SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Faza konsenzusa u OLC paradigmi sastavljanja genoma

Ivan Paljak, Ivan Sekulić
Voditelj: *Mile Šikić*

Zagreb, siječanj 2017.

SADRŽAJ

1. Algoritam Sparc

Cilj ovog projekta bio je implementirati algoritam Sparc (Ye i Ma, 2016) koji se koristi u konsenzus fazi preklapanje-razmještanje-konsenzus (engl. Overlap-Layout-Consensus, OLC) pristupa. Kod OLC pristupa, traženje puta u grafu svodi se na traženje Hamiltonovog puta. To je put koji prolazi kroz sve vrhove u grafu točno jedanput te bi nam takav put otkrio cijeli slijed. Traženje Hamiltonovog puta je NP potpun problem te je potrebno uvesti heuristike za pojednostavljenja grafa. Sparc je algoritam za konsenzus fazu OLC pristupa, temeljen na *k-mer/de Bruijn* (Hannenhalli et al., 1996) grafovima.

1.1. Ulaz u algoritam

Algoritam koristi izlaz iz faze razmještanja OLC pristupa te sva početna očitanja genoma. Očitanja su mapirana na kontigu iz faze razmještanja i pohranjena u datoteku formata .sam, opisanom u nastavku.

Kako bi uspješno izgradili graf i proveli algoritam, potrebno je znati gdje se pojedina očitanja mapiraju na osnovnu kontigu. Te informacije dobivano iz .sam datoteke koju smo generirali alatom *GraphMap* (Sovi'c et al., 2016). Nama najvažnije informacije u datoteci jesu:

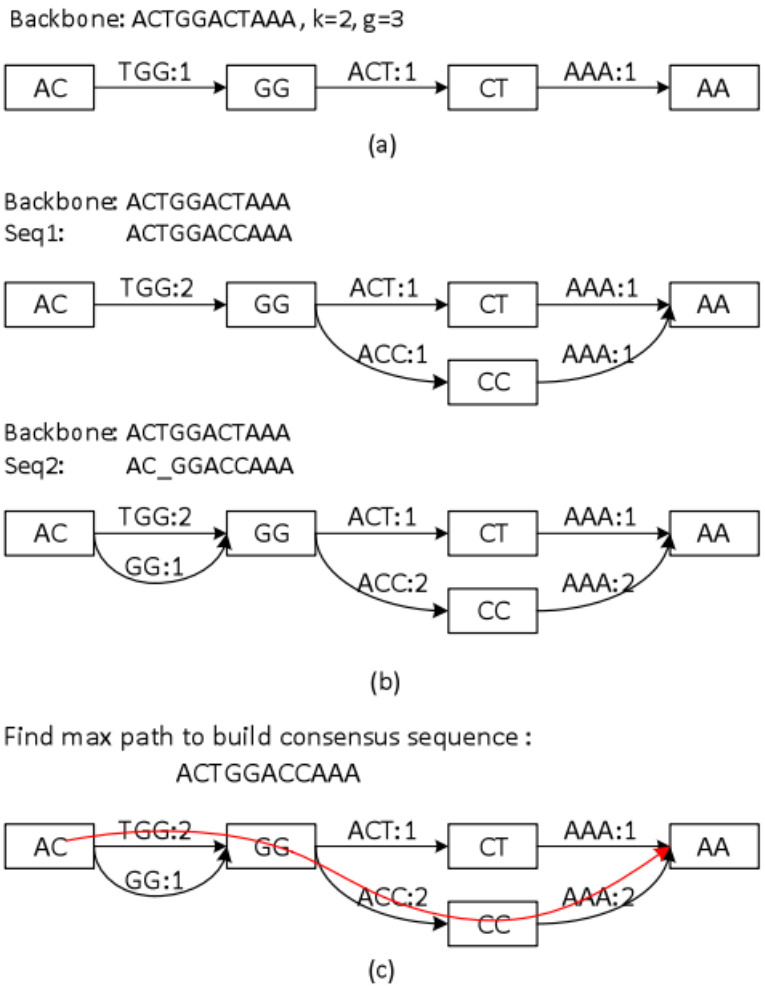
- POS: pozicija na osnovnoj kontizi na kojoj počinje mapiranje pojedinog očitanja
- CIGAR: operacije obavljene nad očitanjem kako bi se dobilo mapiranje (dodavanje, brisanje, pomicanje, itd.)
- SEQ: očitana sekvenca u prvotnom obliku (bez obavljenih CIGAR operacija)
- QUAL: kvaliteta očitanja (iz .fastq datoteke očitanja)

Važnost navedenih informacija bit će jasnija kasnije. Za naše potrebe, spremili smo sve važne informacije u zasebnu datoteku koja se sastoji od osnovne kontige (engl. backbone, layout) u prvoj liniji. U nastavku datoteke su očitanja zapisana kroz tri linije.

Prva linija predstavlja originalno očitanje na koje su primjenjene CIGAR operacije kako bi se dobilo poravnanje s backbone-om. U drugoj liniji nalazi se kvaliteta očitanja (QUAL), a u trećoj je pozicija na backbone-u na kojoj počinje mapiranje izmjenjene sekvence.

1.2. Opis algoritma

Najprije se gradi graf direktno iz backbone-a, kako je prikazano na slici **??**. Linearno se prolazi kroz backbone te se u čvorove stavljaju *k-torke (k-meri)*, a bridovi su definirani kao postojeći, ako su sufiks jednog čvora, a prefiks drugoga. Zatim se prolazi kroz mapirana očitanja te se ažuriraju težine u grafu gdje dolazi do preklapanja. Dodaju se novi čvorovi i bridovi ako se u očitanju javljaju nove k-torke. Konačni korak je pronalazak najtežeg puta u grafu. Primjer izgradnje grafa dan je na slici 1.1. TODO: objasni bolje



Slika 1.1: Izgradnja grafa.

2. Zaključak

U ovom projektu implementirali smo algoritam Sparc.

3. Literatura

Sridhar Hannenhalli, William Feldman, Herbert F Lewis, Steven S Skiena, i Pavel A Pevzner. Positional sequencing by hybridization. *Computer applications in the biosciences: CABIOS*, 12(1):19–24, 1996.

Ivan Sovi'c, Mile Šiki'c, Andreas Wilm, Shannon Nicole Fenlon, Swaine Chen, i Niranjan Nagarajan. Fast and sensitive mapping of nanopore sequencing reads with graphmap. *Nature communications*, 7, 2016.

Chengxi Ye i Zhanshan Sam Ma. Sparc: a sparsity-based consensus algorithm for long erroneous sequencing reads. *PeerJ*, 4:e2016, 2016.

4. Sažetak

Sažetak.