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 **2**

3. Strukture podataka i algoritmi u Kraken-u **3**

3.0.1. Struktura baze podataka 3

3.0.2. Algoritam klasifikacije podataka 3

3.0.3. LCA(*Lower Common Ancestor*) 4

4. Algoritmi indeksiranja i pretrage **5**

4.0.1. Minimizer - algoritam za reduciranje podataka 5

4.0.2. Binarno pretraživanje 5

5. Pseudokod i razrada algoritma **6**

6. Analiza učinkovitosti rješenja **7**

7. Zaključak **8**

Literatura **9**

1. Uvod

Bioinformatika je područje koje se bavi razvojem programa i metoda za interpretiranje bioloških podataka.Kao interdisciplinarna znanost bioinformatika spaja matematiku, statistiku, računalnu znanost te druge prirodno matematičke znanosti. U posljednih nekoliko godina bionformatika doživljava nagli porast, te se uspjevaju mapirati genomi mnogih živih bića. Između svih dijelova bioinformatike najviše se ističu metode analiziranja genoma te određivanja njihovoh položaja u taksonomskom stablu. Upravo u tom polju postoje mnogi alati koji se u mnogočemu razlikuju.
U tom polju veliki posao su napravili Derrick E. Wood i Steven L. Salzberg sa alatom nazvanim Kraken.
Kraken is ultrafast and high accurate program for assigning taxonomic labels to metagenomic DNA sequences.[1]
Kraken je alat za metagenomsku analizu podataka, koji se ističe svojom brzinom i točnošću. Dotadašnji alati su se mogli podjeliti na brze i točne, te je upravo to bio problem mnogim znanstvenicima. Trebalo je naći neku sredinu te spojiti ta dva uvjeta u jedan alat koji će moći sve to i još mnogo toga. Kraken radi na principu poravnanja k-mera te , slanjem upita u bazu podataka genoma, uspoređivanja te postavljanja genoma u taksonomsku granu. Ta metoda rada daje nevjerovatnu preciznost od preko 90%. Ta preciznost ovisi mnogo o već spomenutoj bazi podataka genoma, koja je jako velika (standardna baza je veličine 70ak GB).

[1] Wood and Salzberg: **Kraken: ultrafast metagenomic sequence classification using exact alignments.** *Genome Biology* 2914 15:R46

2. Definiranje problema

Kraken radi na principu razbijanja ulaznog podatka na preklapajuće k-mere.
K-meri se nazivaju svi mogući podnizova ulaznog niza. U bioinformatici se k-merama nazivaju svi mogući podnizovi baza aminokiselina u nekom uzorku genoma.
Svaki k-mer se tada mapira prema algoritmu LCA(*Lower common ancestor*) prema genomima koji su spremljeni u bazi podataka.

Radi brzine izvođenja ta baza se učitava u RAM računala. Standardna baza podataka koju koristi Kraken veličine 70ak GB , te to predstavlja problem na računalima ograničenih resursa. Kao jedno rješenje tog problema može se generirati vlastita baza, ili koristiti baza podataka nazvana MiniKraken. Iako znatno manja, te time omogućava korištenje spomenutog alata, to rješenje uvelike utječe na kvalitetu očitanja.

Iako se iz dane tablice vidi da se preciznost korištenjem MiniKraken baze povećala u odnosu na Kraken bazu podataka. S druge strane senzitivnost jako opada, što nam predstavlja veći problem.
Ideja ovog rada je da se algoritmima sličnim algoritmima straničenja osposobi svako računalo na rad sa standardnom bazom podataka. Algoritam radi na način da je ulazna baza podataka razdijeli na manje jedinice proizvoljne veličine, te se indeksiranjem istih stvori lista indexa koja će se lako moći pretraživati, te tako učitavati samo ona jedinica koja se po k-merima slaže sa našim ulaznim podatkom. Iako se ovaj radi odnosi isključivo na Kraken, izgradit će se pseudokod koji će se moć implementirati u ovisnosti kako izgleda naša baza podataka.

3. Strukture podataka i algoritmi u Kraken-u

3.0.1. Struktura baze podataka

Baza podataka koju koristi Kraken se iz podataka (genoma) povučenih sa NCBI-ove službene stranice. Budući da Kraken radi na principu uspoređivanja k-mera baza podataka se također sastoji od podataka raspoređenim po k-merima. Nakon povlačenja podataka sa NCBI-a koristi se Jellyfish za raspoređivanje k-mera. Kraken zadano koristi 31-mere, tj za svaki podataka traži sve moguće podskupove do 31 element, u ovom slučaju se to odnosi na jednu nukleotidnu bazu. Za indeksiranje i raspoređivanje podataka koristi se koncept minimizera, kojega su razvili Roberts M, Hayes W, Hunt B, Mount s, te Yorke J. Prilikom sortiranja i indeksiranja se stvaraju svojevrsni kontejneri koji sadržavaju one podatke ciji k-meri imaju iste minimizer-e, tako da prilikom klasificiranja podatka se binarno pretražuje samo određeni kontejner.
Tu dolazimo do ideje rješenja našeg problema. Iskorištavanjem već danih indeksa i kontejnera će se svaki kontejner pretvoriti u posebnu bazu podataka te će se ista sama učitavati u RAM. Stvoritelji Krakena su uzeli u obzir i da je moguć slučaj da veliki broj k-mera ima isti minimizer, te su to riješili tako što se nad svakim kontejnerom izvrši XOR operacija koja odbaci pola bitova te tako su uspjelu približno dobiti normalnu distribuciju podataka. To rješenje će pomoći i u našem problemu, tako da nećemo morati paziti na to da svaki kontejner bude približne veličine te da veličina jednog ne bude prevelika za naš RAM.

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.

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 cvora x i y nađe najbliži čvor u stablu koji je hijerarhijom viši, te je zajedniči 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 bioinformatici 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 poravnat 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 zadovoljavaju kriterij kolekcije, tj moraju koliko toliko zadovoljavati normalnu razdiobu u kontejnere. Stoga je potrebno osposobit taj način rada tako da dva slična niza biraju svoje k-mere na način da će se lako naći poravnanje , te s tim uspjeti 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.