvo se određuju i mapiraju svi k-meri samog podatka. Tada se za svaki pojedini k-mer određuje podrijeklo te se, na osnovu LCA(*Lower Common Ancestor*), određuje taksonomsko stablo svakog k-mera. Budući da se radi o jako velikim podacima, te se stvara veliki broj k-mera to taksonomsko stablo vrlo lako može voditi do vrste koja nije nikako povezana s našim podatkom. Zbog toga se za svaki put određuju težine koje tada kompenziraju moguće pogreške. Ako postoji više puteva s istim težinama do listova tada se na njima ponovno radi LCA algoritam.



Slika 3.2: Prikaz klasifikacije podataka

3.0.3. LCA(*Lower Common Ancestor*)

LCA(*Lower common ancestor*) je jedan od osnovnih algoritamskih problema u strukturama stabala. Prvu ideju i definiciju LCA dali su Alfred Aho, John Hopcroft i Jeffrey Ullman 1973. godine u svom radu *On finding lowest common ancestors in trees*. Problem se odnosi na to da se za dva čvora x i y nađe najbliži čvor u stablu koji je hijerarhijski viši, te je zajednički oba čvora. Ovaj problem je važan ne samo zbog svoje složenosti i razumjevanja strukture stabala, nego i zbog svoje primjene. Jedna od primjena je u bioinformatičkoj za određivanje zajedničkog pretka dvije različite vrste. Upravo Kraken koristi LCA kao metodu određivanja navedenog iz k-mera podataka.

4. Algoritmi indeksiranja i pretrage

4.0.1. Minimizer - algoritam za reduciranje podataka

S obzirom na veličinu podataka koji se koriste u metagenomici nailazilo se na problem spremanja istih u RAM. Na ideju rješenja su došli Michael Roberts, Wayne Hayes, Brian R. Hunt, Stephen M. Mount i James A. York u svom radu *Reducing storage requirements for biological sequence comparison*. Način rada se bazira na tome da se radi *seed-and-extend*. Od ulaznog niz se uzimaju određeni dijelovi koji karakteriziraju dani niz, te se poravnava sa karakterističnim podnizovima drugog niza. Kada se uspiju poravnati string se proširuje i traže se druga poravnanja. Ovo uvelika smanjuje bazu podataka koju Kraken koristi, zbog toga što se u kontejnerima koje sadrže podatke istih k-mera nalaze samo vrijednosti podnizova koje se poslije dinamično proširuju.

Nakon poravnanja i sažimanja podataka oni moraju zadovoljavati kriterij kolekcije, tj moraju koliko toliko zadovoljavati normalnu razdiobu u kontejnere. Stoga je potrebno osposobiti taj način rada tako da dva slična niza biraju svoje k-mere na način da će se lako naći poravnanje, te s tim uspjehom usporediti pravilno.

Minimizer algoritam radi tako da se niz sortira leksikografski te se iz toga biraju minimizer vrijednosti. Prvo se izvedu sve moguće k-mere, te se niz poreda leksikografski, tada se od danih k-mera odabiru one najmanje. Ako postoji više najmanjih k-mera, tada su sve one minimizer vrijednosti.

4.0.2. Binarno pretraživanje

5. Pseudokod i razrada algoritma

6. Analiza učinkovitosti rješenja

7. Zaključak

Zaključak.

LITERATURA

Algoritam staničenja velike baze podataka genoma

Sažetak

Sažetak na hrvatskom jeziku.

Ključne riječi: Ključne riječi, odvojene zarezima.

Title

Abstract

Abstract.

Keywords: Keywords.