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

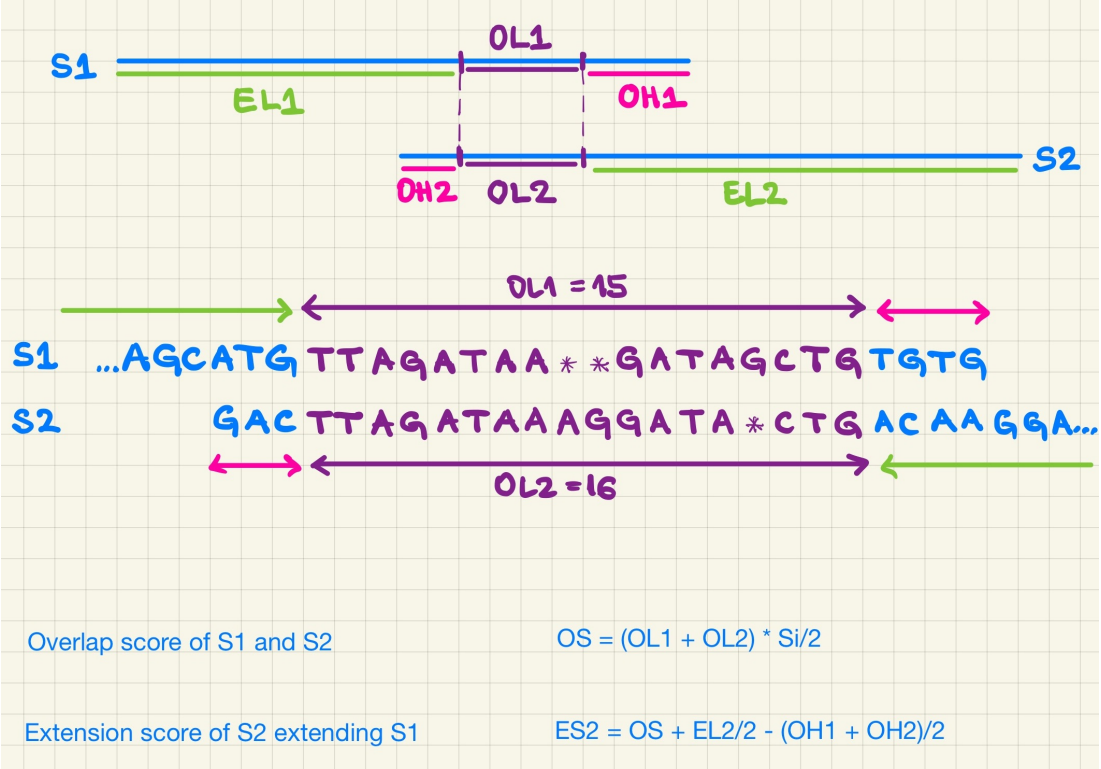
Izlazni podaci:

- 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*.

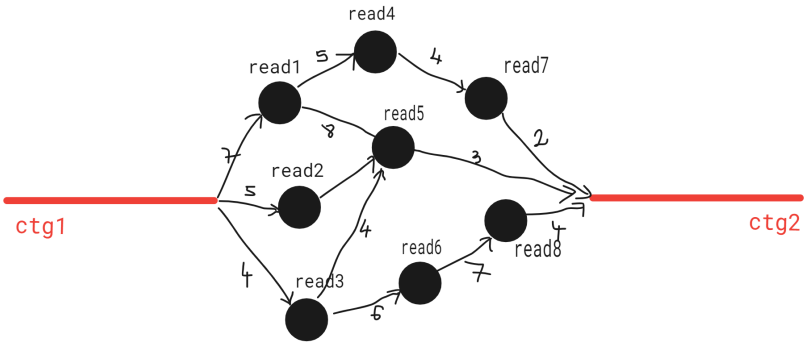



**Slika 2.1:** Postupak računanja vrijednosti preklapanja i nadovezivanja

## 2.1. Koraci algoritma

### Korak 1: Konstrukcija grafa preklapanja

Konstruirati graf preklapanja koji će povezati fragmentirane contige i očitanja




**Slika 2.2:** Primjer grafa preklapanja

### Korak 2: Identifikacija usidrenih čvorova

Identificirati usidrene čvorove u grafu preklapanja.

### Korak 3: Pronalaženje optimalnih staza

Pronaći optimalne staze iz svakog usidrenog čvora do mogućih završnih usidrenih čvorova. Ove staze određuju kako će se fragmentirani contigi povezivati međusobno.
(ctg1 -> read1-> read5 -> ctg2)

### Korak 4: Povezivanje contiga

Na temelju optimalnih staza, povezati fragmentirane contige u cijeli genom. Koristititi informacije o preklapanju i veze između contig-a i očitanja i očitanja i očitanja kako bismo dobili ispravno generirani genom.

# 3. Rezultati testiranja

Analiziramo vrijeme izvođenja algoritma i količinu zauzete memorije.

## 3.1. Vrijeme izvođenja

Konstruiranje grafa na temelju:

- overlap score: 1.415 milliseconds
- extension score: 1.721 milliseconds
- MonteCarlo metoda: 1.515 milliseconds

## 3.2. Količina zauzete memorije

Konstruiranje grafa na temelju:

- overlap score: 475.621 MB
- extension score: 478.988 MB
- MonteCarlo metoda: 476.492 MB

# 4. Zaključak

Napravili smo algoritam za poboljšanje djelomično sastavljenog genoma dugim očitanjima. Korištenjem grafa preklapanja i traženjem optimalnih staza, uspjeli smo povezati fragmentirane contige u cijeli genom. Naš algoritam koristi preklapanja dobivena alatom Minimap2 kao ulazne podatke i generira poboljšani skup sastavljenih contiga u FASTA formatu kao izlaz.

Testiranjem algoritma, proučavali smo vrijeme izvođenja i količinu zauzete memorije. Rezultati su pokazali da je naš algoritam učinkovit i može se primijeniti na velike skupove podataka s minimalnim resursnim zahtjevima.

Zaključno, naš algoritam za poboljšanje djelomično sastavljenog genoma dugim očitanjima pokazao se kao koristan alat u području bioinformatike. Daljnjom primjenom i nadogradnjom ovog algoritma možemo pridonijeti rješavanju problema sastavljanja genoma i razumijevanju složenih genomskih struktura.

# 5. 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.