SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Poboljšanje djelomično sastavljenog genoma dugim očitanjima

Lana Tuković, Ema Vlainić
Voditelj: *Krešimir Križanović*

Zagreb, svibanj 2023.

SADRŽAJ

1. Uvod

Postupak sastavljanja složenih genoma može dati fragmentiran rezultat zbog velikog broja ponavljajućih sekvenci koje otežavaju proces poravnanja. Naš zadatak bio je pokušati međusobno povezati dobivene fragmente - contige u cijeli genom. Algoritam koji smo pri tome koristili zasniva se na konstruiranju grafa preklapanja te pronalaženja optimalnih staza među preklapanjima. Pronađena optimalna staza će nam služiti da povežemo dva contig-a.

2. Opis algoritma

Skupovi očitanja i već sastavljenih contig-a pripremljeni su kao testni podaci. Njihova preklapanja dobivena pomoću alata Minimap2 koristimo kao ulazne točke programa.

Ulazni podaci:

- preklapanja između contig-a i očitanja u PAF formatu dobivena korištenjem alata Minimap2 Li (2023) nad datotekom sa skupom contig-a i datotekom sa skupom očitanja
- međusobna preklapanja očitanja u PAF formatu dobivena korištenjem alata Minimap2 nad dvije iste datoteke sa skupom očitanja

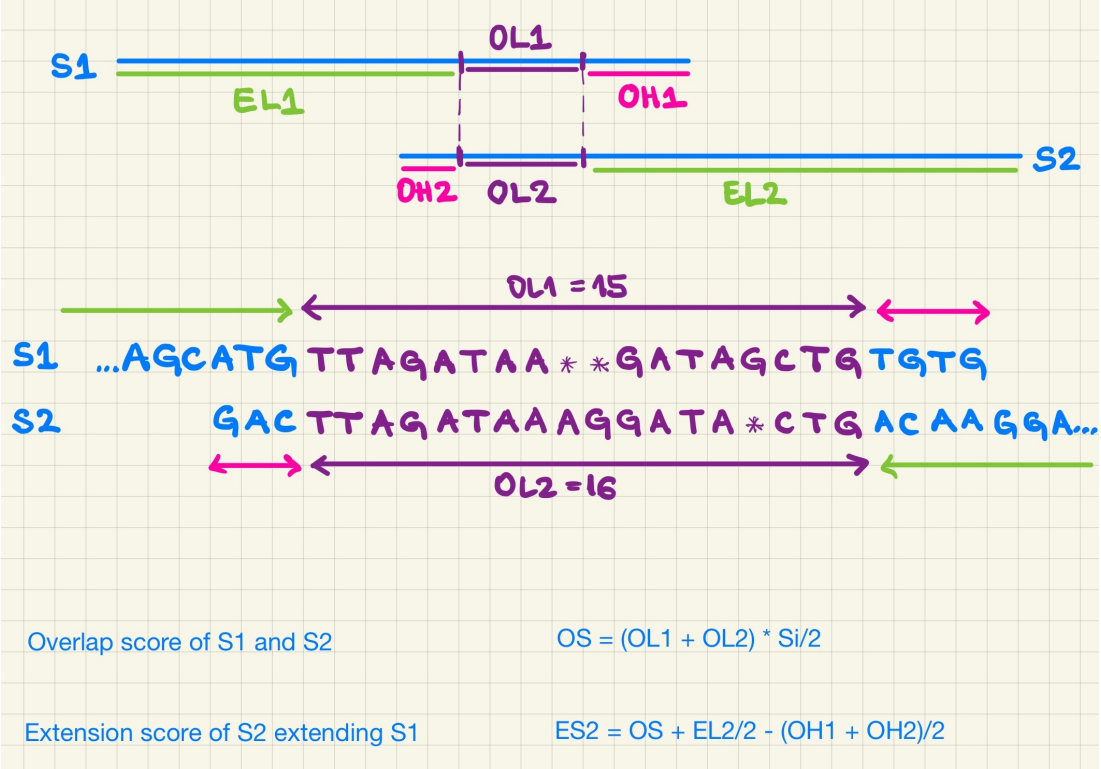
Izlazni podaci:

- poboljšani skup sastavljenih contiga u FASTA formatu

Sada kada imamo sve potrebne podatke možemo konstruirati graf preklapanja. U svom radu Du i Liang (2021) opisali su postupak kontruiranja grafa preklapanja gdje preklapanja povezuju čvorove grafa. Graf preklapanja sastoji se od dvije vrste čvorova: usidreni čvorovi (*anchoring nodes*) koji predstavljaju unaprijed sastavljane contig-e i čvorovi očitanja (*read nodes*). Veza između dva čvorova predstavlja preklapanje tih dvaju čvorova. Kada je graf konstruiran, slijedi traženje optimalnih staza iz svakog usidrenog čvora do skupa mogućih završnih usidrenih čvorova. Postupak traženja staza provodi se tako da se iz svakog čvora širi staza prema susjednom čvoru sa najvećim dobitkom. Optimalne staze u grafu tražili smo na tri načina:

- staza se proširuje na susjedni čvor koji ima najveći *overlap score* - vrijednost preklapanja
- staza se proširuje na susjedni čvor koji ima najveći *extension score* - vrijednost nadovezivanja
- staza se proširuje na susjedni čvor koji je slučajno izabran, a vjerovatnost njegovog izabiranja je proporcionalna njegovoj vrijednosti nadovezivanja - *Monte Carlo* metoda

Na slici 2.1 prikazano je kako se računa *overlap score* i *extension score*.



Slika 2.1: Postupak računanja vrijednosti preklapanja i nadovezivanja

3. Zaključak

Zaključak.

4. Literatura

Huilong Du i Chengzhi Liang. Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. *bioRxiv*, 345983, 2021. doi: 10.1101/345983. URL https://doi.org/10.1101/345983.

Heng Li. Minimap2: Alignment for versatile sequence analysis, 2023. URL https://github.com/lh3/minimap2.