SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 6149

De novo sastavljanje genoma vođeno referencom

Sara Bakić

Zagreb, lipanj 2019.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA ZAVRŠNI RAD MODULA

Zagreb, 13. ožujka 2019.

ZAVRŠNI ZADATAK br. 6149

Pristupnik: **Sara Bakić (0036498154)**
Studij: Računarstvo
Modul: Računarska znanost

Zadatak: **De novo sastavljanje genoma vođeno referencom**

Opis zadatka:

Sastavljanje genoma jedan je od najsloženijih zadataka u području bioinformatike i računalne biologije. S obzirom da je problem NP težak, koriste se brojne heurističke metode. Cilj ovoga rada je testirati važnost pravilne detekcije kimernih i repetitivnih očitanja te pronalaska svih preklapanja u de novo sastavljanju genoma. Koristeći alat minimap2 mapirati očitanja na poznati referentni genom. Nakon toga detektirati kimerna i repetitivna očitanja s posebnim naglaskom na repetitivnim očitanjima koja povezuju udaljene regije genomu. Koristeći preostala očitanja sastaviti genom pomoću alata za sastavljanje genoma Ra. Na kraju usporediti rezultate s i bez detekcije očitanja te isprobati različite parametre za mapiranje.
Programski kod je potrebno komentirati i pri pisanju pratiti neki od standardnih stilova. Kompletnu aplikaciju postaviti na repozitorij Github.
U svezi dobivanja detaljnih informacija obratiti se Robertu Vaseru, mag. ing.

Zadatak uručen pristupniku: 15. ožujka 2019.
Rok za predaju rada: 14. lipnja 2019.

Mentor:

Prof. dr. sc. Mile Šikić

Predsjednik odbora za završni rad modula:

Doc. dr. sc. Marko Čupić

Djelovođa:

Izv. prof. dr. sc. Tomislav Hrkać

Zahvaljujem se svima koji su savjetom i podrškom pomogli mi, kako u izradi ovoga rada, tako i u mom preddiplomskom obrazovanju uopće. Posebna zahvala mojim mentorima Mili Šikiću i Robertu Vaseru na divnoj suradnji, strpljenju i pruženoj pomoći.

SADRŽAJ

1. Uvod

Bioinformatika (*bios* = *život* + *informatika*) interdisciplinarno je znanstveno područje koje povezuje biološke podatke s tehnikama pohrane, distribucije i analize informacija u svrhu istraživanja u brojnim područjima, na čelu s biomedicinom. Obuhvaća biologiju, računarsku znanost, matematiku i statistiku te ih koristi u procesu analize i interpretacije bioloških podataka. Veliki napredak u tehnikama generiranja bioloških podataka te drastičan pad u cijeni procesiranja tih podataka, omogućio je ubrzan razvoj bioinformatike (Lesk, 2013).

Obzirom da je sva genetska informacija sadržana u DNA ili RNA lancu, informacije potrebne bioinformatici su upravo podaci o slijedu nukleotida u tim lancima. Zbog toga jedni od temeljnih oblika bioloških podataka u bioinformatičkim istraživanjima su sekvence genoma ili genomi u cijelosti.

Sekvenciranje je proces određivanja poretka nukleinskih baza – adenina (**A**), citozina (**C**), gvanina (**G**), timina (**T**), uracila (**U**) – u DNA ili RNA lancu. Pomoću tih informacija raznim se tehnikama utvrđuju genomi, srodnosti i evolucije vrsta, novi geni koji se pridružuju bolestima te se pokušavaju pronaći adekvatni lijekovi.

Postoji nekoliko metoda sekvenciranja koje kao zajedničko obilježje imaju ograničenu sposobnost čitanja nukleotida. Ovisno o metodi, duljine pročitanih nukleotida kreću se između pedeset i nekoliko stotina tisuća nukleotida. Kako bi sekvenciranje bilo uspješno nužno je podijeliti lance na manje fragmente te onda primjenom neke metode sekvenciranja očitati slijed nukleotida svakoga fragmenta. Dominantom strategijom danas se smatra *shotgun* sekvenciranje kojim se DNA umnaža i potom na slučajan način lomi u mnogo malih fragmenata koji se potom sekvenciraju. Obzirom da je krajnji cilj imati potpuno sastavljenu sekvencu, kada bi na ovaj način samo jednom razbili DNA na mnoštvo manjih fragmenata bilo bi nemoguće ponovno odrediti stvarni poredak tih fragmenata i sastaviti sekvencu. Upravo zato, shotgun sekvenciranje više se puta ponavlja kako bi se dobila očitanja iz kojih je pomoću preklapajućih krajeva moguće sastaviti kontinuiranu sekvencu.

U okviru *shotgun* sekvenciranja postoje dva temeljna pristupa.

Prvi pristup je hijerarhijsko *shotgun* sekvenciranje koje kao prvi korak podrazumijeva izradu fizičke mape genoma. Pomoću te mape određuje se minimalan broj segmenata potreban za sekvenciranje genoma u cijelosti. Genom se potom dijeli u dugačke fragmente parova baza te se ti fragmenti kloniraju umetanjem u bakterije domaćine koristeći BAC-ove (*Bacterial artificial chromosome*. Naposljetku se dobivene kopije fragmenata *shotgun* metodom usitnjavaju na manje dijelove i sekvenciraju.

Problem kod ovakvog pristupa je izrada fizičke mape genoma jer je to skup i spor proces. Umetanje BAC-ova u bakterije i čekanje da se one razmnože kako bi se onda ti fragmenti izdvajali iz bakterija zahtijeva puno vremena te stručno osoblje.

Iz tih razloga razvijen je drugi pristup, *shotgun* sekvenciranje cijelog genoma. DNA se dijeli u slučajne dijelove raznih duljina koji se potom kloniraju. Fragmenti se zatim sekvenciraju i to uvijek od 5' kraja prema 3' te se rekonstrukcijom očitanja dobiva potpuni slijed.

Obzirom da *shotgun* sekvenciranje cijelog genoma ima veliku vremensku i cjenovnu prednost nad hijerarhijskim *shotgun* sekvenciranjem, dominantno je u današnje vrijeme. Ipak, postoji jedan problem s takvim sekvenciranjem, a to je mogućnost ispravnog sastavljanja genoma, pogotovo onih genoma koji imaju ponavljajuće regije (Šikić i Domazet-Lošo, 2013).

Temeljni fokus u ovom radu je na sastavljanju genoma vođenom referentnom sekvencom.

U drugom poglavlju nalazi se teoretska pozadina problema i pojašnjenja pojmova koja su fokus ovoga rada te koja će se kroz cijeli rad pojavljivati.

U trećem poglavlju prikazani su i ukratko pojašnjeni formati podataka koji se koriste u bioinformatici, a koje sam koristila u svome radu.

U četvrtom poglavlju potanko su objašnjeni algoritmi i korištene metode.

U petom poglavlju nalaze se opis i objašnjenje implementacije.

Šesto poglavlje prikazuje rezultate testiranja.

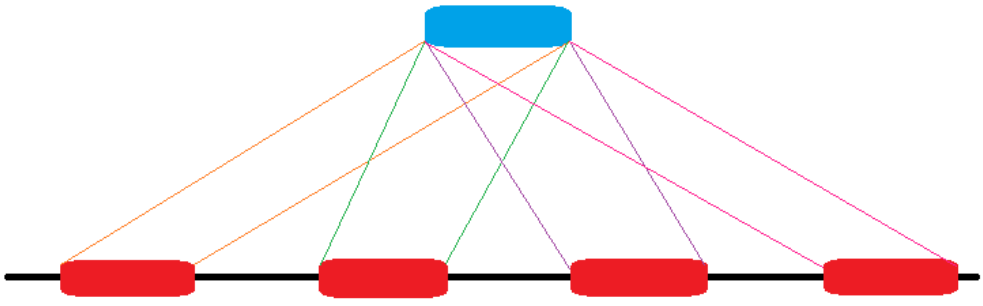
2. Sastavljanje genoma

Sastavljanje genoma jedan je od najsloženijih problema u području bioinformatike. Pripada skupu NP teških problema i zbog toga se u rješavanju koriste brojne heurističke metode. Problem u sastavljanju genoma predstavljaju takozvana ponavljajuća i kimerna očitanja.

2.1. Ponavljajuća očitanja

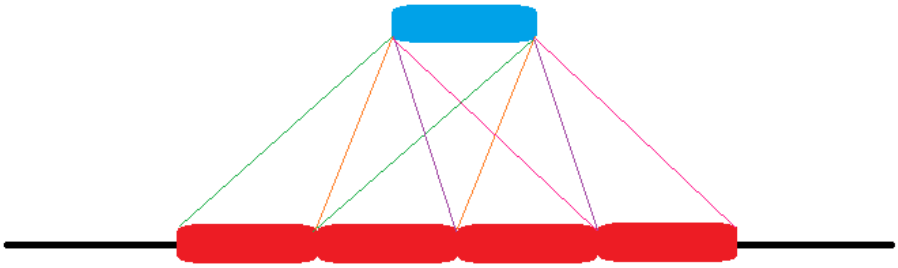
Ponavljajućim očitanjima smatramo ona očitanja koja se mogu mapirati na više mjesta u referentnom genomu zbog postojanja dugačkih istih ili vrlo sličnih sljedova u genomu. Postoje dvije temeljne vrste ponavljajućih regija (Šikić i Domazet-Lošo, 2013).

Prva vrsta su one regije koje se nalaze na udaljenim područjima u genomu i nisu susjedne te se takve regije nazivaju raspršene (interspersed) ponavljajuće regije. Raspršene ponavljajuće sekvence najčešće nastaju kao produkt takozvanih *transposona*, odnosno dijelova DNA koji, uz pomoć enzima, imaju mogućnost umetanja na određena mjesta u kromosomu.



Slika 2.1: Raspršene ponavljajuće regije na referenci

S druge strane, postoje takozvane tandem ponavljajuće regije koje se u genomu pojavljuju jedna za drugom, grupirane na jednom dijelu genoma. One nastaju kao posljedica raznih bioloških mehanizama te se pojavljuju u različitim duljinama i brojnostima unutar genoma (Hauth i Joseph, 2002).

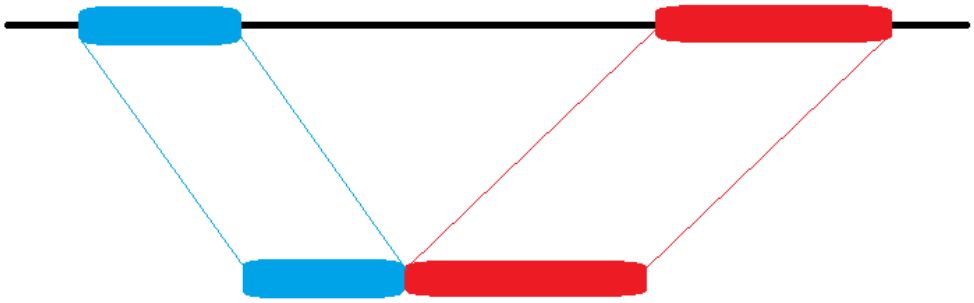


Slika 2.2: Tandem ponavljajuće regije na referenci

Gore objašnjen *shotgun* pristup sekvenciranju cijelog genoma u kombinaciji s ponavljajućim regijama genoma, posebice onima duljima, stvara potencijalne probleme u sastavljanju genoma. Ako prilikom *shotgun* sekvenciranja izgeneriramo sekvence koje su takve da pojedine ponavljajuće regije genoma ne mogu u cijelosti premostiti, nastaje dvosmislenost prilikom odabira sekvence koja stvarno pripada tom dijelu genoma te kvaliteta sastavljanja genoma postaje vrlo upitna. Upravo zato je vrlo bitno kvalitetno detektirati ponavljajuća očitanja i ponavljajuće regije na genomu.

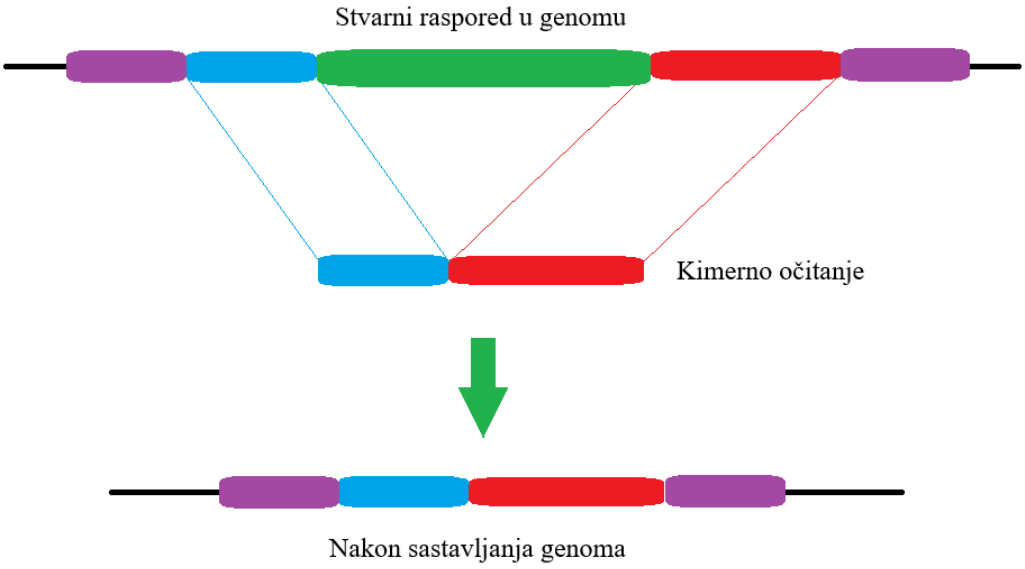
2.2. Kimerna očitanja

Kimernim očitanjima (*kimera – jedinka sastavljena od dijelova različitih rasa*) smatramo ona očitanja koja nastaju pogrešnim spajanjem dva očitanja koja predstavljaju u prirodi veoma udaljene dijelove genoma. Oni nastaju kao posljedica grešaka prilikom identificiranja nukleotida u očitavanjima. Zbog brzine kojom uređaji novijih generacija to rade događaju se greške u očitanjima zbog kojih se pojavljuju lažna preklapanja između prirodno vrlo udaljenih očitanja te se kao posljedica ta očitanja spajaju u jedinstveno očitanje koje bi trebalo predstavljati jedan dio genoma, ali zapravo predstavlja dva vrlo udaljena dijela genoma (atdbio, 2011).



Slika 2.3: Kimerno očitanje nastala spajanjem dvaju očitanja

I takva očitanja stvaraju probleme pri izgradnji genoma. Obzirom da je proces sastavljanja genoma zapravo proces traženja preklapanja krajeva očitanja, kimerno očitanje će na jednom kraju pronaći preklapanje s jednim dijelom genoma, a na drugoj s potpuno drugim dijelom genoma što očito narušava sliku stvarnog genoma.(Institute, 2003)



Slika 2.4: Problemi sa sastavljanjem genoma uzrokovani kimernim očitanjima

Osnovni cilj moga rada jest detekcija kimernih i ponavljajućih očitanja kako bi, uz izostavljanje tih očitanja, što kvalitetnije i točnije sastavili genom.

3. Podaci

3.1. FASTA format

Bioinformatički podaci najčešće se nalaze u FASTA tekstualnom formatu. FASTA je tekstualni format kojim se prikazuju sljedovi nukleotida, a svaki nukleotid predstavlja jedno slovo (Šikić i Domazet-Lošo, 2013).

Slovo reprezentat	Nukleotid
A	Adenin
C	Citozin
G	Gvanin
T	Timin
U	Uracil
R	Adenin ili gvanin
Y	Citozin, timin ili gvanin
K	Gvanin, timin ili uracil
M	Adenin ili gvanin
S	Citozin ili gvanin
W	Adenin, timin ili uracil
B	Ne adenin
D	Ne citozin
H	Ne gvanin
V	Niti timin niti uracil
N	Adenin, citozin, gvanin, timin, uracil
X	Maskiranje
-	Procjep neodređene duljine

Tablica 3.1: Tablica znakova u FASTA formatu

Svaka sekvenca u FASTA formatu sastoji se od dvije linije. Prva linija započinje znakom „>“ i nakon nje slijedi identifikator slijeda. Druga linija sastoji se od samoga slijeda.

>a019fb87-85b7-495a-b6dc-789a2f8c4572_Basecall_2D_000_2d
TACGCATAAGCGCCAAAAGCACAAGATGCTCACCGCCAG...

3.2. FASTQ format

Drugi česti format je FASTQ format koji je vrlo sličan FASTA formatu.

Zapis sekvence sastoji se od 4 linije. Prva linija započinje znakom „@“ nakon koje slijedi identifikator slijeda. U drugoj liniji nalazi se sami slijed. Treća linija započinje znakom „+“ nakon kojeg opcionalno slijedi identifikator slijeda i u četvrtoj liniji nalazi se vrijednost kvalitete slijeda iz druge linije (Šikić i Domazet-Lošo, 2013).

@cluster_2:UMI_ATTCCG
TTTCCGGGGCACATAATCTTCAGCCGGGCGC...
+
9C;=;=<9@4868>9:67AA<9>65<=>591

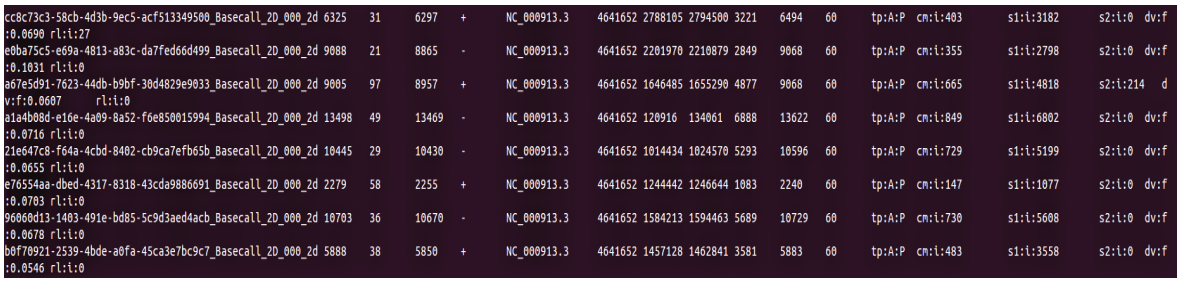
Početni podaci su se nalazili upravo u FASTA i FASTQ formatu te su kao takvi predstavljali ulaz alatu minimap2 koji je na temelju očitanja koja se najčešće nalaze u FASTQ formatu i reference koja je najčešće u FASTA datoteci mapirao očitanja na sekvencu i to prikazao u PAF formatu.

3.3. PAF format

PAF format tekstualni je format koji opisuje aproksimativne pozicije mapiranja između dva seta sekvenci. Sastoji se od ovih informacija odijeljenih tabulatorom (Li, 2018b):

Stupac	Tip	Opis
1	string	Naziv očitanja
2	int	Duljina očitanja
3	int	Početna pozicija mapiranja očitanja
4	int	Krajnja pozicija mapiranja očitanja
5	char	Relativni slijed "+" ili "-"
6	string	Naziv reference
7	int	Duljina reference
8	int	Početna pozicija mapiranja očitanja na referenci
9	int	Krajnja pozicija mapiranja očitanja na referenci
10	int	Broj poklapanja očitanja i reference
11	int	Ukupan broj poklapanja, promašaja, umetanja i brisanja u poravnanju
12	int	Kvaliteta mapiranja

Tablica 3.2: Opis PAF formata



Slika 3.1: Isječak mapiranja u PAF formatu

4. Metode

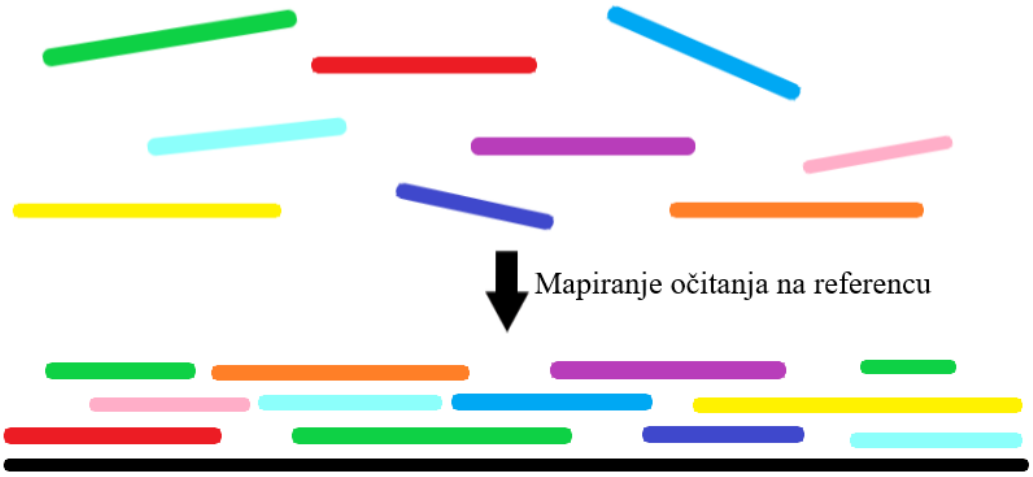
4.1. Priprema podataka alatom minimap2

Minimap2 svestrani je alat korišten za mapiranje DNA ili RNA sekvenci na referentnu bazu podataka.

Nekoliko je najčešćih tipova korištenja alata minimap2:

- mapiranje PacBio ili Nanopore očitanja na ljudski genom
- traženje preklapanja između dugačkih očitanja
- splice-aware poravnanje na referentni genom
- poravnanje jednostrukih ili uparenih Illumina očitanja
- assembly-to-assembly poravnanje
- poravnanje cijelih genoma srodnih vrsta

Minimap2 sam koristila za mapiranje očitanja u FASTA formatu na referentnu bazu podataka u FASTQ formatu. Kao rezultat dobila sam indeks mapiranja u PAF formatu i pomoću njega detektirala ponavljajuća i kimerna očitanja.



Slika 4.1: Mapiranje očitanja na referencu

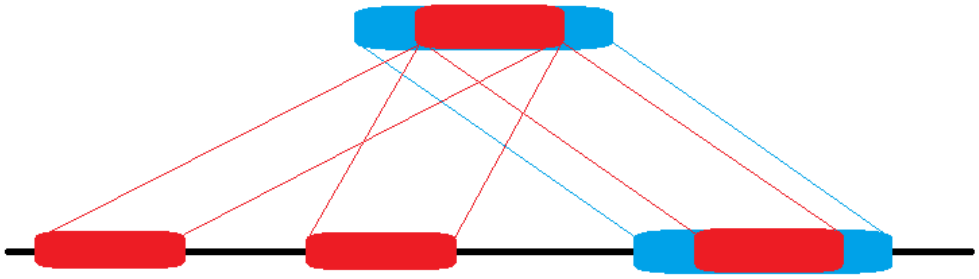
Kako bi korištenje rezultata minimap2 u svrhu detekcije kimernih i ponavljajućih očitanja bilo što razumljivije, ovo je kratak pregled algoritma minimap2: (Li, 2018a)

1. za **I** referentnih baza se pronađu takozvani minimizeri - specifični podnizovi preddefinirane duljine **k** koji služe za detekciju sličnih regija među dvjema sekvencama - i spremaju u indeks u obliku tablice raspršenog adresiranja
2. čita se **K** baza sekvenci i za svaku se ponavljaju koraci 3 do 7
3. svaki minimizer očitanja provjerava se u indeksu minimizera referentne baze i ako se ne nalazi među preddefiniranih **f** najčešćih, skupljaju se njegova pojavljivanja u referenci te se ona nazivaju *seeds*
4. *seeds* sortiraju se po poziciji na referenci te se dinamički povezuju u lance
5. za svaki lanac, počevši od najboljeg po njegovoj lančanoj vrijednosti, provjerava se preklapanje s lancima u postojećem (početno praznom) setu primarnih mapiranja; ako je preklapanje manje od preddefinirane vrijednosti **mask-level** dodaje se u set primarnih mapiranja, inače se deklarira kao sekundarno mapiranje lancu s kojim ima preklapanje veće od zadane vrijednosti
6. uzimaju se sva primarna mapiranja i **N** najboljih sekundarnih očitanja koja imaju lančanu vrijednost veću od **p** posto odgovarajućeg primarnog mapiranja
7. filtriraju se oni *seedovi* koji vode do velikih umetanja i brisanja prilikom poravnanja te se nad preostalima vrši globalno poravnanje; lanci se dijele na manje ako vrijednost poravnanje padne za **z**, zanemarujući dugačke praznine te se lanci i njihova mapiranja ispisuju u odgovarajućem formatu (u našem slučaju PAF formatu)
8. ako postoji više sekvenci vraća se na korak 2 dok god postoji još neobrađenih sekvenci
9. ako postoji više referentnih sekvenci ponavlja se cijeli postupak dok god sve referentne sekvence nisu obrađene

4.2. Detekcija ponavljajućih i kimernih očitanja

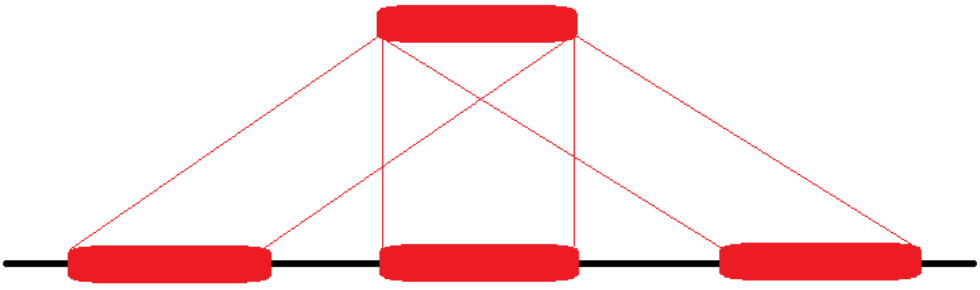
Obzirom da su ponavljajuća očitanja definirana kao ona očitanja koja se mogu mapirati na više mjesta na genomu, algoritam minimap2 prilikom izgradnje indeksa mapiranja detektira sva višestruka mapiranja ponavljajućih sekvenci.

Sekvenca može biti ponavljajuća u dva smisla. Primjerice, sekvenca kao takva može imati mapiranje na referentni genom cijelom svojom duljinom (misli se na barem 90% stvarne duljine sekvence) te unutar sebe sadržavati regiju koja se naknadno u jednom ili više navrata mapira na druga mjesta na genomu.



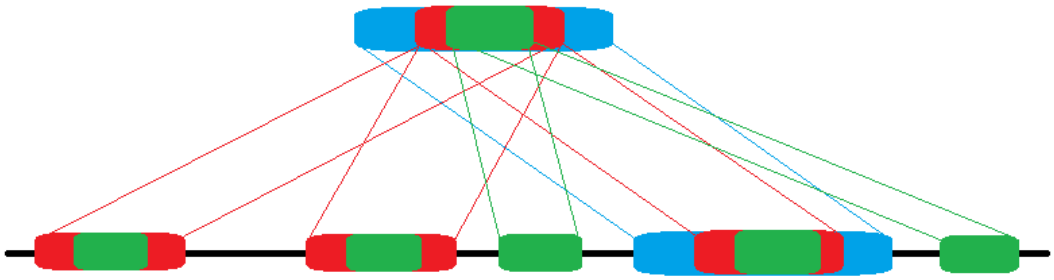
Slika 4.2: Ponavljajuća sekvenca s jednim primarnim mapiranjem na referencu

S druge strane, postoje sekvence koje su u cijelosti ponavljajuće, odnosno njihova najbolja (najdulja) mapiranja su višestruka.



Slika 4.3: Sekvenca u cijelosti ponavljajuća mapirana na referencu

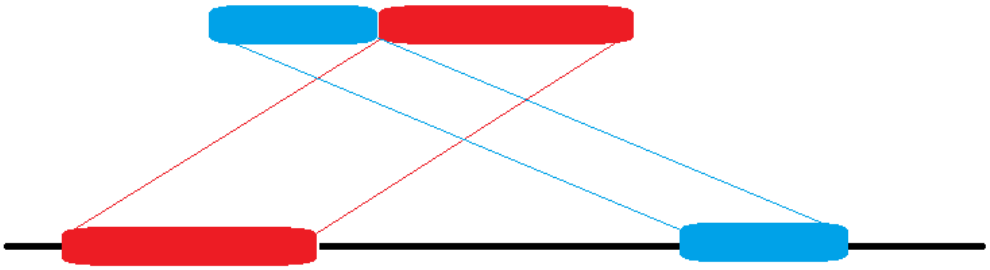
Potencijalno, unutar svake ponavljajuće regije na sekvenci može postojati manja regija koja je također ponavljajuća.



Slika 4.4: Ponavljajuća regija očitanja unutar ponavljajuće regije

U gore pojašnjenom algoritmu minimap2 za ponavljajuće regije posebno su važni koraci 5 i 6 gdje se ponavljajuće regije prepoznaju kao sekundarna mapiranja odgovarajućem primarnom mapiranju te se u 6. koraku najveće ponavljajuće regije uzimaju u obzir prilikom poravnanja sekvence na referencu.

Dok ponavljajuća očitanja možemo zamisliti kao učahurene ponavljajuće regije unutar cijelog očitanja, temeljna karakteristika kimernih očitanja jest da su ona "isjeckana" i tako odvojeno mapirana na različite dijelove genoma.



Slika 4.5: Mapiranje kimernog očitanja na genom

Obzirom da su ona nastala pogrešnim spajanjem prirodno odvojenih očitanja, nemoguće ih je u jednom dijelu kvalitetno mapirati na genom.

Za njih je najvažniji korak 7 u algoritmu minimap2 koji lance dijeli u manje ako vrijednost poravnanja pada, zanemarujući velike praznine.

Nakon parsiranja PAF ispisa minimap2-ovog algoritma, sva dostupna mapiranja spremljena su u mapu koja je kao ključ imala naziv sekvence koja se mapira, a kao vrijednost pridruženu ključu sva pronađena mapiranja tog očitanja.

Kao potencijalna kimerna i ponavljajuća očitanja nameću se ona očitanja za koja je pronađeno više od jednog mapiranja.

Poznavajući početne i krajnje pozicije mapiranja očitanja moguće je jasno odvojiti kimerna od ponavljajućih. Ponavljajuća očitanja imat će višestruka mapiranja kod kojih su područja očitanja koja se mapiraju u odnosu podskupa, dok će kimerna očitanja imati višestruka mapiranja kod kojih su pozicije mapiranja na očitanju praktički bez presjeka.

Primjerice, imamo očitanje koje se mapira na šest različitih mjesta u genomu i to na način prikazan u tablici 4.1.

Početna pozicija na očitanju	Krajnja pozicija na očitanju	Početna pozicija na referenci	Krajnja pozicija na referenci
85	16679	1865475	1881092
7822	8577	1978527	1979239
7822	8577	257932	258644
7822	8577	19820	20532
7822	8577	3583458	3584170
7822	8577	1049808	1050520

Tablica 4.1: Primjer mapiranja ponavljajuće sekvence

Po pozicijama mapiranja očitanja vidljivo je da je prvo mapiranje primarno mapiranje toga očitanja. Gledajući ostala mapiranja lako se uočava da se sva ta mapiranja mogu zamisliti kao podregije regije koja čini primarno mapiranje zbog čega takvo očitanje klasificiramo kao ponavljajuće očitanje.

S druge strane moguće je i očitanje s dva mapiranja prikazana u tablici 4.2.

Početna pozicija na očitanju	Krajnja pozicija na očitanju	Početna pozicija na referenci	Krajnja pozicija na referenci
17	3842	756995	760771
3884	15141	3320818	3331885

Tablica 4.2: Primjer mapiranja ponavljajuće sekvence

Gledajući pozicije mapiranja toga očitanja, lako je uočljivo da pozicije prvog i drugog mapiranja nemaju presjek, odnosno prvi dio očitanja mapira se na jedan dio reference, a drugi, odvojeni dio očitanja mapira se na drugi dio reference. Zbog toga, takvo očitanje klasificiramo kao kimerno očitanje.

4.3. Potencijalni problemi prilikom detekcije kimernih i ponavljajućih očitanja

Nekoliko je slučajeva u kojima nije odmah vidljivo je li neko očitanje kimerno, ponavljajuće ili regularno očitanje.

Prvi takav slučaj je očitanje koje ima ovakva mapiranja:

Početna pozicija na očitanju	Krajnja pozicija na očitanju	Početna pozicija na referenci	Krajnja pozicija na referenci
170	2445	1870705	1872977
2531	3306	3583433	3584182
2531	3306	257920	258669
2531	3306	19808	20557
2531	3306	1978515	1979264
2531	3306	1049783	1050532
2531	3306	279175	279924

Tablica 4.3: Primjer mapiranja kimernog očitanja s ponavljajućom regijom

Ovo očitanje ima dvije vrste mapiranja. Vidljivo je da postoje mapiranja zbog kojih bi očitanje detektirali kao kimerno (dijelovi očitanja mapiraju se na pozicijama 170-2445, 2531-3306), ali postoji i regija koja ima ponavljajuće mapiranje (na pozicijama 2531-3306). Takvo mapiranje se usprkos ponavljajućoj regiji detektira kao kimerno očitanje jer je iz mapiranja očito da je očitanje neregularno u smislu nepripadnosti jednom dijelu genoma. Ponavljajuća regija samostalno je očitanje koje je zapravo ponavljajuće, ali je zbog greške u sekvenciranju spojeno s drugim očitanjem u kimerno.

Drugi problem vezan je uz prokariotske kromosome koji su tipično cirkularni. Zbog cirkularnosti kromosoma, postoje očitanja koja minimap prepoznaje kao očitanja mapirana na "početak" i "kraj" reference.

Početna pozicija na očitanju	Krajnja pozicija na očitanju	Početna pozicija na referenci	Krajnja pozicija na referenci	Duljina reference
46	11545	4630203	4641628	4641652
11575	19354	6	7676	4641652

Tablica 4.4: Mapiranje očitanja na "rubove" cirkularnog kromosoma

Ta se očitanja čine kao kimerna očitanja mapirana na krajnje udaljene dijelove reference, ali zapravo je to jedinstveno očitanje koje se nalazi na istom dijelu cirkularnog kromosoma i zato se ne smatra kimernim očitanjem nego regularnim.

Treći slučaj su očitanja koja djeluju kao kimerna iako zapravo nisu. To su očitanja s više mapiranja koja gledajući samo pozicije mapiranja na očitanju djeluju kao kimerna. S druge strane, kada se pogledaju pozicije mapiranja na referenci, uz uvjet da su mapiranja na istom relativnom slijedu, vidljivo je da je praznina između mapiranja na očitanju otprilike iste duljine kao praznina između mapiranja na referenci. Iz toga se može zaključiti da takvo očitanje nije zapravo kimerno nego regularno koje iz nekog razloga, najčešće grešaka u sekvenciranju, ima jedan dio koji se lošije mapira na referencu.

I na kraju, postoje očitanja koja kao rezultat minimap2-ovog algoritma imaju više mapiranja koja djeluju kao kimerna, ali među tim mapiranjima postoji najveće mapiranje koje sadrži barem 90% duljine cijelog očitanja. Iako se takvo očitanje u početku čini kao kimerno, ono je zapravo regularno očitanje s manjom greškom zbog koje se sitan dio očitanja mapira na drugi dio reference.

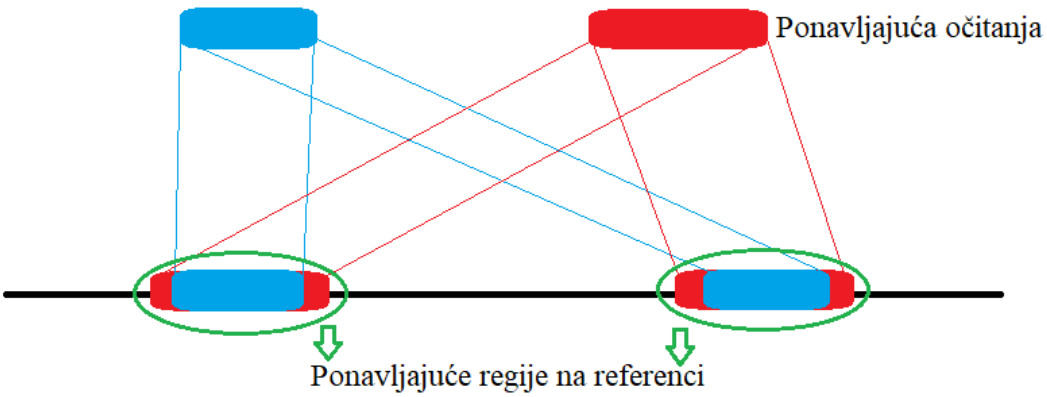
4.4. Detekcija ponavljajućih regija na referenci

Budući da su ponavljajuće regije, kao što im i ime kaže, ponavljajući dijelovi na referenci, za očekivati je da je na jednu regiju potencijalno moguće mapirati više različitih očitanja.

Iz algoritma minimap2 vidljivo je da on, ako postoje, pronalazi višestruka mapiranja pojedinih očitanja. Svako očitanje koje u sebi sadrži regiju koja zapravo pripada ponavljajućoj regiji na referenci će biti mapirana na svaku od tih regija koja je vrlo slična ili ista onom dijelu reference iz kojeg je očitanje zapravo uzeto.

Takva očitanja smo već detektirali i klasificirali kao ponavljajuća očitanja i upravo među njihovim mapiranjima tražimo ponavljajuće regije reference.

U procesu detektiranja ponavljajućih regija na referenci, među ponavljajućim očitanjima tražimo višestruka(barem dvostruka) mapiranja različitih očitanja na iste regije reference te one regije za koje postoje takva mapiranja proglašavamo ponavljajućim.



Slika 4.6: Primjer ponavljajuće regije na referenci detektirane pomoću dva ponavljajuća očitanja koja se mapiraju na isto područje

Poznavajući ponavljajuće regije na referenci i očitanja koja se tamo nalaze te kimerna očitanja, možemo prilikom izgradnje genoma obratiti posebnu pažnju na ta očitanja i te regije kako bismo što uspješnije izbjegli zamke zbog kojih bi sastavljeni genom mogao izgledati vidno drugačije od stvarnog genoma.

Znajući sve to, ulazne FASTA podatke prilagođavamo onome što je detektirano.

4.5. Izlaz programa i sastavljanje genoma

Program datoteku s očitanjima u FASTA ili FASTQ formatu, koja je ulazni podatak programa, parsira uz pomoć biblioteke "bioparser". Iterirajući po vektoru očitanja dobivenom parsiranjem datoteke pomoću jedinstveno identificirajućeg imena očitanja pretražuju se kreirani setovi kimernih i ponavljajućih očitanja te se ovisno o tome je li očitanje detektirano i označeno kao kimerno, ponavljajuće, ponavljajuće sadržano ili je regularno očitanje ono obrađuje u skladu s klasifikacijom. Očitanja se uz informacije o prekidima za kimerna te informacije o ponavljajućim regijama odvajaju u tri FASTA datoteke. Opis obrade i zapisivanja očitanja ovisno o vrsti očitanja je u nastavku.

U prvoj FASTA datoteci nalaze se sva kimerna očitanja. Svako očitanje zapisano je sukladno standardnom FASTA formatu uz neke dodatke. U prvoj linija se osim imena očitanja nalaze i pozicije na očitanju koje su detektirane kao prekidi, odnosno pozicije na kojima su očitanja lažno spojena u jedno očitanje. Datoteka se sprema u direktorij iz kojeg je program pokrenut pod imenom "chimeric_reads.fasta".

U drugoj FASTA datoteci nalaze se ona ponavljajuća očitanja koja se u cijelosti nalaze u nekoj ponavljajućoj regiji na referenci. Ta očitanja ne mogu premostiti ponavljajuću regiju i zbog toga stvaraju probleme prilikom sastavljanja genoma. I ta očitanja osim uobičajenih informacija sadržanih u FASTA formatu imaju pridodane neke dodatne informacije. U prvoj liniji se osim imena očitanja nalaze i pozicije ponavljajućih regija toga očitanja. Datoteka se sprema u direktorij iz kojeg je program pokrenut pod imenom "repeating_reads.fasta".

Treću FASTA datoteku čine sva regularna očitanja, kimerna očitanja razdvojena po pozicijama koje su detektirane kao prekidi te ponavljajuća očitanja koja nisu sadržana u nekoj ponavljajućoj regiji. Prilikom rezanja kimernih očitanja pazilo se na ponavljajuće regije kimernih očitanja te su u datoteku uvrštena samo ona izrezana očitanja koja nisu ponavljajuća ili ako su ponavljajuća se ne nalaze u cijelosti u nekoj ponavljajućoj regiji na referenci. Ponavljajuća očitanja uvrštena u datoteku uz ime sadrže i anotaciju s informacijama o ponavljajućim regijama unutar očitanja. Datoteka s regularnim očitanjima koristi se kao ulaz u alat za sastavljanje genoma Ra, a sprema se u direktorij iz kojeg je program pokrenut pod imenom "cleaned_sequences.fasta".

Ra je asembler koji na temelju ulaznih očitanja u FASTA/FASTQ formatu generira set kontiga visoke točnosti u FASTA formatu (Vaser i Šikić, 2019).

5. Implementacija i korištenje

Cjelokupna implementacija pisana je programskim jezikom C++. Kao ulaz program očekuje datoteku u PAF formatu koja sadrži rezultat mapiranja očitanja na referencu te datoteku u FASTA ili FASTQ formatu koja sadrži početni skup očitanja.

Osim programskog dijela pisanog u C++-u korištene su neke vanjske biblioteke te alati pomoću kojih je izrađeno i testirano programsko rješenje. Neki od tih alata i biblioteka su već spomenuti, a u nastavku je ukupan pregled svih korištenih tehnika te upute za korištenje svih potrebnih komponenti i samoga rješenja.

Za pretprocesiranje podataka korištene su dvije komponente. Prva komponenta je gore spomenuti alat minimap2 koji uzimajući u obzir preddefinirane argumente mapira očitanja na referencu te generira podatke o mapiranjima u PAF formatu.

Ulaz u programski dio rezultat je mapiranja u PAF formatu. Ti su podaci potom parsirani još jednom vanjskom komponentnom, bioparserom. Bioparser je višenamjenski parser bioinformatičkih formata. Osim za parsiranje PAF formata, korišten je i za parsiranje FASTA datoteke kako bi se među svim očitanjima iz FASTA datoteke izdvojila kimerna i ponavljajuća očitanja od onih regularnih. Dobivene PAF i FASTA datoteke se parsirane spremaju u vektore unique_ptr koji pokazuju na strukture u kojima su pohranjene sve potrebne informacije poput imena očitanja i reference, duljina očitanja i reference, pozicija mapiranja, relativnih sljedova i slično te se podaci spremljeni u tom vektoru koriste za detekcije i na kraju generiranja kimernih, ponavljajućih i regularnih FASTA datoteka.

Nakon detekcije i odvajanja kimernih, ponavljajućih i regularnih očitanja, FASTA datoteka s regularnim očitanjima proslijeđena je alatu Ra koji iz pročišćenih očitanja overlap-layout-konsensus pristupom de novo gradi genom te generira set kontiga u FASTA formatu (Vaser i Šikić, 2019).

Naposlijetku, za potrebe testiranja korišten je alat QUAST-LG. QUAST-LG je alat koji evaluira sastavljenost (engl. *assembly*) genoma koristeći razne metrike. Uspoređujući set kontiga izgeneriran alatom Ra s referentom bazom generira niz statistika i grafova koji pokazuju uspješnost sastavljenosti genoma u usporedbi s referencom.

Ovo su upute za korištenje svih navedenih komponenata u operacijskom sustavu Linux.

Za pripremu alata minimap2 za korištenje potrebni su ovi koraci:

```
$ git clone https://github.com/lh3/minimap2
$ cd minimap2 && make
```

Nakon toga je minimap2 spreman za korištenje. Minimap2 kao ulaz očekuje datoteku s očitanjima i datoteku s referencom. Poziva se na ovaj način:

```
$ minimap2 reference.fasta reads.fasta > mapping.paf
```

Nakon toga slijedi detekcija kimernih i ponavljajućih očitanja. Kompletna implementacija nalazi se na platformi Github te se priprema za rad na ovaj način:

```
$ git clone --recursive
https://github.com/lbcb-edu/BSc-thesis-18-19.git
detection
$ cd detection
$ mkdir build
$ cd build
$ cmake -DCMAKE_BUILD_TYPE=Release ..
$ make
```

Detaljne upute za instalaciju, korištenje i te podaci su dostupni na `https://github.com/lbcb-edu/BSc-thesis-18-19/tree/sbakic`, a ovo su upute za pozivanje programa:

```
$ detection mapping.paf sequences.fasta
```

Za instalaciju alata Ra potrebno je izvršiti ove korake:

```
$ git clone --recursive
https://github.com/rvaser/ra.git ra
$ cd ra
$ mkdir build
$ cd build
$ cmake -DCMAKE_BUILD_TYPE=Release ..
$ make
```

Potom se Ra poziva naredbom:

```
$ ra -x {ont, pb} reads.fasta > contigs.fasta
```

6. Testiranje i rezultati

Rezultati su testirani gore spomenutim alatom QUAST-LG. Ono što je zapravo testirano jest kako se kontizi koje Ra izgradi nakon uklanjanja kimernih i ponavljajućih očitanja podudaraju s postojećom referencom. Gledaju se brojne metrike pomoću kojih se ocjenjuje uspješnost sastavljenosti genoma te se generira opći izvještaj o uspješnosti sastavljenosti genoma (Gurevich et al., 2018a).

Kako bi bilo što lakše pratiti rezultate testiranja u nastavku je kratki pregled najvažnijih dijelova QUAST-ovog izvještaja. Izvještaj možemo podijeliti u tri tematska dijela.

Na samom početku nalaze se osnovne informacije o kontizima, broj kontiga, duljinama, udjelima G i C nukleotida u ukupnoj duljina kontiga u odnosu na udjele G i C nukleotida u ukupnoj duljini reference te duljinama i brojevima kontiga koji zadovoljavaju 50, odnosno 75 posto ukupne duljine.

U ovom dijelu izvješća nalaze se i informacije o sastavljenosti genoma, uspješnosti poravnanja kroz informacije o udjelima podudaranja, nepodudaranja i praznina na 100kbp i slični podaci vezani uz poravnate kontige. Iz tih informacije je moguće odrediti mjeru *identity* koja pokazuje udio genoma koji ima ispravno podudaranje nukleotida po poziciji.

Nakon tih podataka slijede informacije o pogrešno sastavljenim dijelovima (engl. *misassemblies*). Taj dio izvješća posebno je zanimljiv jer su pogrešno sastavljeni dijelovi greške najčešće uzrokovane ponavljajućim regijama i kimernim očitanjima, odnosno greškama u sekvenciranju. Pogrešno sastavljanja izuzetno su problematična u vidu korištenja sastavljene reference za daljnja istraživanja (Gurevich et al., 2018b).

QUAST-LG neku poziciju proglašava pozicijom pogrešnog sastavljanja ako zadovoljava jedan od ovih kriterija (Gurevich et al., 2018a):

- lijevo bočno očitanje se poravnava više od 1 kbp (kbp=kilobase pairs) od desnog bočnog očitanja na referenci (relokacija)
- bočna očitanja se preklapaju više od 1kbp (relokacija)
- bočna očitanja se poravnavaju na različite kromosome (translokacija) ili slijedove (inverzija)

QUAST-LG u izvješću pruža podatke o broju pogrešnih sastavljanja detaljizirano podijeljenih na gore definirane relokacije, translokacije i inverzije te podatke o kontizima koji ih sadržavaju.

U zadnjem dijelu su informacije o kontizima koji nisu uspješno poravnati, bilo djelomično ili u potpunosti.

Nakon tabličnog prikaza uspješnosti sastavljenosti genoma slijede grafički prikazi istih vrijednosti od kojih se kao najzanimljiviji izdvajaju slijedeći grafovi:

- graf udjela G i C nukleotida u referenci i u kontizima
- broj pogrešnih sastavljanja po relokacijama, translokacijama i inverzijama
- Feature-Response krivulja koja prikazuje ukupan broj poravnatih baza u kontigu podijeljen s ukupnom duljinom reference ovisno o broju značajki kontiga
- graf kumulativne duljine reference i kumulativne duljine poravnatih kontiga

Grafički prikazi QUAST-LG-ovom izvještaju se baziraju na gore opisanim mjerama. Ipak, treba pojasniti mjeru koja se ne pojavljuje eksplicitno u tabličnom dijelu izvješća, ali pogled na nju može dati uvid u kvalitetu sastavljanja. FRCurve(engl. Feature Response Curve) je mjera koja pokazuje kako rezultat sastavljanja odgovara na skup značajki. U konkretnom slučaju QUAST-LG-ovog izvješća pokazuje udio točno poravnatih baza u kontizima obzirom na broj značajki u kontigu (Gurevich et al., 2018a).

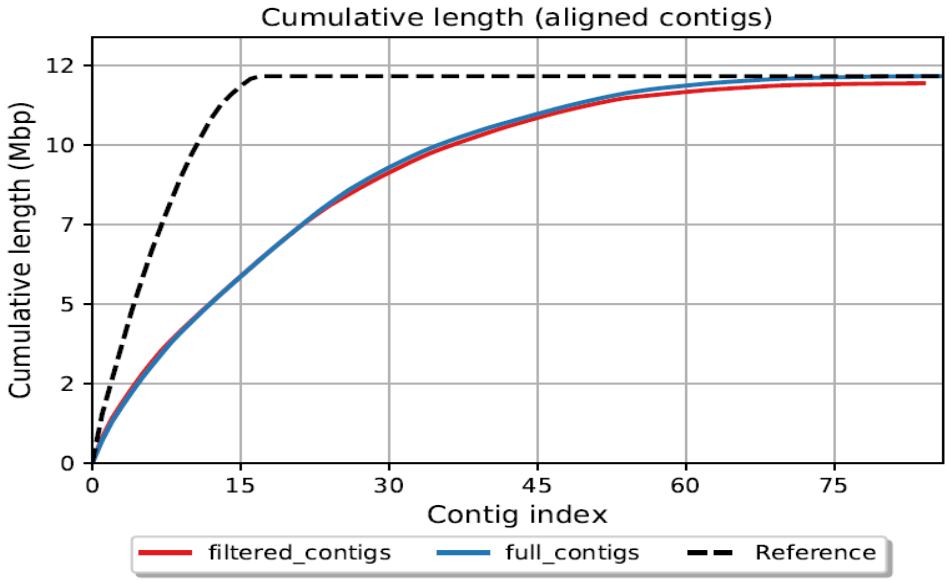
Testovi su rađeni na tri različita skupa podataka, *Saccharomyces Cerevisiae*, *Escherichia Coli* i *Klebsiella Pneumoniae*.

Prvi test rađen je na skupu podataka *Saccharomyces Cerevisiae*. Nakon detekcije kimernih i ponavljajućih očitanja, 1164 kimerna očitanja su izrezana po točkama prekida, 2569 ponavljajućih očitanja koja nisu sadržana niti u jednoj ponavljajućoj regiji na referenci je anotirano po pozicijama ponavljajućih regija unutar očitanja te je 3555 ponavljajućih očitanja koja se u cijelosti sadržana u nekoj ponavljajućoj regiji izbačeno iz ulaznog skupa očitanja. Sastavljena su dva genoma, jedan iz očitanja dobivenih nakon detekcije kimernih i ponavljajućih očitanja te jedan iz svih očitanja bez detekcije te su neki od važnijih rezultata i grafova prikazani.

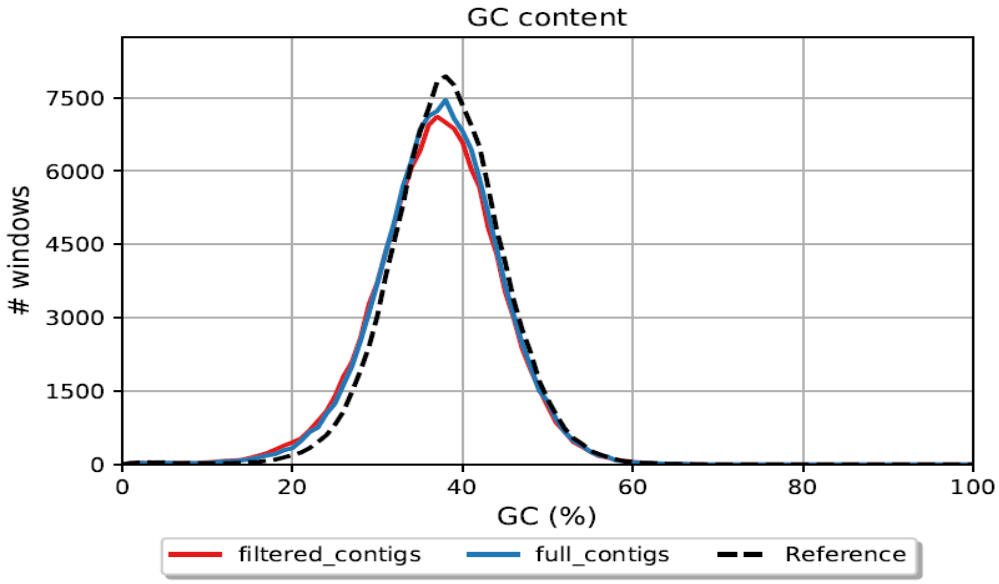
	Filtrirana očitanja	Sva očitanja
broj kontiga	68	69
broj misassembly-ja	16	17
sastavljenost genoma(%)	93.47	95.84
broj neporavnatih kontiga	0	0
identity(%)	98.8	98.9
duljina sastavljenog genoma	11 992 503	12 197 762
duljina ukupnog poravnanja	11 930 259	12 152 908
duljina reference	12 157 105	

Tablica 6.1: Rezultati testiranja za *Saccharomyces Cerevisiae*

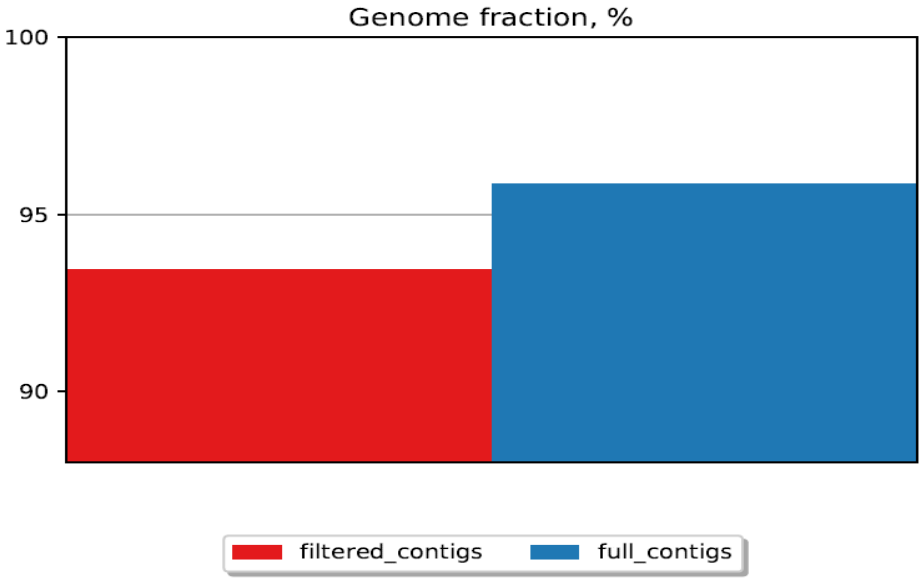
Vidljivo je da je uz manji broj kontiga smanjen i broj pogrešno sastavljenih dijelova, grešaka uzrokovanih ponavljajućim i kimernim očitanjima. Sastavljenost genoma malo je manja nego u slučaju sastavljanja svih očitanja, ali to je potencijalno zbog razlike u pogrešnim sastavljanjima, a identity je približno isto uz malo kraću ukupnu duljinu genoma sastavljenog iz filtriranih očitanja.



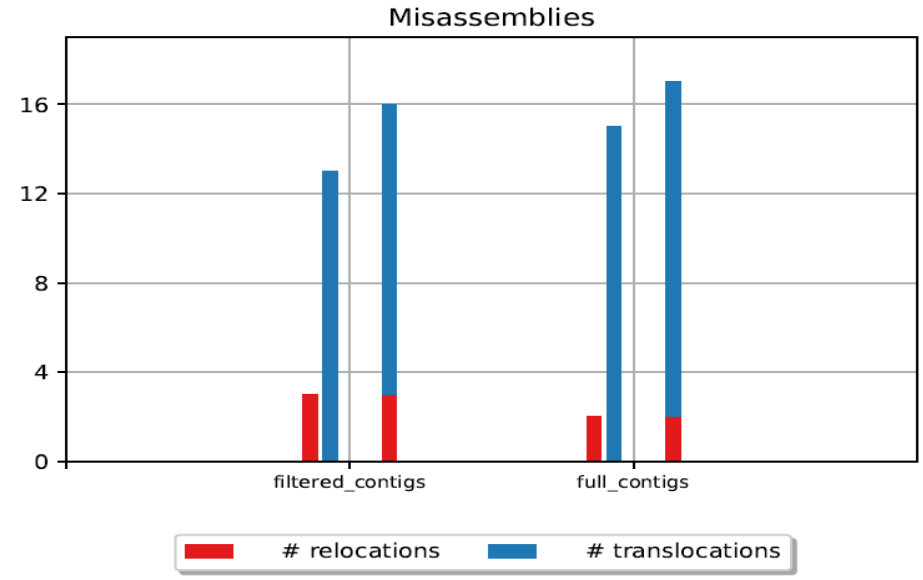
Slika 6.1: Odnos kumulativnih duljina kontiga dobivenih iz filtriranih i nefiltriranih očitanja te usporedba s referencom za *Saccharomyces Cerevisiae*



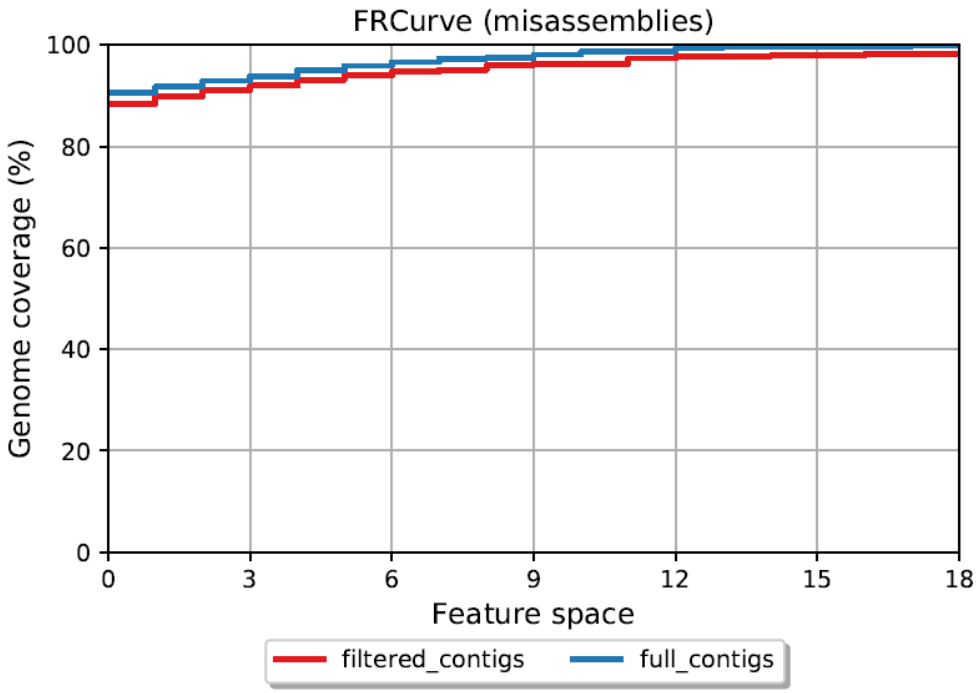
Slika 6.2: Odnos udjela G i C nukleotida u u kontizima dobivenim iz filtriranih i nefiltriranih očitanja te usporedba s referencom za *Saccharomyces Cerevisiae*



Slika 6.3: Odnos sastavljenosti genoma kontiga dobivenih iz filtriranih i nefiltriranih očitanja za *Saccharomyces Cerevisiae*



Slika 6.4: Odnos broja pogrešnih sastavljanja za kontige dobivene iz filtriranih i nefiltriranih očitanja za *Saccharomyces Cerevisiae*



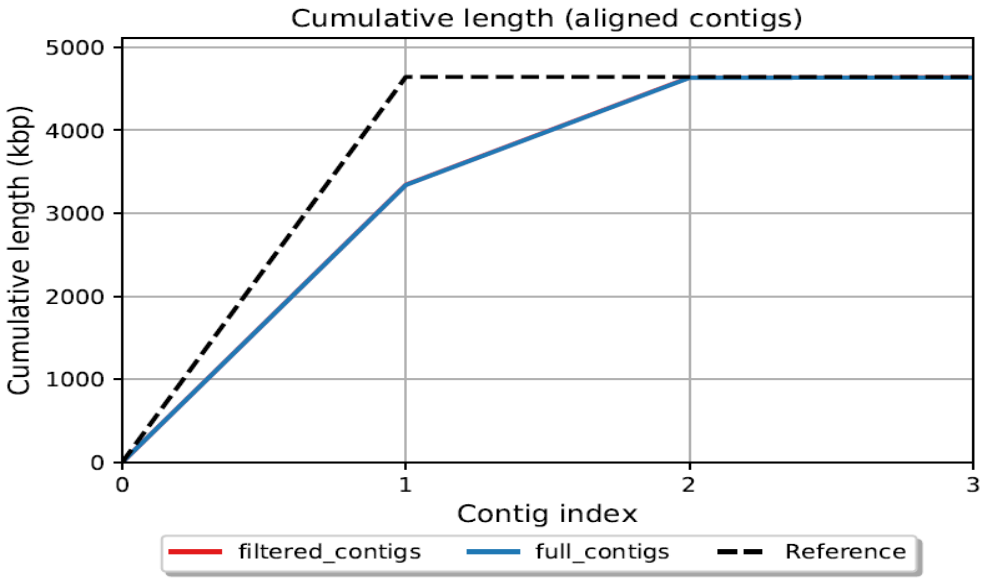
Slika 6.5: Odnos Feature Response-a za kontige dobivene iz filtriranih i nefiltriranih očitanja za *Saccharomyces Cerevisiae*

I iz grafičkih podataka je vidljivo da je uz manji broj kontiga smanjen broj pogrešno sastavljenih dijelova. Ipak, kao problem ostaje malo slabija sastavljenost genoma u smislu duljine te sporija konvergencija potpunoj pokrivenosti genoma vidljiva na grafu Feature Response krivulje.

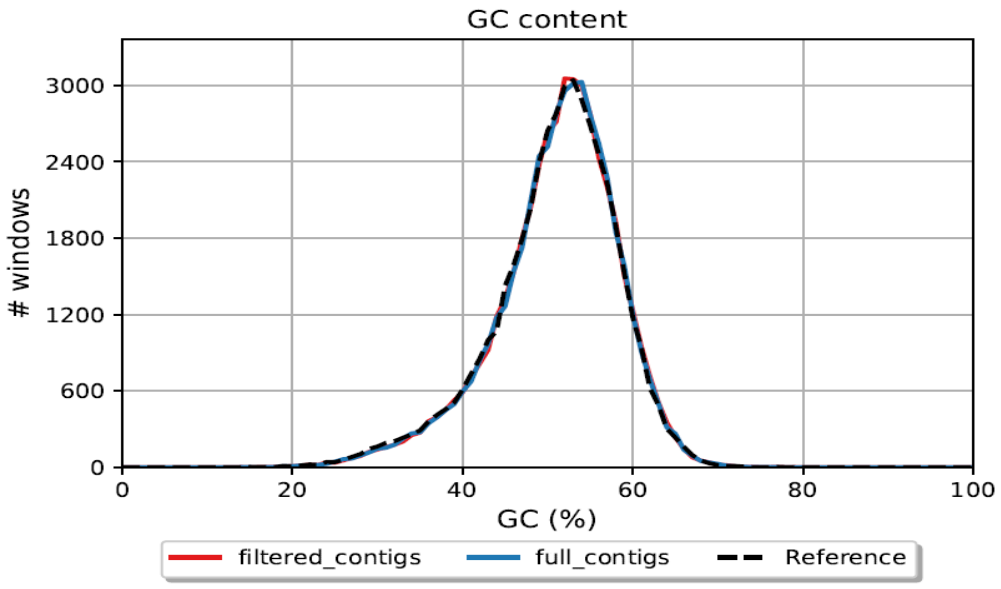
Na isti način je proveden test na skupu podataka *Escherichia Coli*. Iz skupa svih očitanja izdvojeno je 71 očitanje kao kimerno te je na točkama prekida izrezano na nova regularna očitanja, 9 očitanja prepoznato je kao ponavljajuće očitanje sadržano u nekoj ponavljajućoj regiji na referenci i kao takvo je izbačeno iz konačnog skupa očitanja koji je ulaz alatu Ra, a 57 očitanja je klasificirano kao ponavljajuće očitanje koje nije u cijelosti sadržano u nekoj ponavljajućoj regiji na referenci te je anotirano i proslijeđeno sastavljanju genoma. Nakon testiranja alatom QUAST-LG dobiveni su ovi rezultati:

	Filtrirana očitanja	Sva očitanja
broj kontiga	1	1
broj misassembly-ja	2	2
sastavljenost genoma(%)	99.988	99.957
broj neporavnatih kontiga	0	0
identity(%)	99.5	99.5
duljina sastavljenog genoma	4 635 596	4 630 485
duljina ukupnog poravnanja	4 635 595	4 630 463
duljina reference	4 641 652	

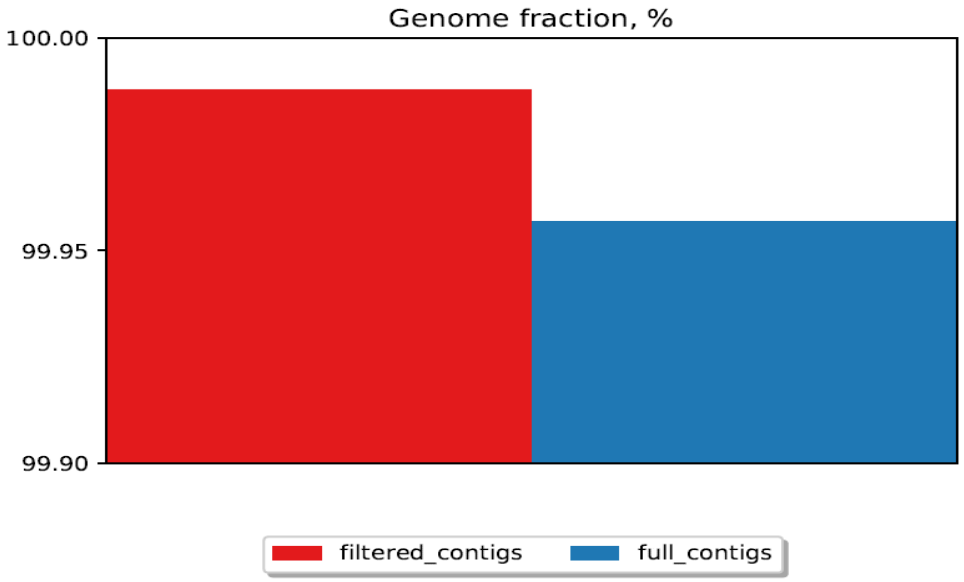
Tablica 6.2: Rezultati testiranja za *Escherichia Coli*



Slika 6.6: Odnos kumulativnih duljina kontiga dobivenih iz filtriranih i nefiltriranih očitanja te usporedba s referencom za *Escherichia Coli*

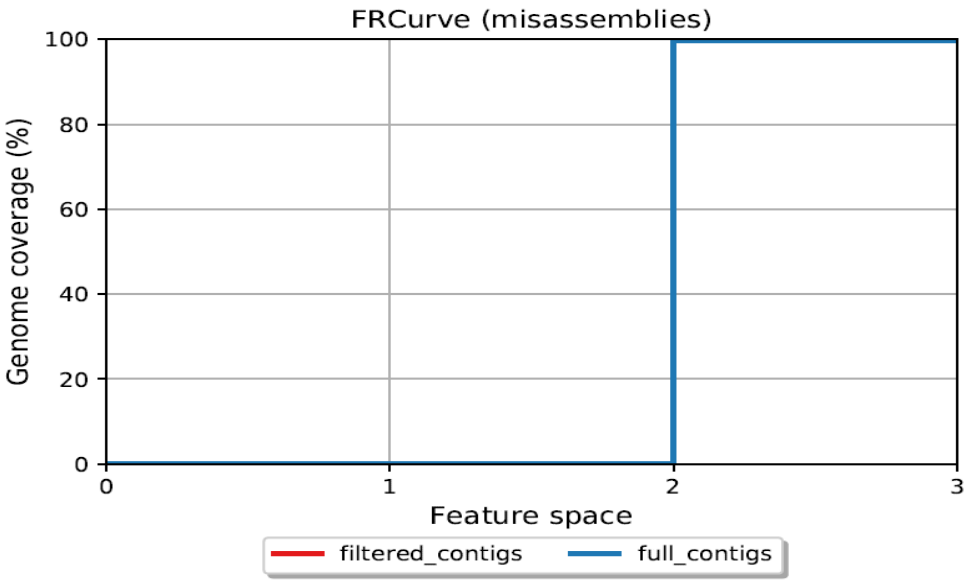


Slika 6.7: Odnos udjela G i C nukleotida u u kontizima dobivenim iz filtriranih i nefiltriranih očitanja te usporedba s referencom za *Escherichia Coli*



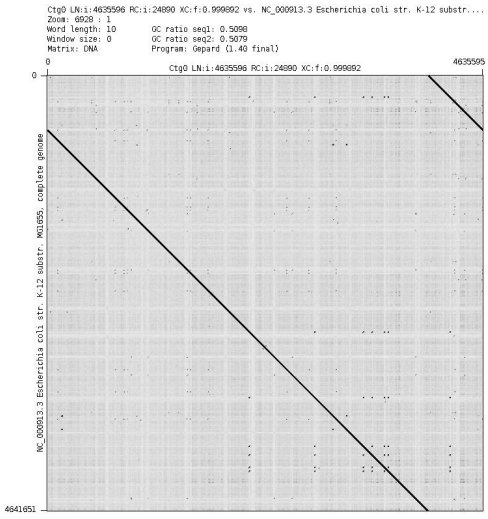
Slika 6.8: Odnos sastavljenosti genoma kontiga dobivenih iz filtriranih i nefiltriranih očitanja te usporedba s referencom za *Escherichia Coli*

Slika 6.9: Odnos broja pogrešnih sastavljanja za kontige dobivene iz filtriranih i nefiltriranih očitanja za *Escherichia Coli*

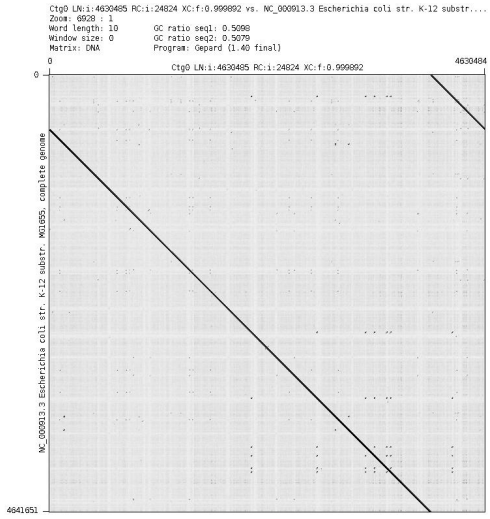


Slika 6.10: Odnos Feature Response-a za kontige dobivene iz filtriranih i nefiltriranih očitanja za *Escherichia Coli*

Rezultati ovoga testiranja pokazuju da se broj kontiga prije i nakon filtriranja nije promijenio. Ipak, treba imati na umu da je *Echerichia Coli* relativno "malena" (samo 4641652 baza) te da je prepoznat malen broj kimernih (71) i ponavljajućih (66) očitanja. Sastavljenost genoma malo je bolja nakon filtracije, a tvrdnja da je genom zaista dobro sastavljen, usprkos dva pronađena pogrešna sastavljanja može se potvrditi uspoređivanjem sastavljenog genoma i postojeće reference u alatu gepard. Gepard kao ulaz očekuje dvije datoteke u FASTA ili FASTQ formatu sa sastavljenim genomima pomoću kojih iscrtava točkasti graf s pozicijama jednog genoma na x-osi te pozicijama drugog genoma na y-osi. Dobro podudaranje zapravo znači poklapanje na istoj poziciji što je praktično pravac nagnut pod kutom od 45° u odnosu na x-os (Krumsiek et al., 2007).



(a) Poravnanje genoma izgrađenog iz filtriranih očitanja na referencu



(b) Poravnanje genoma izgrađenog iz svih očitanja na referencu

Uz jednake iznose identity-ja od 99.5% iz slika je vidljivo da se sastavljeni genomi i postojeća referenca gotovo savršeno poklapaju stoga se da zaključiti da su prepoznata pogrešna sastavljanja neke manje greške koje nisu bitno utjecale na krajnji izgled genoma.

Posljednji test rađen je na skupu podataka *Klebsiella Pneumoniae*. Iz početnog skupa podataka je detektirano i po pozicijama prekida izrezano 5345 kimernih očitanja. 721 ponavljajuće očitanje se nije nalazilo niti u jednoj ponavljajućoj regiji na referenci te je anotirano i dodano u skup očitanja za sastavljanje, a 272 očitanja su bila ponavljajuća i u potpunosti sadržana u nekoj ponavljajućoj regiji na referenci. Testiranje je dalo ovakve rezultate:

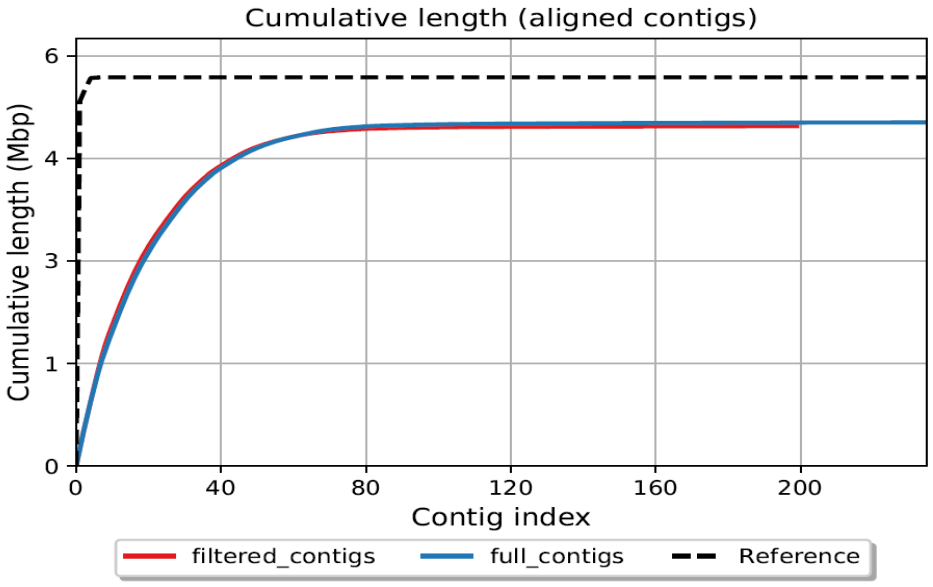
	Filtrirana očitanja	Sva očitanja
broj kontiga	16	9
broj misassembly-ja	71	85
sastavljenost genoma(%)	87.29	86.94
broj neporavnatih kontiga	0	0
identity(%)	98.687	98.686
duljina sastavljenog genoma	5 481 106	5 559 831
duljina ukupnog poravnanja	4 968 356	5 024 577
duljina reference	5 682 322	

Tablica 6.3: Rezultati testiranja za *Klebsiella Pneumoniae*

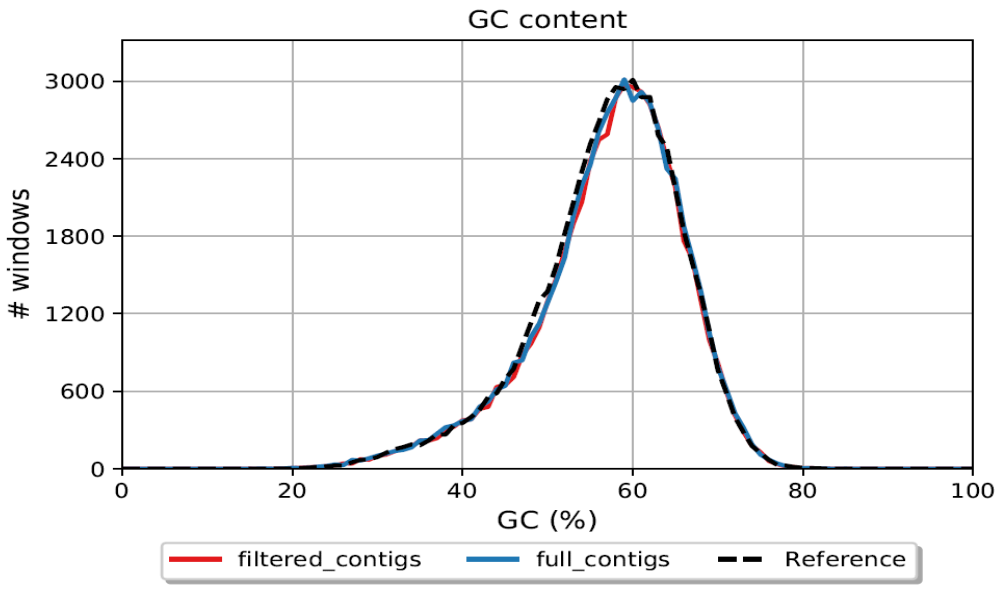
Skup podataka za *Klebsiella Pneumoniae* dao je rezultate koji nakon filtriranja podataka imaju više kontiga te malu razliku u sastavljenosti genoma. S druge strane, broj pogrešnh sastavljanja nakon filtriranja vidljivo je manji u odnosu na originalni skup očitanja.

Postoji više potencijalnih razloga za ovakve rezultate. Jedan od mogućih razloga slabija je kvaliteta referentne baze podataka ili baze podataka očitanja zbog čega prilikom mapiranja očitanja na referencu dolazi do, u prosjeku, lošijih mapiranja i zbog toga se više očitanja detektira kao kimerna očitanja što kasnije može stvarati probleme prilikom sastavljanja genoma.

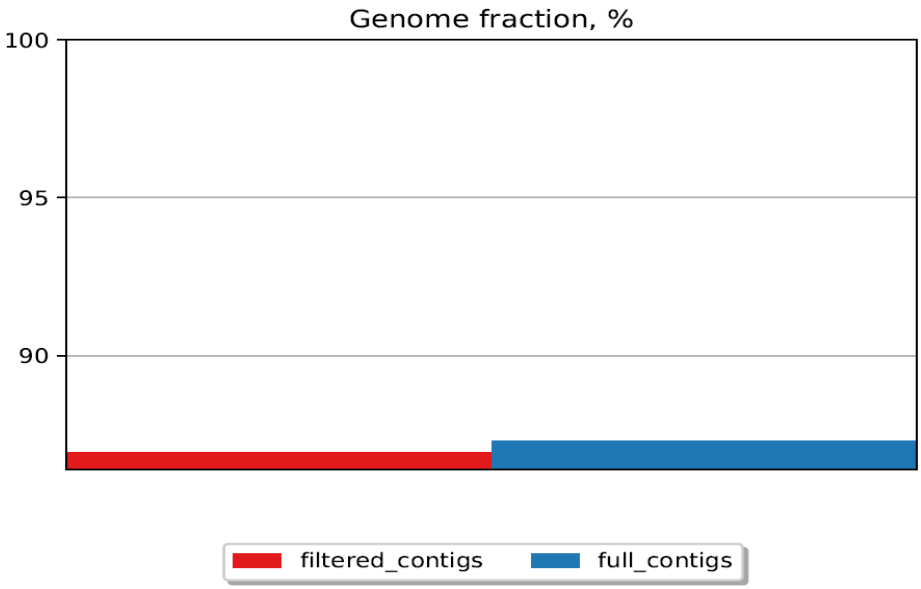
Ipak, ono što je pozitivni rezultat ovoga testa jest činjenica da je, usprkos povećanju broja kontiga, smanjen broj pogrešno sastavljenih dijelova u izgrađenom genomu.



Slika 6.12: Odnos kumulativnih duljina kontiga dobivenih iz filtriranih i nefiltriranih očitanja te usporedba s referencom za *Klebsiella Pneumoniae*

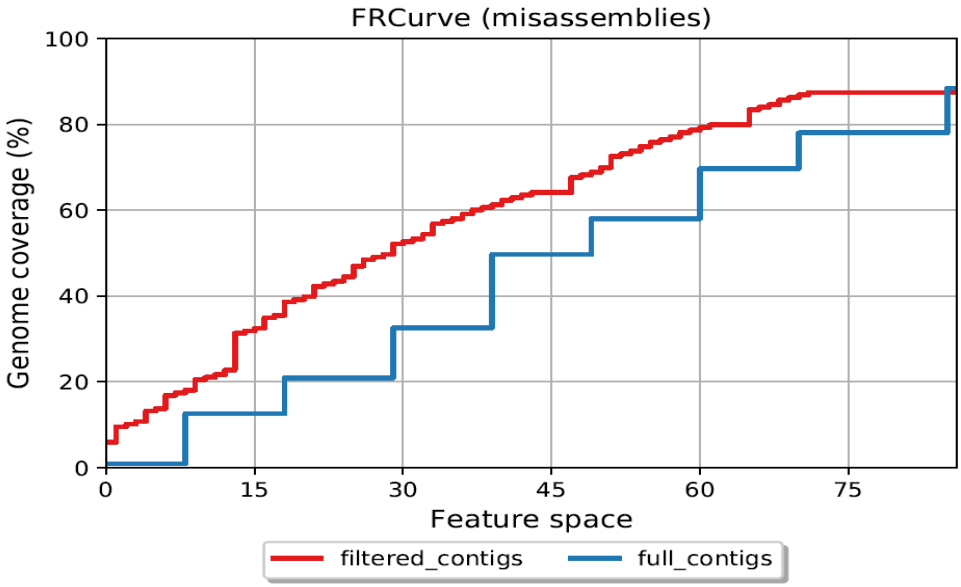


Slika 6.13: Odnos udjela G i C nukleotida u u kontizima dobivenim iz filtriranih i nefiltriranih očitanja te usporedba s referencom za *Klebsiella Pneumoniae*



Slika 6.14: Odnos sastavljenosti genoma kontiga dobivenih iz filtriranih i nefiltriranih očitanja te usporedba s referencom za *Klebsiella Pneumoniae*

Slika 6.15: Odnos broja pogrešnih sastavljanja za kontige dobivene iz filtriranih i nefiltriranih očitanja za *Klebsiella Pneumoniae*



Slika 6.16: Odnos Feature Response-a za kontige dobivene iz filtriranih i nefiltriranih očitanja za *Klebsiella Pneumoniae*

Obzirom da je većina grafičkih prikaza već objašnjena, kao posebno zanimljiv ostaje graf Feature Response krivulje. Naime, na grafu je vidljivo da za filtrirana očitanja prekrivenost genoma ima puno bolji odziv u odnosu na kontige sastavljene iz svih očitanja za isti broj značajki.

6.1. Daljnji rad

Kroz ovih nekoliko testova pokazano je da za de novo sastavljanje genoma postoje određeni benefiti u pronalasku i anotaciji kimernih i ponavljajućih očitanja.

Pokazalo se, doduše, da postoje i određeni problemi na koje ovakav način detekcije nailazi ako su podaci za mapiranje lošije kvalitete.

Kako bi detekcija što više pomogla prilikom sastavljanja genoma, bilo bi idealno osim podataka o mapiranjima očitanja na postojeću referencu koristiti još neke informacije o samim očitanjima i na taj način još kvalitetnije selektirati stvarna kimerna i ponavljajuća očitanja među ostalim očitanjima.

7. Zaključak

Bioinformatika kao interdisciplinarno područje okuplja širok spektar znanja i stručnjaka koji provode kompleksna istraživanja s ciljem pronalaska što točnijih i konkretnijih informacija vezanih uz bioznanost. Bioznanost kao vrlo osjetljivo područje zahtijeva podatke koji su vrlo visoke kvalitete i preciznosti.

S padom cijena istraživanja počeo je ubrzani razvoj s novim metodama sekvenciranja. To je omogućilo veći broj istraživanja, ali je uzrokovalo i veću stopu pogrešaka u procesu sekvenciranja. Problemi odabira odgovarajućih očitanja na ponavljajućim područjima genoma te izbjegavanje pogrešnog sastavljanja genoma uslijed kreiranja "umjetnih" očitanja, problemi su koji nisu izbjegnuti u novoj generaciji sekvenciranja. Takvi problemi nameću se kao ugrožavajući faktor za kvalitetu sastavljenih genoma, a samim time i kvalitetu istraživanja kojima se podvrgavaju tako dobiveni genomi.

Kada bi se i takvi problemi uspješno rješavali, primjena rezultata istraživanja bila bi još veća u svakodnevici te je zato vrlo bitno razvijati bioinformatičke alate i algoritme u smjeru što veće preciznosti.

Upravo zato važno je raditi na alatima i metodama koje će efikasno i precizno detektirati kimerna i ponavljajuća očitanja te koristiti znanja o njima kako bi novo sastavljeni genom bio što točniji i upotrebljiviji u daljnjim testiranjima i istraživanjima. Takvi alati u suradnji s alatima za mapiranja i sastavljanje genoma mogu dovesti do genoma visoke kvalitete i upotrebljivosti što se nameće kao vrlo poželjan ishod.

Ovaj rad pokazuje kako detekcija kimernih i ponavljajućih očitanja daje rezultate prilikom sastavljanja genoma. Uz detekciju i kvalitetno iskorištavanje informacija o kimernim i ponavljajućim očitanjima moguće je smanjiti broj pogrešno sastavljenih dijelova genoma te tako popraviti kvalitetu sastavljanja. Naravno, kako bi rezultati detekcije bili što preciziniji potrebno je nadalje razvijati metode i testirati pristupe detekciji takvih očitanja.

LITERATURA

atdbio. Next generation sequencing, 2011. URL https://www.atdbio.com/content/58/Next-generation-sequencing.

Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, i Glenn Tesler. Quast 5.0.2 manual, 2018a. URL http://quast.bioinf.spbau.ru/manual.html#sec3.

Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, i Glenn Tesler. Quast: quality assessment tool for genome assemblies, 2018b. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624806/.

Amy M. Hauth i Deborah A. Joseph. Beyond tandem repeats: complex pattern structures and distant regions of similarity, 2002. URL https://academic.oup.com/bioinformatics/article/18/suppl_1/S31/231738.

Mile Šikić i Mirjana Domazet-Lošo. *Bioinformatika*. 2013. URL https://www.fer.unizg.hr/_download/repository/bioinformatika_skripta_v1.2.pdf.

J. Craig Venter Institute. Genome sequencing, 2003. URL http://www.genomenewsnetwork.org.

Jan Krumsiek, Roland Arnold, i Thomas Rattei. Gepard: a rapid and sensitive tool for creating dotplots on genome scale, 2007. URL https://academic.oup.com/bioinformatics/article/23/8/1026/198110.

Arthur M. Lesk. *Bioinformatics*. 2013. URL https://www.britannica.com/science/bioinformatics.

Heng Li. Minimap2: pairwise alignment for nucleotide sequences, 2018a. URL https://github.com/lh3/minimap2.

Heng Li. Paf: a pairwise mapping format, 2018b. URL https://github.com/lh3/miniasm/blob/master/PAF.md.

Robert Vaser i Mile Šikić. Yet another de novo genome assembler, 2019. URL https://www.biorxiv.org/content/10.1101/656306v1.article-metrics.

De novo sastavljanje genoma vođeno referencom

Sažetak

De novo sastavljanje genoma jedan je od najkompleksijih problema u bioinformatici. Brojni su problemi prilikom sastavljanja genoma, uzrokovani, kako biološkim specifičnostima, tako i greškama u procesu sekvenciranja. Probleme stvaraju takozvana ponavljajuća i kimerna očitanja. U ovom radu prikazano je kako detektiranje kimernih i ponavljajućih očitanja uz prikladna postupanja s njima pomaže ispravnom sastavljanju genoma. Prikazane su poteškoće na koje se nailazi prilikom sastavljanja genoma te kako uz postojeće alate minimap2 i Ra i ispravnu detekciju kimernih i ponavljajućih očitanja što bolje sastaviti genom te rezultati testiranja sastavljenosti genoma. Osim toga, pojašnjeni su i formati korištenih podataka te temeljne značajke korištenih alata.

Ključne riječi: kimerno, onavljajuće, genom, minimap2, ra, FASTA, FASTQ, PAF, bioinformatika.

Reference-Guided de novo Genome Assembly

Abstract

De novo genome assembly is one of the most complex problems in Bioinformatics. There are many problems, either caused by biological specificities or mistakes made during sequencing process, that complicate de novo genome assembly process. The problems are mainly caused by chimeric and repeating reads. This thesis describes how detecting and annotating chimeric and repeating reads helps the assembly process. Typical issues that occur during the assembly process and how successful existing tools Ra and minimap2 with the detection of chimeric and repeating reads assembly the reads are shown. Besides that, thesis describes file formats used in this work and thorough features of used tools

Keywords: chimers, repeats, genome, minimap2, Ra, FASTA, FASTQ, PAF, Bioinformatics