

Asembliranje sekvenci

Seminarski rad u okviru kursa

Uvod u bioinformatiku

Matematički fakultet

Aleksandra Karadžić, Dragutin Ilić

karadzic.matf@gmail.com, dragutin_ilić@yahoo.com

18. maj 2017.

Sažetak

U ovom tekstu je ukratko prikazana osnovna forma seminarskog rada. Obratite pažnju da je pored ove .pdf datoteke, u prilogu i odgovarajuća .tex datoteka, kao i .bib datoteka korišćena za generisanje literature. Na prvoj strani seminarskog rada su naslov, apstrakt i sadržaj, i to sve mora da stane na prvu stranu! Kako bi Vaš seminarski zadovoljio standarde i očekivanja, koristite uputstva i materijale sa predavanja na temu pisanja seminarskih radova. Ovo je samo šablon koji se odnosi na fizički izgled seminarskog rada (šablon koji *morate* da ispoštujete!) kao i par tehničkih pomoćnih uputstava. Molim Vas da kada budete predavali seminarski rad, imenujete datoteke tako da sadrže temu seminarskog rada, kao i imena i prezimena članova grupe (ili samo temu i prezimena, ukoliko je sa imenima predugačko). Predaja seminarskih radova biće isključivo preko web forme, a NE slanjem mejla.

Sadržaj

1	Uvod	3
2	Opis algoritma	4
2.1	Konstrukcija niza super-reč	5
2.2	Izračunavanje preklapanja koristeći niz super-reč	6
2.3	Ubrzanje paralelizacijom	6
2.4	Konstrukcija grafa preklapanja	6
2.5	Generisanje fajla konsenzus sekvenci kontiga i ocene kvaliteta	7
3	Prvi naslov	7
4	Drugi naslov	7
5	n-ti naslov	7
5.1	... podnaslov	7
5.2	... podnaslov	7
6	Poslednji naslov	7
7	Zaključak	8

Literatura	8
A Dodatak	8

1 Uvod

Sekvenciranje genoma, odnosno određivanje rasporeda nukleotida (nt) u genomu, je jedan od fundamentalnih zadataka u bioinformatici. Dužina genoma varira u zavisnosti od organizma (dužina ljudskog genoma je oko 3 milijarde nt, a jedan od najdužih genoma pripada amorfnom jednoćelijskom organizmu *Ameoba dubia* koja je oko 200 puta duži). Najveća prepreka je zapravo činjenica da još uvek nisu razvijene tehnologije koje omogućavaju čitanje nukleotida u genomu od početka do kraja (slično kao čitanje stranica knjige sleva udesno). Trenutno zastupljeno rešenje je sekvenciranje manjih fragmenata DNK koji se nazivaju **ridovi** (*eng. reads*). Uzima se mali uzorak tkiva ili krvi koji sadrži milione kopija DNK. Biohemijskim procesima se DNK razbija na fragmente, čijim sekvenciranje se dobijaju ridovi. Ne zna se iz kog dela genoma je dobijen određeni rid, pa se koristi tehnika preklapanja ridova da bi se rekonstruisao genom. Ovaj ceo proces se naziva i **asembliranje genoma** (*eng. genome assembly*).

Prvo sekvenciranje genoma je odrađeno 1977. godine od strane Frederika Sanger (eng. Frederick Sanger). U Sangerovom metodi dužina ridova je bila između 500 i 1000 ridova. Većina programa zasnovana na ovoj tehnici (*eng. de novo assembly programs*) se bazira na strategiji "preklapanje-raspored-konsenzus" (*eng. overlap-layout-consensus strategy*), u kojoj se preklapanja između ridova izračunavaju brzim tehnikama upoređivanja, raspored kontiga (*eng. contigs*) se generiše pomoću preklapanja u opadajućem rasporedu kvaliteta, a konsenzus sekvenci kontiga se dobija brzim metodama višestrukog poravnavanja. Glavni razlog uspešnosti ovih programa jeste ta da je veličina ridova dovoljna za ustanovljavanje razlika između pravih i lažnih preklapanja.

Prilikom sekvenciranja genoma nailazimo na otežavajuće situacije. Prvo, DNK se sastoji od dve niti, pa ne možemo znati iz koje od njih je rid izveden (nemamo informaciju da li da koristimo dobijen rid ili njegov obrnuti komplement prilikom asebliranja određene niti u genomu). Drugo, tehnologije koje se koriste nisu savršene, pa dobijeni ridovi često sadrže greške (čime se otežava preklapanje ridova). Treće, neki regioni genoma mogu da ostanu nepokriveni ridovima, čime je onemogućena rekonstrukcija celog genoma. Za otklanjanje ovih problema se koriste razne tehnike kao na primer: pristup razbijanja ridova (*eng. read breaking approach*) se koristi za problem nepokrivenosti genoma ridovima, asebliranje kontiga (nepridržanih delova genoma), a ne čitavih hromozoma, se vrši zbog navedenih problema, tehnika uklanjanja mehurova (*eng. bubble removal*) kojom se rešavaju ridovi sa greškom (lažni ridovi) itd.

Sa ulaskom u drugu generaciju (masovno paralelnih) sekvencijalnih tehnologija, počelo se sa proizvodnjom na milijarde kratkih ridova dužine od 50 do 150 nukleotida. Ogromni skupovi podataka vremenski i prostorno usložnjavaju izračunavanje preklapanja dinamičkim algoritmima. Ridovi male dužine otežavaju otkrivanje tačnih od lažnih preklapanja prilikom generisanja rasporeda kontiga. Kako bi se ovo izbeglo, razvijeni su algoritmi za asebliranje zasnovani na detekciji preklapanja kao tačnih pogodataka fiksne dužine. Na osnovu ovakvih preklapanja formira se *de Bruijn*-ov graf, u kome svaki jedinstveni string dužine k (k -mer) koji se pojavljuje u ridu, predstavlja granu koja spaja čvorove označene sa $k-1$ -mer prefiksom i $k-1$ -mer sufiksom. Pomoću ovog grafa i distribucije frekvencije stringova u ridovima razlikujemo stringove koji imaju grešku od onih koji nemaju. Stringovi sa greškama se ili ne koriste ili se prepravljaju i zatim koriste pri asebliranju.

Kako bi koristili sekvenciranje ridova direktno, u ovom radu je predstavljen efikasan metod za izračunavanje preklapanja za kratke ridove. Metod dozvoljava promašaje u preklapanjima, ali ne dodavanja i brisanja, koji se inače i javljaju rede od promašaja. U grafu preklapanja, svaki čvor predstavlja rid, dok grana predstavlja preklapanje. U ovom grafu algoritam pronalazi jedinstvenu putanju ridova koja reprezentuje kontigu. Konsenzus sekvence svake kontige je dobijen izračunavanjem poravnanja višestrukih ridova koji nisu razdvojeni nukleotidima koji oni ne sadrže.

U nastavku će biti predstavljeni detalji algoritma kao i opis programa PCAP.Solexa u kome je algoritam implementiran.

2 Opis algoritma

Na početku ćemo opisati metod koji se koristi za izračunavanje preklapanja između ridova. Za realizaciju ovog metoda je korišćena struktura podataka koju ćemo zvati **“niz super-reč”** (*eng. superword array*), koja je ime dobila na sličan način kao i druga struktura podataka zvana **“sufiksni niz”** (*eng. suffix array*). Reč dužine w je string od w karaktera, a super-reč sa v reči je string sa $v \cdot w$ karaktera dobijena konkatencijom tih reču po redu u kome se javljaju. Reči u super-reči ćemo indeksirati pozicijama od 1 do v . Pozicije u rečima (indeksiranje kreće od jedinice) se dele u dve grupe: čekirane i nečekirane pozicije. Za dve reči dužine w kažemo da se podudaraju ukoliko imaju identične nukleotide na svakoj čekiranoj poziciji tih reči. Na primer ako imamo dve reči dužine 15, **ACCATACCATAGCAC** i **ACTATTCCATAACAC**, i ako pretpostavimo da su nečekirane pozicije 3,6 i 12 tada za ove dve reči možemo reći da se podudaraju. Dužina reči w se često namešta da bude 12 ili veća, kako bi broj čekiranih pozicija u reči bio 12, čime se osigurava da lookup tabela za sve stringove ove dužine (ako je reč duža od 12 brišemo sve nukleotide iz nečekiranih pozicija) može stati u glavnu memoriju. Broj reči v se bira tako da super-reč dužine $v \cdot w$ bude manja od dužine svakog rida (dužinu rida ćemo označavati sa r). Dužina super-reči se još naziva i minimalna dužina preklapanja. Adekvatan primer za ridove dužine 150 bi bio da postavimo da w na 15 i v na 6 čime dobijamo super-reči dužine 90. Rid dužine r ima $r - w \cdot v + 1$ super-reči na pozicijama 1,2, ..., $r - w \cdot v + 1$. U našem primeru bi rid dužine 150 imao 61 super-reč na pozicijama 1,2, ...,61. Za dve super-reči kažemo da su identične ako im se sve reči podudaraju, odnosno svi nukleotidi na čekiranim pozicijama su im identični. Za jednu super-reč kažemo da je manja (veća) od druge ako string nukleotida na svakoj čekiranoj poziciji prve super-reči, u leksikografskom poredku, dolazi pre (posle) stringa nukleotida na čekiranim pozicijama druge super-reči. Svaki rid dobija jedinstveni nenegativni ceo broj koji se zove indeks rida. Pomoću njega i početne pozicije super-reči u ridu se izračunava jedinstveni indeks svake super-reči u svakom ridu. U algoritmu se koristi i funkcija koja pomoću datog indeksa super-reči efikasno računa poziciju te super-reči u indeksu i indeks rida. Sada imamo sve podatke kako bi definisali **“niz super-reč”** strukturu podataka. **“Niz super-reč”** za skup ridova je niz svih indeksa super-reči gde su super-reči sortirane od manje ka većoj.

Na slici 2 prikazano je osam super-reči u sortiranom poredku. Svaka super-reč se sastoji od četiri reči ($v=4$) dužine $w = 15$, gde je svaka čekirana pozicija označena bitom 1, a nečekirana 0. Poslednja pozicija svake reči je označena taraba znakom. Super-reči su grupisane tako

```

11011011111011111011011111011111011011111011111011011110111110111
# # # #
ATNAGCCCAGTTATCCTAGTCAGACTCAGGTTNCATCATTTCNTCCGANCAGACTGACCAG
ATGAGGCCAGTCATCCTTGTGAGACTTAGGTTACATCATTTCATCCGAACAACTGANCAG
ATTAGNCCAGTAATCCTCGTAAGACTAAGGTTACANCATTCTTCCGACCANACTGATCAG
*
ATCAGTCCAGTNATCCTTTTTAGACTTAGGTTGCAACATTTCGTCCGAGCATACTGACCAG
ATGAGACCAGTNATCCTATTGAGACTGAGGTTACATCATTCTTCCGACCAAACACTGAACAG
*
ATGAGACCAGTNATCCTNTTNAGACTCAGGTTCCAACATTGCTCCGANCATACTGAGCAG
ATNAGTCCAGTAATCCTTTTAAGACTTAGGTTGCAGCATTGGTCCGAGCANACTGACCAG
ATCAGNCCAGTNATCCTATTAGACTNAGGTTGCATCATTGATCCGATCACACTGANCAG

```

Slika 1: Prikaz grupisanja i sortiranja super-reči

da se u svakom bloku nalaze identične super-reči (prve tri, sledeće dve i poslednje tri). Vidimo da nukleotidi na nečekanim pozicijama mogu biti ili različiti ili nedefinirani (označeni sa **N**) u istom bloku. Pvi blok dolazi pre drugog zato što na poziciji označenoj sa zvezdicom se u prvom bloku nalazi **G** nukleotid a u drugom **T** (**G** leksikografski dolazi pre **T**). Slično zvezdica između drugog i trećeg bloka označava zašto drugi dolazi pre trećeg.

2.1 Konstrukcija niza super-reč

Niz super-reč se konstruiše pomoću bucket sorta u v krugova sortiranja (podsetimo se da se svaka super-reč sastoji od v reči od kojih svaka ima tačno 12 čekiranih pozicija). Najpre se niz inicijalizuje indeksima super-reči u rastućem poretku po vrednostima indeksa. Zatim se iz svake super-reči uzima reč na poziciji p gde p ide od v do 1 (krećemo se sdesna u levo u svakoj super-reči) i niz sortiramo po leksikografskom poretku tih reči. Kako idemo sdesna u levo nakon v koraka bucket sort garantuje da će “niz super-reč” biti sortiran. Za sortiranje se koriste dve pomoćne strukture podataka, lookup tabela i niz lokacija, koji služe kao kofe (*eng. bucket*) u koje smeštamo reči. Niz lokacija je niz celih brojeva koji je indeksiran od 0 do najvećeg indeksa super-reči. Lookup tabela se inicijalizuje negativnim brojevima što označava praznu kofu, a ima onoliko polja koliko ima kodova za sve stringove dužine 12. Bucket sortiranje se obavlja tako što se elementi “niza super-reč” sleva na desno čitaju, smeštaju u kofe i zatim se kopiraju iz kofa nazad u “niz super-reč” sdesna u levo. Detaljnije, trenutni element, indeks super-reči, se smešta u kofu iz “niza super-reč” na sledeći način. Indeks super-reči se koristi kako bi se našla reč na poziciji p . Zatim tu reč koja je dužine w transformišemo u string dužine 12 korišćenjem bitovskih operacija kako bi uklonili nukleotide na nečekanim pozicijama. Znamo da postoji indeks u lookup tabeli koji predstavlja kod tog stringa dužine 12, pa vrednost lookup tabele na tom indeksu čuvamo u nizu lokacija na poziciji koja ima isti indeks kao i indeks super-reči, a u lookup tabelu upisujemo sam indeks super-reči. Kada smo sve elemente “niza super-reč” ubacili u kofe, čitamo ih počevši od najveće

kofe i kopiramo ih u “niz super-reč” sdesna ulevo.

2.2 Izračunavanje preklapanja koristeći niz super-reč

Nakon konstrukcije “niza super-reč” delimo ga u sekcije tako da se u svakoj sekciji nalaze identične super-reči. Ove sekcije se obrađuju kako bi se generisala preklapanja između ridova. To radimo na sledeći način. Posmatraćemo trenutnu sekciju koja ima s super-reči. Svaka dva rida koja imaju super-reč u ovoj sekciji mogu da se poklapaju. Tako da imamo $(s * (s - 1))/2$ potencijalnih preklapanja u ovoj sekciji. Kako bi uprostiti problem mi nećemo računati sva ova preklapanja, već ćemo računati njegov podskup (koji ćemo dobiti za linearno vreme), a ostala preklapanja će moći da budu zaključena iz ovog podskupa. Ovo se može uraditi, tako što sortiramo super-reči u sekciji po dužini njihovih prefiksa u ridu i selektovanjem susednih parova super-reči kako bi izračunali preklapanja. Ako super-reči pripadaju različitim ridovima (podsetimo se da ovo možemo proveriti koristeći funkciju koja od indeksa super-reči računa poziciju super-reči u ridu i indeks rida kome ta super-reč pripada) onda računamo preklapanje između njih. Ukoliko je broj nukleotida koji se razlikuju u preklapanju ograničen odsecanjem, preklapanje čuvamo.

2.3 Ubrzanje paralelizacijom

Veliki skupovi ridova se dele u manje podskupove (obeležениh brojevima od 0 do $j-1$). Svaki rid iz istog podskupa ima super-reč kod koje ostatak pri deljenju koda reči na poziciji p (p je između 1 i v) sa j daje oznaku tog podskupa. Naglasimo da rid može pripadati u više podskupova. Ovi podskupovi se dodeljuju j procesorima kako bi se paralelno izračunala preklapanja. Svaki procesor računa preklapanja u podskupu konstruišući niz super-reč za svaki podskup. Ovim postupkom dobijamo različite fajlove preklapanja za svaki podskup. Ovi fajlovi se potom skeniraju kako bi se uklonila redundantna preklapanja. Na taj način dobijamo skup fajlova neredundantnih preklapanja.

2.4 Konstrukcija grafa preklapanja

Fajlovi preklapanja se čitaju redom jedan po jedan i svako preklapanje sa dobrim procentom preklapanja (broj nukleotida koji su se razlikovali u preklapanju je manji od granice) čuvamo u glavnoj memoriji. Graf preklapanja se konstruiše da čvorovi budu ridova, a grane između čvorova su preklapanja između dva rida koju povezuje ta grana. Graf se potom ispituje kako bi se našle dugačke putanje preklapanja neponavljajućih čvorova. Čvor je ponavljajući ako postoje dve dugačke putanje koje se završavaju u čvoru, ali nemaju ponavljanja između njihovih prefiksni putanja. Ovo se izvodi na principu dijekstrinog algoritma. Svaka dugačka putanja neponavljajućih čvorova se predstavlja kao kontiga ridova. Stvaranje rasporeda svake kontige se obavlja na jednom procesoru sa velikom količinom glavne memorije. Na kraju ovog koraka se dobijaju fajlovi koji predstavljaju raspored kontiga.

2.5 Generisanje fajla konsenzus sekvenci kontiga i ocene kvaliteta

Fajlovi rasporeda kontiga se paralelno procesiraju. Ridovi se raspoređuju kako bi formirali višestruka poravnanja ridova. Ova poravnanja se koriste za generisanje konsenzus sekvenci za kontige. Za svaki fajl rasporeda kontiga se generiše fajl konsenzus sekvenci kontiga kao i fajl ocene kvaliteta baza konsenzus kontiga. Na osnovu ovih fajlova se generiše jedan fajl konsenzus sekvenci kontiga i jedan fajl ocene kvaliteta baze konsenzus kontiga.

3 Prvi naslov

Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst.

Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst.

Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst.

4 Drugi naslov

Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst.

Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst.

5 n-ti naslov

Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst.

5.1 ... podnaslov

Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst.

5.2 ... podnaslov

Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst.

6 Poslednji naslov

Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst. Ovde pišem tekst.

7 Zaključak

Ovde pišem zaključak. Ovde pišem zaključak. Ovde pišem zaključak.
Ovde pišem zaključak. Ovde pišem zaključak. Ovde pišem zaključak.
Ovde pišem zaključak. Ovde pišem zaključak. Ovde pišem zaključak.
Ovde pišem zaključak. Ovde pišem zaključak. Ovde pišem zaključak.

A Dodatak

Ovde pišem dodatne stvari, ukoliko za time ima potrebe. Ovde pišem
dodatne stvari, ukoliko za time ima potrebe. Ovde pišem dodatne stvari,
ukoliko za time ima potrebe. Ovde pišem dodatne stvari, ukoliko za time
ima potrebe. Ovde pišem dodatne stvari, ukoliko za time ima potrebe.