Poboljšanje djelomično sastavljenog genoma dugim očitanjima

Temeljeno na radu:
Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads

Huilong Du, Chengzhi Liang
bioRxiv 345983

Projekt iz Bioinformatike

Autori:
- Domagoj Latečki
- Juraj Fulir
- Rudolf Lovrenčić

Izgradnja puteva (Deterministički)

Deterministički pristup

- Zabrana ponavljanja čvorova
- Gradnja počinje od svakog očitanja povezanog na neki *anchor*
- Odabire se čvor s navećom vrijednošću odabrane metrike
- Ako čvor s najvećom vrijednošću metrike nije zadovoljavajuć, odabire se sljedeći najbolji

Parametri

- Metrika po kojoj se uzimaju najbolji čvorovi

Parametri (*Monte Carlo*)

Monte Carlo pristup

- Zabrana ponavljanja čvorova
- Vjerojatnost odabira susjeda proporcionalna korištenoj metrici
- U slučaju “slijepog čvora” pokušaj povratka
- Ako se istroše pokušaji povratka, put se odbacuje

Parametri

- Metrika po kojoj se uzimaju najbolji čvorovi
- Maksimalna duljina puta (bp)
- Broj pokušaja izgradnje puta
- Broj dozvoljenih povrataka

Konsenzusi između kontiga

Grupiranje

- Prozori
  - Usporedba za graničnu duljinu puta

Konsenzusi

- Filtriranje *rijetkih* puteva
- Konsenzus unutar grupe
  - Prosječna duljina puteva unutar grupe
- Konsenzus između grupa
  - *Valid path number*
  - Nadmetanje uzastopnih grupa

Parametri

- Prag razlike duljina najduljeg i najkraćeg puta za svrstavanje u jednu grupu (bp)
- Širina prozora (bp)
- Omjer vrha i dna za grupiranje puteva

Konačan scaffold

Izgradnja konačnog puta

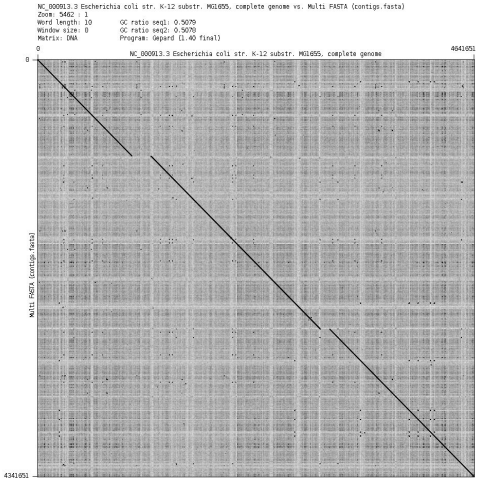
- Konsenzus najučestalijeg para kontiga uzima se kao početni put
- Put se proširuje s obje strane idućim *najučestalijim spojivim* konsenzusom
- Rezultat je lanac kontiga povezanih očitanjima

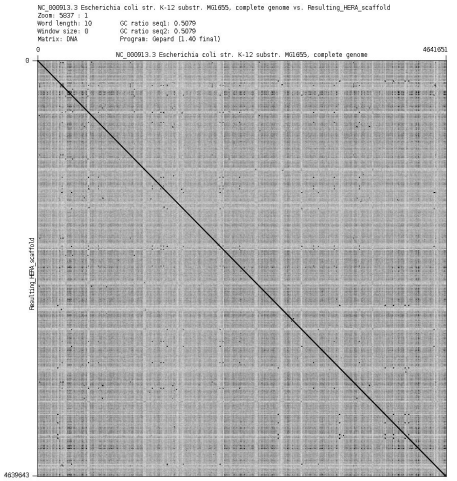
Zapis sekvence

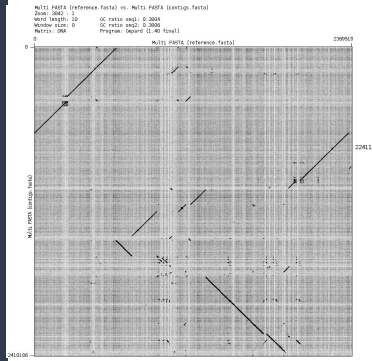
- Učitavaju se sekvence kontiga i očitanja (memorija)
- Obilaskom čvorova konačnog puta, sekvence se režu, okreću, invertiraju

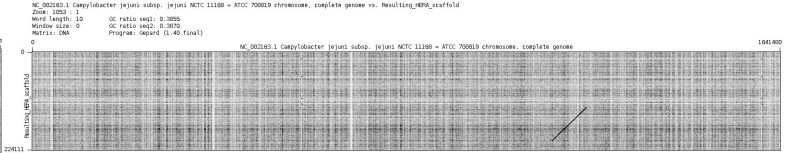
Rezultati

- Testirano na djelomice sastavljenom genomu *E. Coli*
- Rezultati analizirani alatom *Gepard*
- Genom uspješno potpuno sastavljen u **345 sekunde**
- Pritom izgrađeno **10.411 puteva** kroz graf preklapanja
- Na CJejuni ostvaruje loš rezultat (2 kontiga)









Hvala na pažnji!

Pitanja?