Asembliranje sekvenci

Seminarski rad u okviru kursa
Uvod u bioinformatiku
Matematički fakultet

Aleksandra Karadžić, Dragutin Ilić
karadzic.matf@gmail.com, dragutin_ilic@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

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 od kraja (slično kao čitanje knjige). 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 rekontruisao genom. Ovaj ceo proces se naziva i **asembliranje genoma (*eng. genome assembly*)**.

Prvo sekvenciranje genoma je odrađeno 1977. godine od strane Frederika Sangera (eng. Frederick Sanger). U Sangerovom metodu 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 tehinikama 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 odakle je rid izveden (nemamo inforamciju da li da koristimo dobijen rid ili njegov obrnuti komplement prilikom asembliranja određene niti u genomu). Drugo, tehnologije koje se koriste nisu savršene, pa dobijeni ridovi često sadrže greške (ovim je otežano 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, asembliranje kontiga (neprkidnih delova genoma), a ne čitavih hromozoma,tehnika uklanjanja mehurova (*eng. bubble removal*) kojom se rešavaju ridovi sa greškom (lažni ridovi).

Sa ulaskom u drugu genereciju (masovno paralelnih) sekvencijalnih tehnologija, počelo je sa proizvodnjom na milijarde kratkih ridova dužine od 50 do 150 nukleotida. Ogromni skupovi podataka otežavaju izračunavanje preklapanja dinamičkim algoritmima u smislu vremenske i prostorne složenosti. Ridovi male dužine otežavaju otkrivanje tačnih od lažnih preklapanja prilikom generisanja rasporeda kontiga. Kako bi se ovo izbeglo, razvijeni su algoritmi za asembliranje zasnovani na detekciji preklapanja kao tačnih pogodataka fiksne dužine. Na osnovu ovakvih preklapanja formira se *de Bruijin* graf, u kome se svaki jedinstevni 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 korišćeni pri asembliranju.

Kako bi koristili sekvenciranje ridova direktno, u ovom radu je pred-

stavljen efikasan metod za izračunjavanje preklapanja za kratke ridove. Metod dozvoljava promašaje u preklapanjima, ali ne dodavanja i brisanja, koji se inače i javljaju ređe 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 predstavljni detalji algoritma kao i opis programa PCAP.Solexa u kome je algoritam implementiran.

Primer 1.1 *Problem zaustavljanja (eng.* halting problem*) je neodlučiv [?].*

Primer 1.2 *Za prevođenje programa napisanih u programskom jeziku C može se koristiti GCC kompajler [?].*

Primer 1.3 *Da bi se ispitivala ispravost softvera, najpre je potrebno precizno definisati njegovo ponašanje [?].*

Reference koje se koriste u ovom tekstu zadate su u datoteci *seminarski.bib*. Prevođenje u pdf format u Linux okruženju može se uraditi na sledeći način:

```
pdflatex TemaImePrezime.tex
bibtex TemaImePrezime.aux
pdflatex TemaImePrezime.tex
pdflatex TemaImePrezime.tex
```

Prvo latexovanje je neophodno da bi se generisao *.aux* fajl. *bibtex* proizvodi odgovarajući *.bbl* fajl koji se koristi za generisanje literature. Potrebna su dva prolaza (dva puta pdflatex) da bi se reference ubacile u tekst (tj da ne bi ostali znakovi pitanja umesto referenci). Dodavanjem novih referenci potrebno je ponoviti ceo postupak.

Broj naslova i podnaslova je proizvoljan. Neophodni su samo Uvod i Zaključak. Na poglavlja unutar teksta referisati se po potrebi.

Primer 1.4 *U odeljku 3 precizirani su osnovni pojmovi, dok su zaključci dati u odeljku 7.*

Još jednom da napomenem da nema razloga da pišete:

```
\v{s} i \v{c} i \'c ...
```

Možete koristiti srpska slova

```
š i č i ć ...
```

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

2 Slike i tabele

Slike i tabele treba da budu u svom okruženju, sa odgovarajućim naslovima, obeležene labelom da koje omogućava referenciranje.

Primer 2.1 *Ovako se ubacuje slika. Obratiti pažnju da je dodato i*

```
\usepackage{graphicx}
```

Na svaku sliku neophodno je referisati se negde u tekstu. Na primer, na slici 1 prikazane su pande.

Primer 2.2 *I tabele treba da budu u svom okruženju, i na njih je neophodno referisati se u tekstu. Na primer, u tabeli 1 su prikazana različita poravnanja u tabelama.*

Slika 1: Pande

Tabela 1: Razlčita poravnanja u okviru iste tabele ne treba koristiti jer su nepregledna.

centralno poravnanje	levo poravnanje	desno poravnanje
a	b	c
d	e	f

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.

3.1 Prvi podnaslov

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

3.2 Drugi podnaslov

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.

4.1 ... podnaslov

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.

5.2 ... podnaslov

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. 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.

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.