

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Poboljšanje djelomično sastavljenog genoma dugim očitanjima

Dan Ambrošić
Stjepan Dugonjić
Mihaela Bošnjak

Zagreb, siječanj 2019.

SADRŽAJ

1. Opis problema i podataka	1
2. Opis implementacije	2
2.1. Graf preklapanja	3
2.2. Stevo	3
2.3. Mihaela	3
3. Rezultati	4
4. Zaključak	5
Literatura	6

1. Opis problema i podataka

Cilj ovog projekta je poboljšati djelomično sastavljen genom koristeći duga očitavanja. Ukoliko genom ima puno ponavljajućih sekvenci, pogotovo ako su iste duže od duljine očitavanja, teško ga je sastaviti u potpunosti. Ulazni podaci su skup sastavljenih sekvenci (contig-a), koje su dobivene nekim od alata za sastavljanje genoma, te skup dugih očitavanja. Zadatak je napisati program koji pokušava sastaviti contig-e u jednu sekvencu koristeći dobivena očitavanja.

Za provjeru rada implementacije korištena su 3 genoma:

1. EColi - sintetski podaci
2. CJejuni - stvarni podaci
3. BGrahamii - stvarni podaci

Svi podaci se sastoje od datoteke s očitanjima te datoteke s već sastavljenim sekvencama koje su u FASTA formatu. Uz njih algoritmu su potrebna i preklapanja između contig-a i očitavanja, te međusobna preklapanja očitavanja koja su dobivena alatom Minimap2 ¹ u PAF formatu. Konačno, dobivena je i datoteka koja sadrži referentnu sekvencu s kojom se uspoređuje krajnji rezultat.

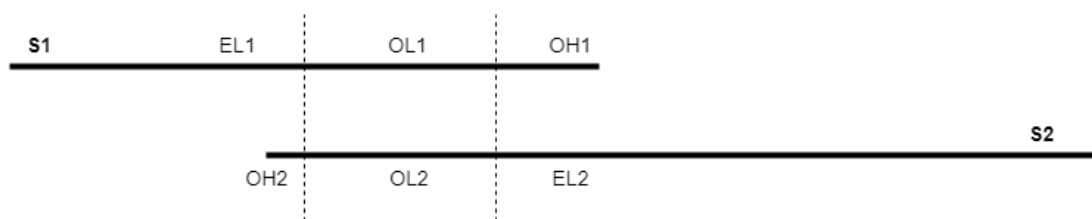
¹<https://github.com/lh3/minimap2>

2. Opis implementacije

Implementacija uglavnom slijedi algoritam HERA (engl. *Highly Efficient Repeat Assembly*) opisan u radu [?] uz nekoliko manjih modifikacija. Algoritam se sastoji od tri glavnih koraka. Prvo se gradi graf preklapanja u kojem se traže putevi između contig-a. Nakon što su pronađeni mogući putevi, za svaki par contig-a se traži reprezentativna sekvenca te se u konačnici sastavlja jedna sekvenca između povezanih contig-a.

Alat Minimap2, ukoliko pronađe preklapanje između dvije sekvence (očitanje ili contig), da informacije o indeksima početka i kraja preklapanja za obje sekvence. Prvi korak je odbacivanje preklapanja koja zadovoljavaju barem jedan od sljedećih uvjeta:

- preklapanje je između dvije iste sekvence,
- preklapanje u kojem jedna sekvenca u potpunosti sadrži drugu,
- SI (engl. *sequence identity*) mjera je ispod određene granice (primjerice 40%).



Slika 2.1: Primjer preklapanja između S1 i S2

Za preostale sekvence se računaju mjere preklapanja i produživanja. Na slici 2.1 je prikazano preklapanje između sekvenca S_1 i S_2 gdje se sekvenca S_2 nalazi nakon sekvenca S_1 . U sredini između isprekidanih linija je regija preklapanja čija je duljina OL_1 i OL_2 , ovisno koju sekvencu gledamo, a izvana su OH (engl. *overhang length*) i EL (engl. *extension length*). Koristeći navedene duljine, računa se mjera preklapanja OS i mjere produživanja ES_1 i ES_2 prema izrazima 2.1 -

2.3. Pri tome je potrebno voditi računa preklapaju li se sekvence na istim lancima ili na suprotnim. Primjerice da se S_1 i S_2 preklapaju na suprotnim lancima, potrebno bi bilo zamijeniti vrijednosti OH_2 i EL_2 .

$$OS = \frac{(OL_1 + OL_2) * SI}{2} \quad (2.1)$$

$$ES_1 = OS + \frac{EL_1}{2} - \frac{OH_1 + OH_2}{2} \quad (2.2)$$

$$ES_2 = OS + \frac{EL_2}{2} - \frac{OH_1 + OH_2}{2} \quad (2.3)$$

2.1. Graf preklapanja

Nakon što su sve mjere izračunate može se izgraditi graf preklapanja. Čvorovi u grafu predstavljaju contig-e i očitavanja, a grane preklapanja, s time da grana može biti povezana na glavu ili rep čvora. Drugim riječima, svaki čvor ima svoje prefikse i sufikse. Sljedeći korak je traženje puteva između dva contig-a koji se obavlja pretraživanjem u dubinu (engl. *DFS*) uz nekoliko pravila. Postupak kreće iz svakog contig-a i zaustavlja se kada dođe do očitavanja koje je povezano s drugim contig-om ili je trenutni put veći od maksimalne dopuštene duljine. Dodatno ograničenje je da se jedno očitavanje ne može više puta pojaviti u istom putu.

Za izgradnju puteva koriste se tri pristupa. Prvi pristup iz početnog contiga izabere sve sufikse, ali za svako sljedeće produživanje bira ono koje ima najveću mjeru preklapanja OS . Ukoliko se dođe do čvora koji nema produživanja, vraća se jedan korak nazad i bira sljedeće najbolje produživanje. Drugi pristup radi isto kao prvi jedino koristi mjeru produživanja ES . Konačno, treći pristup nasumično bira produživanja proporcionalno mjeri produživanja ES , s time da je potrebno odrediti koliko puta će se iz svakog contig-a pokušati pronaći put ovom metodom. Po završetku postupka pronalaženja puteva između contig-a potrebno je samo odbaciti duplikate i moguće je prijeći na sljedeći korak algoritma.

2.2. Stevo

2.3. Mihaela

3. Rezultati

4. Zaključak

LITERATURA

Poboljšanje djelomično sastavljenog genoma dugim očitanjima

Sažetak

Blah blah blah

Ključne riječi: klucne rijeci