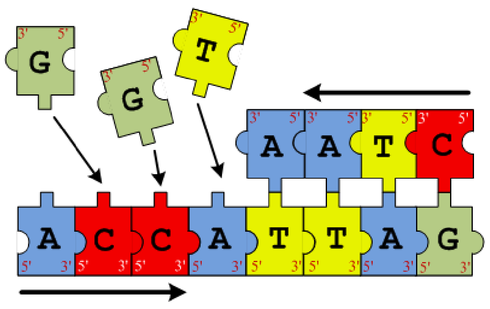
Algoritmi za *de novo* sastavljanje genoma iz sekvenciranih podataka treće generacije

Ivan Sović[1]
mentori: izv.prof.dr.sc. Mile Šikić[2], prof.dr.sc. Karolj Skala[1]
[1]Centar za informatiku i računarstvo, Institut Ruđer Bošković
[2]ZESOI, Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva

1. Uvod

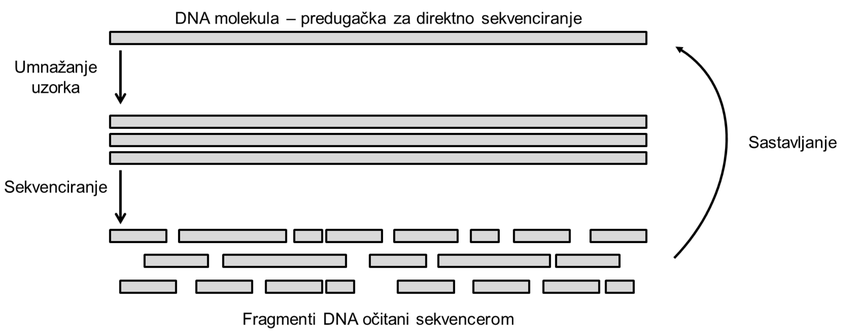
Oxford Nanopore Technologies (ONT) 2014. godine predstavio je revolucionarni uređaj za sekvenciranje treće generacije – MinION je veličine dlana, poveziv putem USB 3.0, cijene $1000 dolara, generira **dugačka očitanja**, no uz **vrlo veliku pogrešku (~15–35%)**.
Iako dugačka (1kbp-300kbp), očitanja su u pravilu znatno kraća od veličine genoma složenijih organizama (čovjek ~3Gbp), te je genom potrebno rekonstruirati (sastaviti) računalnim metodama.

ONT MinION Mk1



2. Opis problema

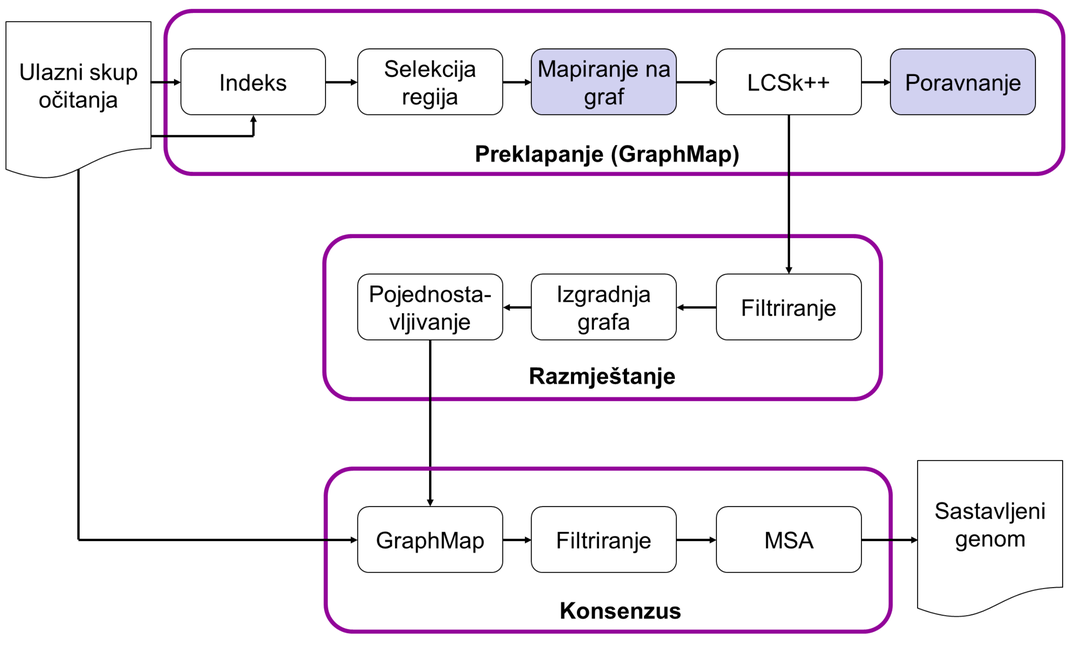
Cilj ovog istraživanja je razvoj nove metode za *de novo* sastavljanje genoma koja uspješno može rukovati ONT podatcima, ali bez vremenski zahtjevnog pred-koraka ispravljanja pogrešake. Kod *de novo* pristupa referentni genom organizma nije unaprijed poznat. Poteškoće u sastavljanju uzrokuju: (I) **pogreške u očitanjima** nastale prilikom sekvenciranja, (II) **ponavljajuće regije** u genomima, (III) stohastički procesi prilikom pripreme uzoraka za sekvenciranje.



3. Metodologija

Problem će biti riješen modularno, primjenom *Preklapanje-Razmještanje-Konsenzus* pristupa (eng. *Overlap-Layout-Consensus*, OLC). Istraživanje je podijeljeno u tri faze:

1. **Ispitivanje postojećih metoda** (eng. *benchmark*)
2. **Razvoj nove metode mapiranja očitanja s velikom pogreškom – GraphMap**; prilagodba za preklapanje očitanja
3. **Razvoj konsenzus metode** - ispravljanje pogrešaka u sastavljenom genomu

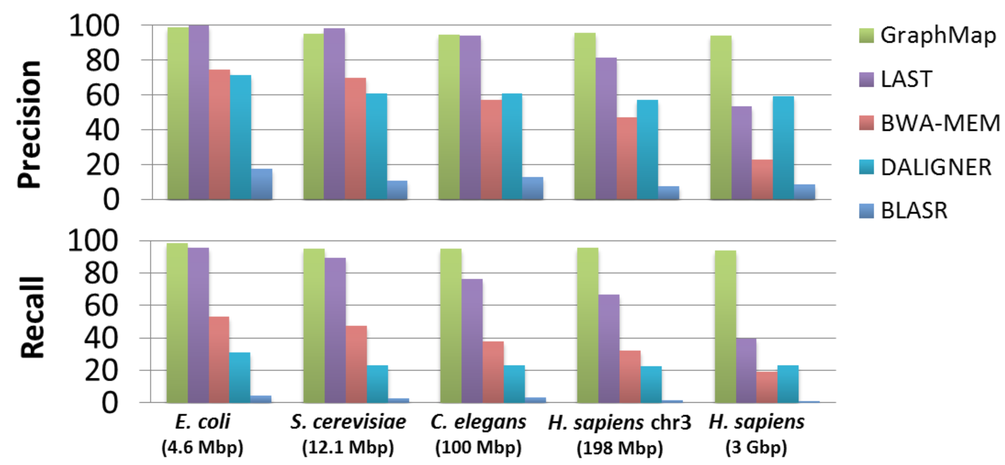


4. Rezultati

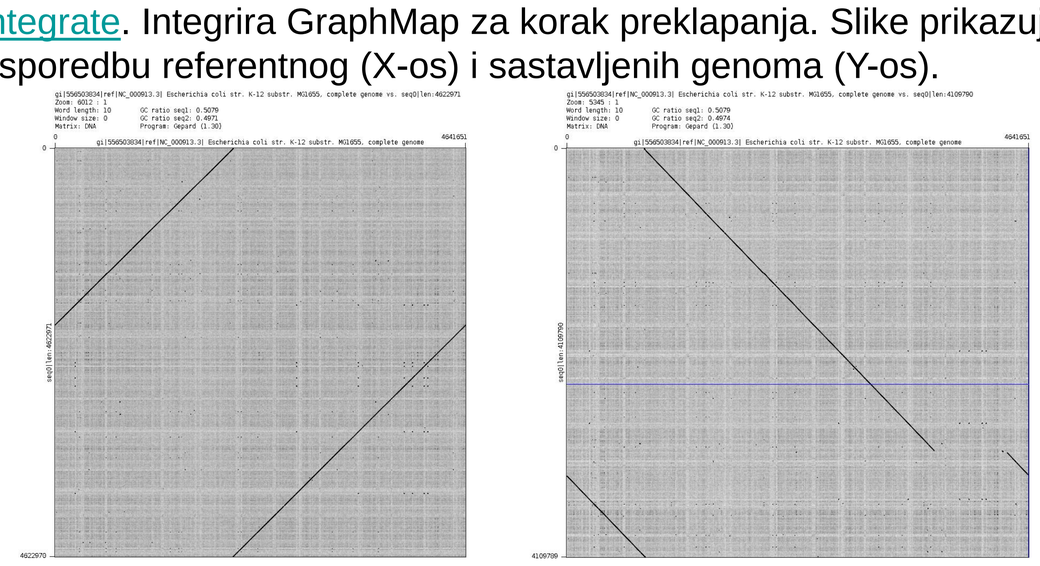
Ispitivanje postojećih metoda
Usporedno je ispitano 5 ne-hibridnih alata za sastavljanje genoma (eng. *assembler*) na 5 stvarnih ONT skupova podataka [1]. Tablica prikazuje rezultate na jednom od skupova podataka kojeg čine ONT očitanja 30x pokrivenosti *E. Coli* K-12 genoma [2].

Assembler	# ctg.	N50	Avg. Identity	#SNP	#Indel	CPU vrijeme [s]	Maks. memorija [GB]
LQS	3	4603990	98.49	4568	65283	2538.6	7.89
Falcon	124	13838	94.97	3206	59638	6.4	9.19
PBcR	1	4329903	94.03	7209	262357	13.7	5.94
Canu	10	4465231	95.77	5213	185027	11.2	3.49
Miniasm	3	3362269	84.04	247849	367927	0.015	1.96

Implementacija GraphMap metode za mapiranje/preklapanje dugačkih greškovitih očitanja
GraphMap je otvoreno dostupan na github.com/isovic/graphmap [3]. Slika prikazuje usporedbu s konkurentnim metodama za mapiranje podataka treće generacije na 5 simuliranih skupova podataka.



Preliminarni rezultati sastavljanja *E. Coli* K-12 genoma
Ra – novi alat za sastavljanje genoma, github.com/mariokostelac/ra-integrate. Integrira GraphMap za korak preklapanja. Slike prikazuju usporedbu referentnog (X-os) i sastavljenih genoma (Y-os).



Uspješno sastavljen genom iz ONT podataka — Greška u sastavljanju genoma iz PacBio podataka

5. Zaključak

U ovome radu demonstrirano je kako je moguće sastaviti potpuni bakterijski genom iz sekvenciranih podataka s vrlo velikom pogreškom bez koraka ispravljanja pogreške u podatcima. Prije dovršetka ovog rada potrebno je: (I) unaprijediti metodu preklapanja očitanja za dodatno povećanje preciznosti, (II) implementirati konsenzus metodu i (III) unaprijediti metodu razmještanja kako bi bila stabilna na varijacije u ulaznim podatcima.

Reference

[1] Sović, I., Križanović, K. et al. Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads. Biorxiv, doi:10.1101/030437, http://biorxiv.org/content/early/2015/11/13/030437. (2015)
[2] Loman, N.J. et al. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat. Methods, 12, 733–5. (2015)
[3] Sović, I., Šikić, M. et al. Fast and sensitive mapping of error-prone nanopore sequencing reads with GraphMap. Biorxiv, doi:10.1101/020719, http://biorxiv.org/content/early/2015/06/10/020719. (2015)

Dan doktorata Fakulteta elektrotehnike i računarstva Sveučilišta u Zagrebu, 3. ožujka 2016.