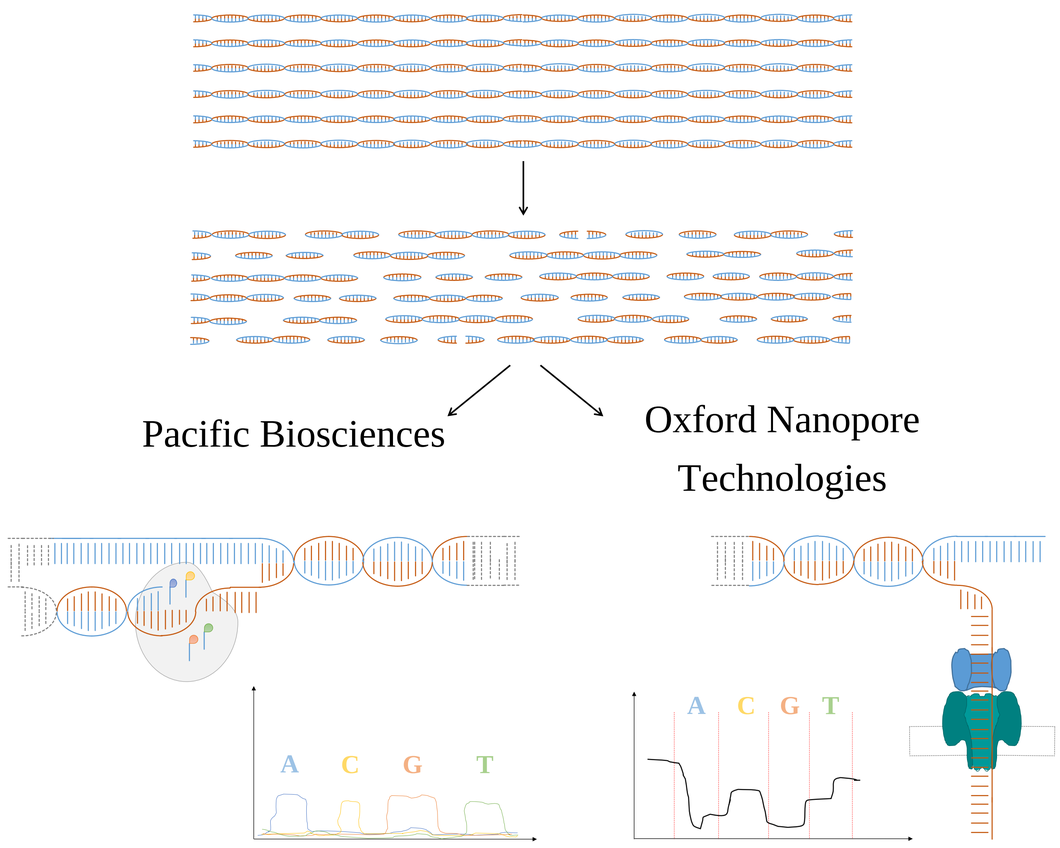
Algoritmi za de novo sastavljanje velikih genoma

Robert Vaser, mag. ing.
mentor: prof. dr. sc. Mile Šikić
Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva

1. Uvod

Treća generacija tehnologija za sekvenciranje omogućila je manju fragmentiranost sastavljenih genoma zahvaljujući dugačkim očitanjima čiji je jedini nedostatak velik udio pogreške. Unatoč tome, algoritmi temeljeni na teoriji grafova uspješni su u sastavljanju kraćih i srednje dugačkih genoma bez prethodnog ispravljanja pogrešaka u očitanjima, ali zahtijevaju značajne resurse za veće genome poput biljaka i sisavaca.

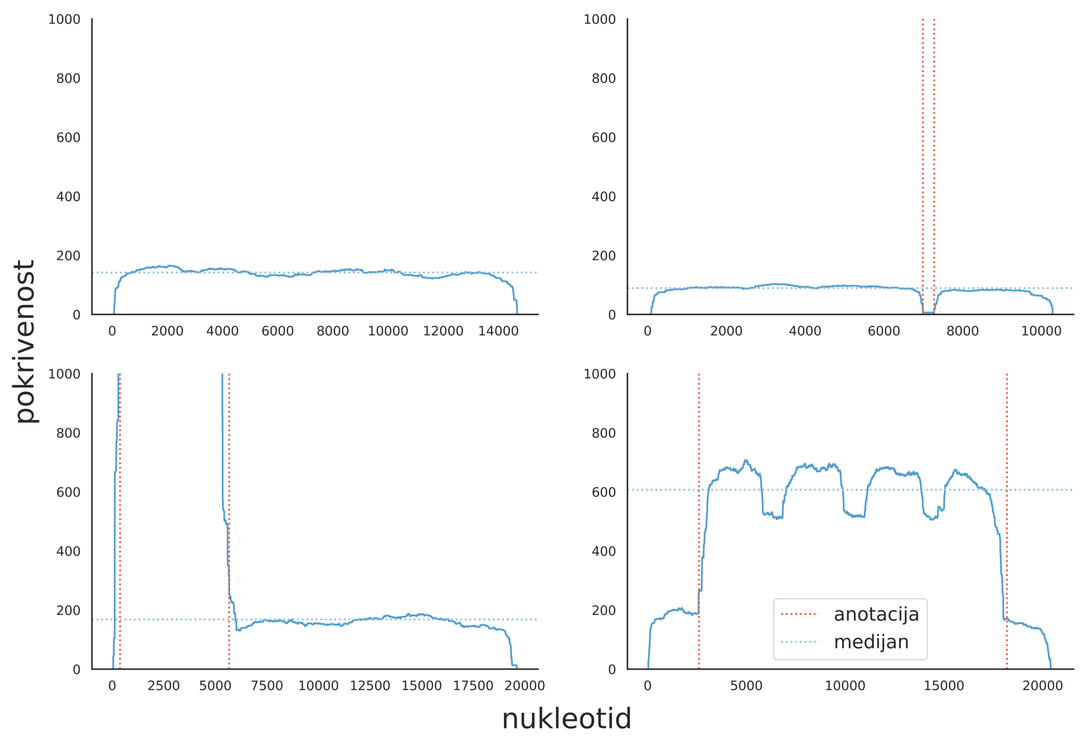


2. Opis problema

Za dani skup očitanja s prosječnom pogreškom od 12.5% te prosječne duljine od 15000 nukleotida, potrebno je rekonstruirati puno veći sekvencirani genom pomoću paradigme preklapanje-razmještaj-konsenzus. Navedeno treba biti izvedeno u što kraćem vremenu s minimalnom potrošnjom memorije.

3. Metodologija

Najprije je potrebno pronaći preklapanja između svakog para očitanja što se obično radi pomoću kratkih podnizova duljine *k*. Dobivena preklapanja primarno služe za izgradnju usmjerenog grafa preklapanja izbacivanjem svih sadržanih očitanja, ali također koriste se za izgradnju gomila s kojima je moguće odrediti očitanja koja petljaju dobiveni graf. Graf preklapanja se najprije pojednostavljuje izbacivanjem tranzitivnih bridova, puteva koji naglo završavaju te struktura koja podsjećaju na mjehuriće. Dodatno, crtanjem grafa u 2D koordinatnom sustavu omogućuje detekciju i uklanjanje bridova koji spajaju udaljene dijelove genoma. Svi putevi pojednostavljenog grafa koji nemaju grananja koriste se u završnoj fazi u kojoj se primjenjuje višestruko poravnanje očitanja kako bi se ispravila pogreška sekvenciranja.



4. Rezultati

Organizam	Pokrivenost	NG50	Točnost	Vrijeme (min)	Memorija (GB)
Klebsiella pneumoniae (O)	122	5456762	0.9841	120.04	3.24
Klebsiella pneumoniae (P)	45	5305421	0.9917	52.493	0.96
Saccharomyces cerevisiae (O)	59	639233	0.9742	91.24	2.89
Saccharomyces cerevisiae (P)	127	829833	0.9974	135.91	4.73
Drosophila melanogaster (O)	32	5674834	0.9885	674.48	11.35
Drosophila melanogaster (P)	109	7850171	0.9911	2486.14	32.09

5. Zahvala projektu

Ovaj rad financiran je sredstvima Hrvatske zaklade za znanost pod projektom De novo sastavljanje genoma i metagenoma (IP-2018-01-5886) te sredstvima Europskog fonda za regionalni razvoj pod projektom DATACROSS (KK.01.1.1.01.0009).