

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DOKUMENTACIJA

Poboljšanje djelomično sastavljenog genoma dugim očitanjima

Bruno Kovač, Tonko Sabolčec, Fabijan Čorak

Voditelj: doc. dr. sc. Krešimir Križanović

Zagreb, siječanj 2020.

SADRŽAJ

1. Uvod	1
2. Postupak	2
2.1. Izgradnja grafa	2
2.2. Obilazak grafa	2
2.3. Obrada putova	3
2.4. Primjer (?)	3
3. Rezultati	4
4. Zaključak	5
5. Literatura	6
6. Sažetak	7

1. Uvod

Sekvenciranje genoma svodi se na kombiniranje očitavanja u jednu cjelinu. Ovaj rad pretpostavlja da su očitavanja već sastavljena, ali djelomično - u fragmente. Jedan takav fragment naziva se *contig*. Dakle, zadatak se svodi na što bolje povezivanje *contiga*, što smo učinili postupkom opisanim u [1], koji se oslanja na duga očitavanja (?). Taj rad definira nekoliko mjera preklopljenosti očitavanja koje kombiniraju duljinu područja *overlap* (*OL*), *overhang* (*OH*) i *extension* (*EL*). Mjere su ovdje definirane za dva očitavanja S_1 i S_2 ; pripadnost područja određena je indeksom.



Slika 1.1: Preklop dvaju očitavanja s naznačenim područjima

- *sequence identity* (*SI*) - omjer ukupnog broja podudarajućih znakova u *overlap* područjima i duljine duljeg od tih dvaju područja

$$SI = \frac{\text{broj_podudaranja}}{\max(OL_1, OL_2)}$$

- *overlap score* (*OS*)

$$OS = (OL_1 + OL_2) \frac{SI}{2}$$

- *extension score* (*ES*) - uz S_2 kao produžetak od S_1

$$ES_2 = OS + \frac{EL_2 - OH_1 - OH_2}{2}$$

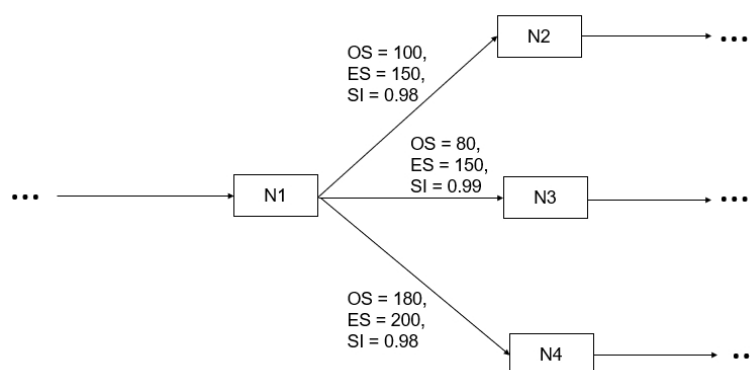
2. Postupak

Sastavljanje očitavanja u niz apstrahirano je izgradnjom i obilaskom grafa.

2.1. Izgradnja grafa

Svaki *contig* i svako očitavanje čine jedan čvor grafa. Čvor koji predstavlja *contig* zovemo *anchor*. Brid postoji između svaka dva čvora čiji je *SI* veći od nekog minimuma (u radu je uzeta vrijednost 0.97). Pritom svaki brid nosi informacije o preklopljenosti čvorova koje povezuje (*SI*, *OS*, *ES*). Te mjere računaju se na temelju informacija o preklopljenosti dobivenih korištenjem alata *minimap2* opisanog u [2].

2.2. Obilazak grafa



Slika 2.1: Odabir sljedećeg čvora u obilasku

Kroz graf se traže putovi čije su krajnje točke *anchor* čvorovi. Za to se koriste tri načina obilaska. Pritom se obišeni čvorovi pamte i postavlja se maksimalna dubina do koje se pretražuje.

1. Iz *anchor* čvora pretraga se nastavlja u sve susjedne čvorove. Iz svakog sljedećeg čvora, pretraga se nastavlja u onaj susjed s kojim je najveći *overlap score*, a

kojim se u konačnici dolazi do *anchor* čvora. Ako je *overlap score* jednak, gleda se *sequence identity*. Ako je pak i ta mjera jednaka, gleda se duljina očitavanja.

2. Kao i prethodni način, ali umjesto mjere *overlap score* gleda se *extension score*.
3. U svakom čvoru susjed se odabire probabilistički - s vjerojatnošću odabira proporcionalnom mjeri *extension score*, sve dok se ne dosegne *anchor*. Postupak se pokreće iz svakog *anchor* čvora proizvoljan broj puta. Ovo je tzv. Monte Carlo metoda.

2.3. Obrada putova

2.4. Primjer (?)

slika opis

3. Rezultati

Implementacija je ispitana na uzorcima ..., ... i ... na jednoj dretvi uz procesorsku moć od ... GHz.

uzorak / organizam (?)	vrijeme (s)	memorija (GiB)
E. Coli	X	Y
A	B	C

4. Zaključak

Zaključak.

5. Literatura

- [1] Huilong Du i Chengzhi Liang. Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. *bioRxiv*, 2018. doi: 10.1101/345983. URL <https://www.biorxiv.org/content/early/2018/06/13/345983>.
- [2] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 05 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty191. URL <https://doi.org/10.1093/bioinformatics/bty191>.

6. Sažetak

Sažetak.