UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

MASTER RAD

Integracija rezultata različitih programa za sklapanje genoma

Autor:
Anamaria PIRI

Mentor:
dr Jovana KOVAČEVIĆ, docent

Članovi komisije:

dr Nevena Veljković, naučni savetnik, INN Vinča, Univerzitet u Beogradu

dr Saša Malkov, vanredni profesor, Matematički fakultet, Univerzitet u Beogradu

Beograd, 2019

UNIVERZITET U BEOGRADU

Matematički fakultet

Sažetak

Integracija rezultata različitih programa za sklapanje genoma

Anamaria PIRI

Danas je u upotrebi na desetine programa za sklapanje genoma, asemblera, koji su često zasnovani na De Brujinovim grafovima. Cilj ovog master rada je integracija rezultata različitih asemblera radi što kvalitetnijeg rekonstruisanja DNK sekvence. Praktični deo rada uključuje konfigurisanje alata za integraciju GAM NGS za rad sa odabranim asemblerima: Velvet, ABySS i SPAdes. Performanse kombinovanog asemblera su upoređene sa pojedinačnim alatima. Testiranje je izvedeno na sirovim podacima preuzetim iz repozitorijuma sekvenci SRA i rezultati su upoređeni sa odgovarajućim referentnim genomom. Za prikaz rezultata je implementirana grafička reprezentacija. Aplikacija je razvijena upotrebom programskog jezika Python.

Sadržaj

Poglavlje 1

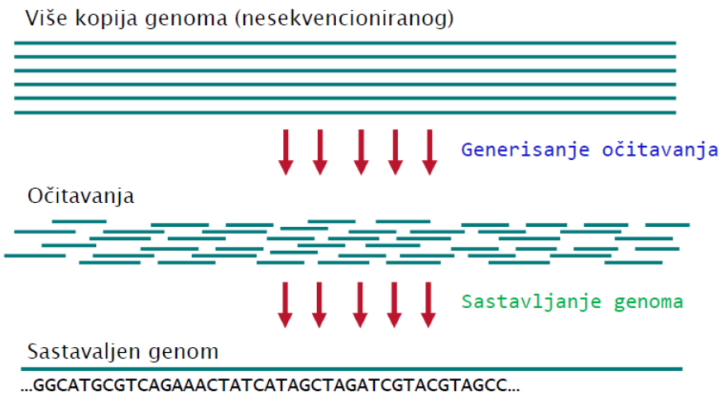
Uvod

1.1 Sekvenciranje i sklapanje genoma

1.1.1 Sekvenciranje celokupnog genoma

Genom je celokupan genetski materijal jednog organizma. Kod većine organizama genetski materijal je sadržan u DNK sekvenci, što je kompletna lista nukleotida (A, C, G i T za genome DNK) koji čine sve hromozome jedne jedinke. Sa računarske strane posmatrano, ovaj niz se može predstaviti kao niska karaktera nad azbukom {A, C, G, T}. Određivanje redosleda nukleotida u DNK sekvenci izvodi se tehnologijom sekvenciranja, dok se sklapanje delova DNK sekvence u genom ili druge manje (egzon, hromozom) celine vrši pomoću bioinformatičkih alata.

Moderne mašine za sekvenciranje ne mogu da pročitaju ceo genom, redom nukleotid po nukleotid, od početka do kraja (kao što bismo pročitali knjigu). Umesto toga, sekvencer generiše kratke podniske genoma, takozvana očitavanja, iz velikog broja kopija genoma. Ove mašine obično isecckaju genom na nasumičnim mestima, na kraće delove, i na osnovu ovih delova generišu kratka očitavanja. Ilustracija sekvenciranja je prikazana na slici 1.1. Sekvenceri generišu kolekciju podniski, koje treba spojiti, i tako rekonstuisati originalnu DNK sekvencu. Sklapanje možemo da posmatramo kao računarski problem rekonstrukcije niske na osnovu ovih kratkih očitavanja.

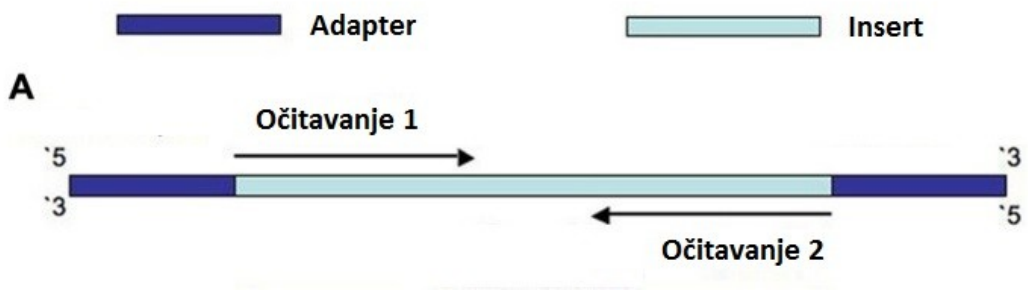


SLIKA 1.1: Ilustracija sekvenciranja

Izvor: [1] str. 50, slika 3.4

1.1.2 Sekvenciranje nove generacije

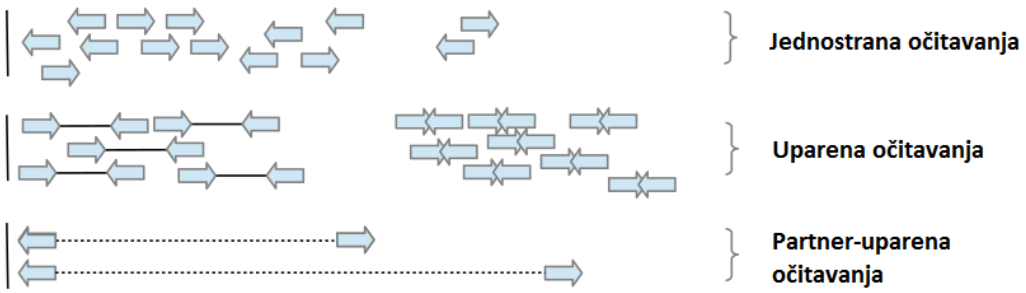
Sekvenciranje nove generacije, NGS (*Next Generation Sequencing*) predstavlja najmoderniju metodu za sekvenciranje genoma [2]. Tehnologije za sekvenciranje ne mogu pročitati celokupne genome odjednom, nego samo kratke delove, koje nazivamo očitavanja. Sekvenceri na nasumično odabranim pozicijama fragmentišu DNK na kratke delove i dodaju takozvane *adapter* sekvence na krajeve. Adapter sekvence su kratke hemijski sintetizovane sekvence nukleotida koje se mogu vezati za krajeve nepoznatih DNK sekvenci i neophodne su u nekim koracima sekvenciranja. Fragmenti dobijeni posle sekvenciranja se sastoje od tri dela: jedna podniska DNK sekvence (koja se naziva *insert*) i dve adapter sekvence (slika 1.2) [3].



SLIKA 1.2: Dobijeni fragmenti se sastoje od inserta i dve adapter sekvence

Izvor: [4] str. 2, slika 2

Sekvenceri pročitaju deo fragmenta bez adaptera do određene, unapred zadate dužine koja predstavlja *dužinu očitavanja*. Čitanje istog dela fragmenta se vrši veći broj puta kako bi se dobila potrebna *dubina pokrivanja*. Kažemo da je dubina pokrivanja neke pozicije u DNK sekvenci velika ako je nukleotid na toj poziciji pročitan veliki broj puta u jedinstvenim očitavanjima. Ovako pročitane delove, tipične dužine 50-400 baznih parova nazivamo kratka očitavanja (*short reads*), dok u slučaju očitavanja u oba smera nastaju očitavanja sa uparenim krajevima (*paired-end reads*), skraćeno uparena očitavanja. Uparena očitavanja mogu da budu različitih dužina. Posebna vrsta uparenih očitavanja su partner-uparena očitavanja. Ona obuhvataju veće regione, veličine 2-20 hiljada baznih parova. Sve vrste očitavanja možemo da vidimo na slici 1.3.



SLIKA 1.3: Vrsta očitavanja

Izvor: [5] str. 1030, slika 2

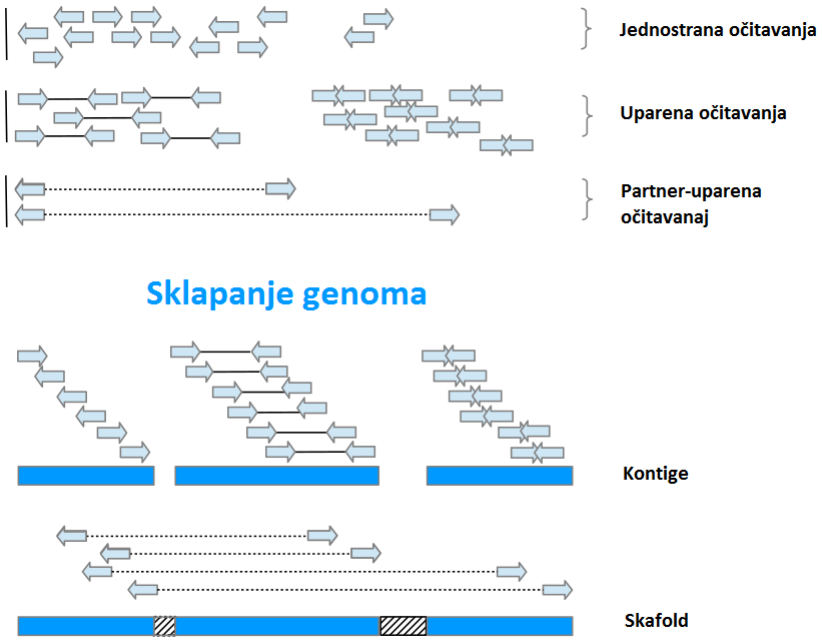
Za NGS tehnologije je karakteristično da generišu vrlo veliki broj očitavanja, pa je za bilo koji organizam moguće dobiti visoko pokrivanje genoma očitavanjima.

1.1.3 Proces sklapanja

Sklapanje DNK sekvenci podrazumeva poravnanje i integrisanje očitavanja radi rekonstrukcije originalne DNK sekvence. *De novo* asembleri su programi koji vrše sklapanje tako što proširuju kratka očitavanja spajanjem susednih očitavanja u dužu sekvencu, bez korišćenja referentne sekvence. Očitavanja se na ovaj način proširuju u kontinuiranu sekvencu, takozvanu kontigu (*contig*). Rezultat sklapanja u obliku kontiga je ilustrovan na slici 1.4. Kontige obično predstavljaju jednu konsenzus nisku[1] i ne sadrže nijedan polimorfizam[2]. Povezivanjem kontiga pomoću partner-uparenih očitavanja dobijemo skafolde (*scaffolds*). U skafoldima između kontiga su praznine, koje odgovaraju neuspešno pročitanim delovima genoma. Redosled povezanih kontiga u skafoldima odgovara redosledu kojim su kontige prisutne u genomu.

[1] niska sastavljena od najfrekventnijih nukleotida na pozicijama poravnatih sekvenci

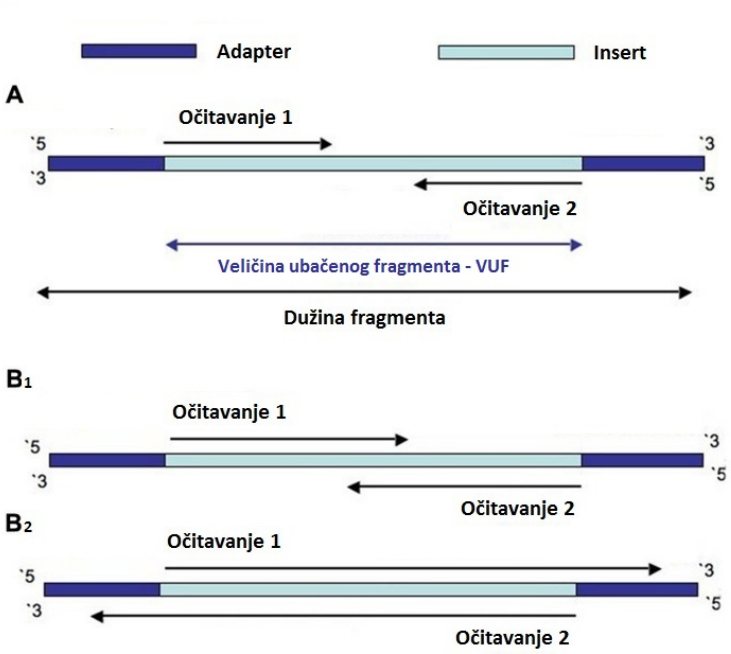
[2] DNK polimorfizam je razlika u sekvenci nukleotida raznih skupova hromozoma u diploidnim i poliploidnim organizmima. Može da bude jednonukleotidni polimorfizam, delecija, insercija i slično



SLIKA 1.4: Proces sklapanja

Izvor: [5] str. 1030, slika 2

Svaki asembler teži da dobije što tačniji rezultat i što duže kontige (što kontinualniji rezultat (*contiguity*)), a time i potpuniju reprezentaciju DNK sekvenci. Većina asemblera opredeljuje se za dobijanje dužih kontiga čak i na račun smanjene tačnosti [2]. Kada želimo duže kontige, onda je za asemblere koje rade sa kratkim očitavanjima važno da je pokrivenost očitavanjima visoka. U suprotnom, rezultat sklapanja je veoma fragmentisan. Kada na raspolaganju postoje uparena očitavanja može da se obezbedi veći kvalitet i duže kontige. Uparena očitavanja pružaju informacije za razrešavanje konflikata između moguće veze kontiga i samim tim spajanje kontiga u duže. Dužine praznina između kontiga (dužine delova genoma koji nisu sklapani) procenjuju se pomoću očekivane veličine ubačenih fragmenata (VUF - *insert size*), pri čemu se pod ubačenim fragmentom podrazumeva deo fragmenta između adaptera. Ilustracija VUF-a prikazana je na slici 1.5. Ova veličina zavisi od koncentracije korišćenog DNK uzorka [4]. VUF je promenljiva i obično na širokom intervalu varira. Specijalno, u nekim slučajevima je dužina očitavanja veća od ubačenog fragmenta pa se u očitavanju dobija i deo adaptera sa suprotne strane. Zbog toga prvi korak u asembliranju mora da bude traženje ostataka adaptera u očitavanju i njihovo uklanjanje, što za posledicu može da ima skraćivanje nekih očitavanja ili njihovo odbacivanje (slika 1.5 pod B_2).



SLIKA 1.5: Veličina ubačenog fragmenta (VUF):
(A) Fragment sa ubačenim fragmentom koji je duži nego očitavanja
(B_1 i B_2) Fragment sa ubačenim fragmentom koji je kraći nego očitavanja

Izvor: [4] str. 2, slika 2-3

Sekvenciranje i sklapanje celog genoma omogućavaju ([5]):

- Posmatranje na koji način i koliko neki organizmi mogu da se prilagode okruženju u kom se nalaze (*selectively important variation*).
- Razmatranje osnova fenomena koji se javljaju kod parenja životinja u bliskom srodstvu (*inbreeding*).
- Istraživanja na polju personalizovane medicine, jer predstavljaju osnovu za predviđanje osetljivosti pojedinca na neke bolesti i njegove reakcije na određene lekove.
- Određivanje jednonukleotidnih polimorfizama (*SNP - single nucleotide polimorfism*) pomoću kojih se mogu detektovati funkcionalne varijacije u genomu što je od značaja za istraživanja u evolucionoj biologiji.

1.1.4 Izazovi prilikom sklapanja sekvence genoma

Idealno, reprezentaciju genoma sačinjava kolekcija parova simbola iz azbuke {A, C, G, T} (parovi simbola odgovaraju baznim parovima iz DNK sekvenci) sa svih hromozoma izabranog organizma, nalik na mapu genetskog sadržaja. U realnosti postoje različiti izazovi koji stoje na putu određivanja mape genetskog sadržaja. Prvenstveno, zbog individualnih genetskih varijacija ni za

jednu vrstu živih bića DNK sekvenca nije jedinstvena u svim njenim primercima. Razlog tome mogu biti insercije i delecije (*indeli*), polimorfizmi, varijacije broja kopija gena ili druge promene na kratkim DNK segmentima. Štaviše, ćelije istog organizma mogu da imaju razlike u genetskom sadržaju zbog telesnih mutacija (*somatic mutation*) [5]. Iz tog razloga, rezultat sklapanja genoma predstavlja samo jednu konsenzus niska svakog hromozoma.

Pored manjkavosti predstavljenja genoma postoje i druge činjenice koje predstavljaju izazov i za naprednije algoritme, čak i u slučaju velikih projekata. Sekvenciranje pomoću današnjih algoritama, zbog prethodnih izazova i zbog komplikovane strukture same DNK sekvence (kao što je na primer nesavršeno ponavljanje[3]), nije savršeno.

1.2 *De novo* sklapanje

Postoji veliki broj aplikacija za *de novo* sklapanje celog genoma, odnosno za sklapanje genoma bez referentne sekvence, i novi programi se stalno pojavljuju ([6, 8, 9]). Ovi asembleri obično koriste kao osnovnu ideju jedan od sledećih algoritama: graf preklapanja (*overlap graph, extension-based method*) ili De Brojnov graf sa Ojlerovom putanjom.

1.2.1 Vrsta *de novo* sklapanja

1. Graf preklapanja (*overlap graph, extension-based method*): U prvu grupu spadaju programi koji koriste graf preklapanja. U OLC[4] strategiji najpre se izračuna preklapanje između svaka dva očitavanja i predstavi se grafom u kom svaki čvor predstavlja jedno očitavanje, a grane postoje između čvorova koji se preklapaju [10]. Ovi algoritmi su najefikasniji kada je mali broj očitavanja među kojima je visok stepen preklapanja, u suprotnom su skloni greškama. Rezultati pokazuju da rade veoma loše i sa malim procentom grešaka u očitavanjima. Čak i bez grešaka u očitavanjima pojavljaju se problemi kod regiona koji se ponavljaju ili sadrže visok nivo polimorfizama.

2. De Brojnov graf sa Ojlerovom putanjom: Druga grupa softvera obuhvata one koji su bolje prilagođeni sklapanju kratkih očitavanja. *De novo* sklapanje genoma je postupak sastavljanja DNK sekvence kada nije na raspolaganju referentna sekvenca [6, 9]. Ova grupa asemblera koristi algoritme zasnovane na De Brojnovom grafu. Problem rekonstrukcije niske moguće je rešiti na dva načina pomoću ove strukture. Jedan je pronalaženje Hamiltonove putanje, a drugi je pronalaženje Ojlerove putanje u De Brojnovom grafu [10]. Metode koje koriste De Brojnov

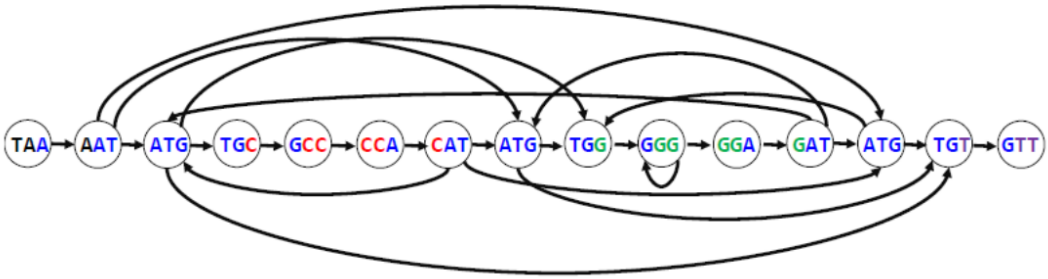
[3]Ponavljanjem u DNK sekvenci nazivamo situaciju kada se šablon od nekoliko nukleotida više puta pojavljuje u nizu. Savršena ponavljanja su manje verovatna, ponovci se obično razlikuju jedan od drugog na nekim pozicijama

[4]OLC je skraćeno od *overlap layout consensus*, gde *overlap* označava izgradnju grafa preklapanja, *layout* spajanje putanja u grafu u kontige, a *consensus* izbor najverovatnije sekvence nukleotida za svaku kontigu [10]

graf obično zahtevaju veliku količinu RAM memorije. U zavisnosti od količine podataka posle sekvenciranja, za sklapanje nekih genoma je ponekad potreban i terabajt unutrašnje memorije. Ideja za sklapanje kratkih očitavanja koja se najčešće sreće jeste korišćenje De Brojnovih grafova kod kojih su očitavanja podeljena u k-grame (podniske očitavanja dužine k, obično $25 - 50$ baznih parova). Svako očitavanje dužine n je prelomljeno na $(n - k + 1)$ preklapajućih k-grama koje dobijamo pomeranjem prozora dužine k duž ulazne sekvence. K-gramski sastav predstavlja kolekciju podniski polazne sekvence pri čemu su sve podniske dužine k i u kolekciju su uključeni duplikati. K-grami su kraći nego očitavanja i kao takvi su pogodniji za heširanje (*hashing*), čime su i operacije nad De Brojnovim grafom manje računarski zahtevne. Zbog toga je većina asemblera, koji sprovode ovu strategiju, zasnovana na ovom pristupu.

1.2.2 Detaljnije o rekonstrukciji niski

Prvi pristup za rekonstrukciju niske (DNK sekvence u našem slučaju) bi bio svođenje ovog problema na problem nalaženja Hamiltonove putanje. To je putanja u grafu koja posećuje svaki čvor tačno jednom. Podaci se mogu modelovati grafom na sledeći način. K-grami formiraju čvorove grafa. Oni su povezani sa usmerenom granom, ako dele jedan (*k-1*)-gram, odnosno sufiks dužine (*k-1*)-gram izlaznog čvora je jednak prefiksu ulaznog čvora. Primer prethodno opisanog grafa je dat na slici 1.6. Kako svaki čvor predstavlja jedan k-gram i potrebno nam je da svi k-grami budu uključeni u rekonstruisanu nisku tačno jednom, rešenje predstavlja Hamiltonovu putanju ovog grafa. Nalaženje Hamiltonove putanje u grafu je NP-kompletan problem, a algoritmi koji koriste ovaj pristup nisu efikasni [11].



SLIKA 1.6: Graf koji odgovara trigramskom sastavu niske TAATGCCATGGGATGTT

Izvor: [11] str. 126, slika 3.7

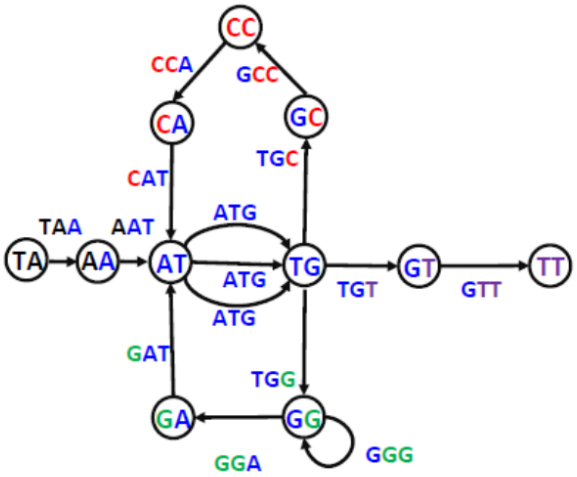
Drugi pristup svodi problem rekonstrukcije DNK sekvence na nalaženje Ojlerove putanje koja predstavlja putanju u grafu koja prolazi kroz sve grane tačno jednom. Ideja se sastoji u izmeni grafovske reprezentacije sekvence tako što se grane obeležavaju k-gramima umesto čvorova. Dakle, svaka grana je obeležena jednim k-gramom. Izlazni čvor odgovarajuće grane

k-grama je obeležen prefiksom k-grama, odnosno početnim (*k-1*)-gramom, dok je ulazni čvor obeležen sufiksom istog tog k-grama. Ilustracija je prikazana na slici 1.7.

SLIKA 1.7: Deo grafa u kojim grane su obeležene trigramima (k-grami) a čvorovi dvagramima (sufiksi i prefiksi)

Izvor: [11] str. 131, slika 3.12

Čvorove koje imaju istu oznaku treba spojiti u jedan, pri čemu su zadržane sve grane koje su ulazile u taj čvor ili izlazile iz njega. Tako je dobijen De Brojnov graf kao što se vidi na slici 1.8. De Brojnov graf predstavlja usmereni graf koji kompaktno predstavlja preklapanje između k-grama. DNK sekvenca genoma je Ojlerova putanja. Algoritam za nalaženje Ojlerove putanje nije NP-kompletan i rešenje može efikasno da se nađe. Efikasan algoritam za rekonstukciju genoma jeste algoritam za pronalaženje Ojlerove putanje, a njegova dobra implementacija, koristeći pametnu strukturu podataka, radi u linearnom vremenu [11].



SLIKA 1.8: De Brojn graf

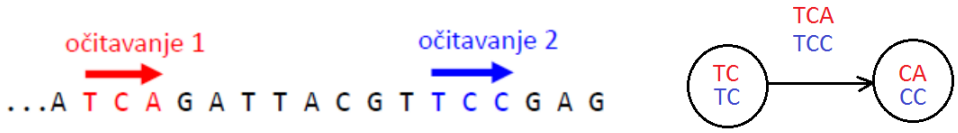
Izvor: [11] str. 134, slika 3.14

Pre procesa sklapanja graf mora da bude očišćen od čvorova i grana koji su rezultat grešaka u sekvenciranju DNK sekvenci. Zbog komplikacija kod sekvenciranja i sklapanja, u realnim primenama ne možemo da dobijamo Ojlerovu putanju celog grafa, već samo njegovih podgrafova. Ove manje putanje predstavljaju kontige. Nakon inicijalne izrade kontiga, postupak sklapanja se nastavlja korišćenjem uparenih očitavanja i partner-uparenih očitavanja da bi iskombinovali kontige u skafoldima, što je prikazano na slici 1.4.

1.2.3 De Brojnov graf na osnovu uparenih očitavanja

Zbog manjeg VUF-a, *de novo* sklapanje kratkih očitavanja predstavlja teži zadatak nego *de novo* sklapanje dugačkih očitavanja. Sa druge strane, biolozi još nisu našli način na koji bi sa visokom tačnošću generisali dugačka očitavanja [5]. U slučaju da je VUF mali, tada je dužina očitavanja manja nego dužina nekog ponovka[5] u DNK sekvenci, pa pozicioniranje ponovaka i utvrđivanje broja ponavljanja predstavlja posebno težak zadatak. Danas postojeće dužine očitavanja nisu dovoljno velike da razreše većinu ponovaka u genomu [5].

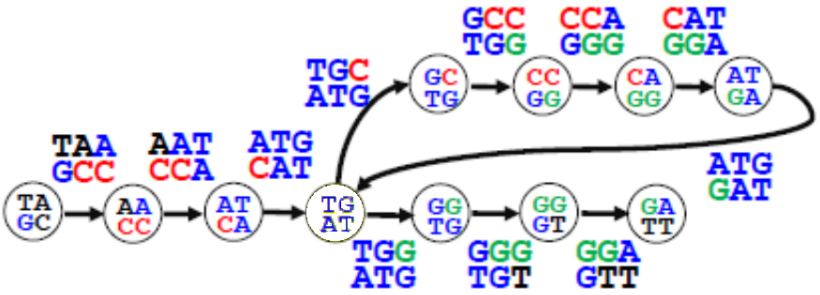
Jedno rešenje za gorepomenute probleme bi bilo korišćenje uparenih očitavanja čime virtualno povećavamo dužinu očitavanja [5]. S obzirom da se uparena očitavanja nalaze na fiksnom rastojanju u genomu, možemo ih posmatrati kao duga očitavanja sa prazninom između krajeva. Početak i kraj ovog dugog virtualnog očitavanja je poznat, ali segment na sredini nije. Ova očitavanja sa prazninama u sebi sadrže više informacija nego jednostrana očitavanja, zbog čega ih većina postojećih programa i koristi. Sledeći korak unapređenja prethodnog algoritma bi stoga bio upareni De Brojnov graf na osnovu uparenih očitavanja.



SLIKA 1.9: TCA i TCC čine jedan upareni trigram

Analogno običnom De Brojnovom grafu konstruišemo upareni De Brojnov graf, s tim što umesto k-grama koristimo uparene k-grame [11]. Primer para očitavanja je na slici 1.9. Svako rešenje problema rekonstrukcije niske pomoću uparenih očitavanja odgovara jednoj Ojlerovoj putanji u De Brojnovom grafu uparenih očitavanja. Ilustracija uparenog De Brojnovog grafa je na slici 1.10.

[5]Ponovci su fenomen kada se šablon od nekoliko nukleotida pojavljuje više puta u nizu.

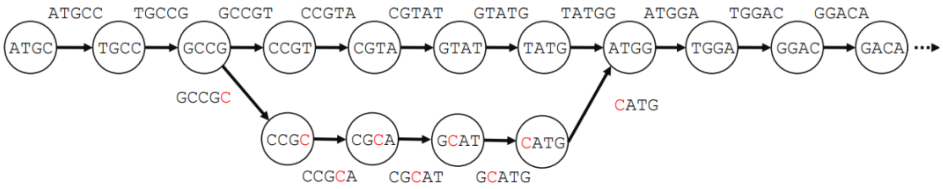


SLIKA 1.10: Upareni De Brojnov graf

Izvor: [11] str. 156, slika 3.34

1.2.4 Problemi i posledice kod sekvenciranja i sklapanja

- Kod genomskih uzoraka koji se daju sekvenceru može da dođe do kontaminacije DNK uzorkom drugih organizama, bilo pri uzimanju uzorka ili tokom laboratorijskih procedura. Sekvenciranje kontaminiranih uzoraka proizvodi očitavanja koja sadrže greške [5]. Prilikom konstrukcije De Brojnovog grafa ove greške se pojavljuju, na primer, kao balončići koji se stvore u grafu. Primer ovog fenomena je prikazan na slici 1.11. Balončić se sastoji od dve kratke različite putanje sa istim početnim i krajnim čvorovima u De Brojnovom grafu.



SLIKA 1.11: Balončić u De Brojnovom grafu usled pojave greške u očitavanju nukleotida T nukleotidom C

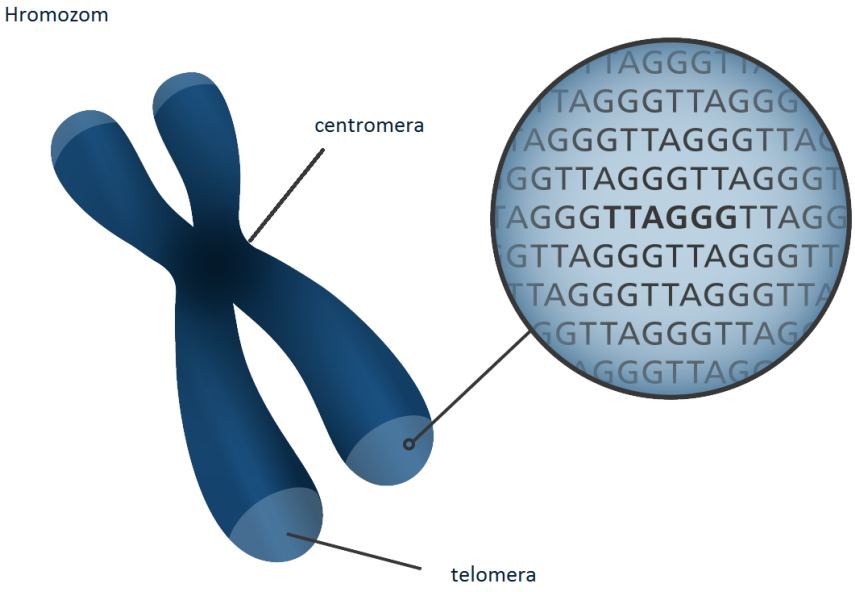
Izvor: [1] str. 60, slika 3.20

- DNK se sastoji iz dva lanca i ne može se unapred reći sa koje strane je pročitano izabrano očitavanje. Ne znamo unapred da li treba da koristimo dobijeno očitavanje pri sklapanju ili njegovo obrnuto komplementarno očitavanje koje odgovara njegovom paru na drugom lancu DNK sekvence. Primer na slici 1.12, obrnuto komplementarna niska niske $GCTCATT$ je $AATGAGC$.

SLIKA 1.12: Obrnuto komplementarne niske

Izvor: http://rosalind.info/glossary/reverse-complement/

- Može se dogoditi da neki delovi genoma ne budu pokriveni ni sa jednim očitavanjem. Savršena pokrivenost genoma očitavanjima (svaki k-gram iz genoma je očitan) nije moguća u većini realnih projekata, pa je samim tim nemoguće rekonstruisanje celog genoma.
- Ne može se pretpostavljati da je frekvencija očitavanja uniformna, odnosno, da su isti broj puta očitani svi delovi. Ovo komplikuje utvrđivanje da li se neko očitavanje više puta pojavljuje u DNK niski, ili je jedan region više puta očitan od drugih regiona.
- Stuktura samog genoma otežava sekvenciranje i sklapanje. Centromere, telomere i drugi repetativni regioni su komplikovani (slika 1.13), i što se tiče nukleotidskog sastava, i što se tiče položaja u prostoru. Oni nisu dobro opisani ni kod genoma čoveka ili miša, o kojima najviše znamo danas [5].



SLIKA 1.13: Hromozom

Izvor: https://www.yourgenome.org/facts/what-is-a-telomere

- Geni od velikog značaja koji se brzo razvijaju su u velikoj meri polimorfni i imaju puno paraloga[6]. Njih je posebno teško sklapati i obično su manje prisutni u finalnom rezultatu sklapanja.

1.2.5 Rešenje nekih problema

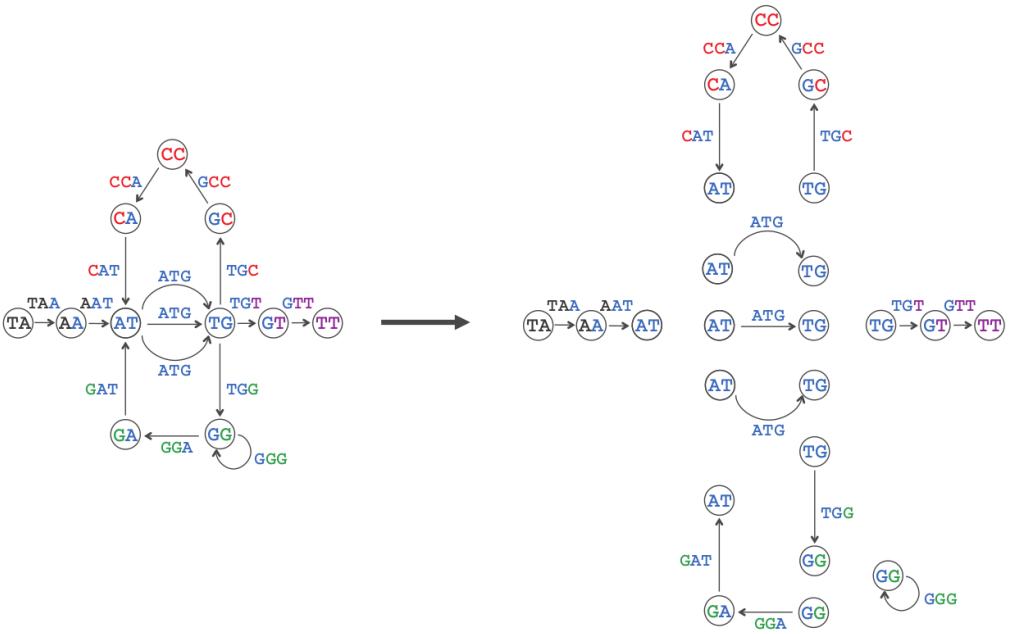
Za izabran k-gram definišemo pokrivenost kao broj očitavanja kojima on pripada. Problem predstavlja što kod realnih podataka retko imamo savršenu pokrivenost, odnosno sekvenceri ne uspevaju da pročitaju neke k-grame genoma zbog prethodno spomenutih problema prilikom sekvenciranja. Efekat se može oblažiti rastavljanjem očitavanja na manje k-grame što se koristi kod velikog broja modernih asemblera, čime težimo ka savršenoj pokrivenosti [5]. Sa druge strane, veličinu k treba pažljivo izabrati, zato što De Brojnov graf dobijen sa prekratkim k-gramima sadrži veliki broj grana, koji ga čine komplikovanim, a na osnovu njega je teško sklapanje genoma. Ni posle rastavljanja očitavanja na manje k-grame problem savršene pokrivenosti u mnogim slučajevima nije rešen u potpunosti - i dalje ima praznina u pokrivenosti k-grama. Rezultat nesavršene pokrivenosti je De Brojnov graf sa granama koje nedostaju a koje bi se u slučaju savršene pokrivenosti nalazile u De Brojnovom grafu što rezultira time da ne možemo da nađemo Ojlerovu putanju u celom grafu. Sklapanje hromozoma u celosti je u ovom slučaju nemoguće, a najviše što se može rekonstuisati iz ovakvih podataka su kontige. Kao što je definisano u poglavlju 1.1.3, kontige su dugački, neprekidni segmenti genoma. Kontiga je maksimalna putanja u grafu takva da svaki njen unutrašnji čvor ima tačno jednu ulaznu i tačno jednu izlaznu granu, i ne može se produžiti tako da važi prethodni uslov. Ove putanje predstavljaju niske nukleotida koje su delovi bilo kog sklapanja genoma za izabrani broj k (dužina k-grama).

U praksi moramo da rastavimo genom na kontige čak i u slučaju savršene pokrivenosti sa očitavanjima, zato što ponovci u genomu sprečavaju da imamo jedinstvenu Ojlerovu putanju. Na slici 1.14 primeru De Brojnovog grafa odgovaraju dve Ojlerove putanje:

1. $TAATGCCATGGGATGTT$
2. $TAATGGGATGCCATGTT$

pa je rastavljanje na kontige neophodno.

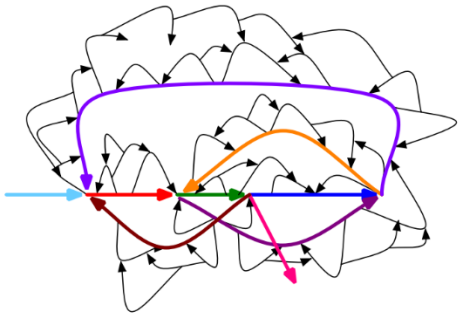
[6]Paralozi su geni koji imaju duplikate u jednom genomu



SLIKA 1.14: Rastavljanje De Brojnovog grafa na kontige

Izvor: [11] str. 160, slika 3.38

Zbog očitavanja koja sadrže greške u sebi, De Brojnov graf sadrži balončiće. Asembleri teže ka tome da eliminišu balončiće iz De Brojnovog grafa. Skoro svako očitavanje ima greške i De Brojnov graf ponekad ima čak i milion balončića. Zbog balončića nekad se ne mogu prepoznati preklapanja među očitavanjima. U slučaju više grešaka dolazi do eksplozije balončića (slika 1.15) koji previše zakomplikuje graf, čineći sklapanje težim zadatkom.



SLIKA 1.15: Eksplozija balončića

Izvor: [11] str. 161, slika 3.39

Postoje heuristički algoritmi za uklanjanje balončića koji pokušavaju da sklone deo putanje u balončićima koji sadrži grešku. Cilj heuristike je da se brzo dođe do rešenja, koje ne mora biti nužno najbolje, a ponekad, nažalost, uklone ispravnu putanju i ostave putanju sa greškom, i tako ne poprave grešku. U strukturi genoma regioni koji se ponavljaju i skoro su identični, osim par nukleotida, generišu balončić u De Brojnovom grafu. Ove grane

opisuju strukturu genoma i ne bi ih trebalo eliminisati, iako kreiraju balončić. Ako se one ipak eliminišu, onda dolazi do nove greške u sklapanju. Moderni asembleri pokušavaju da razlikuju balončiće koji su se zbog greške pojavili (njih treba otkloniti) i one koji su prisutni zbog stukture samog genoma (njih ne treba otkloniti).

1.2.6 Indikatori grešaka

Postoje razne metrike za merenje kvaliteta asemblera. Jedna osnovna metrika je procenat genoma koji je sadržan u rezultatu sklapanja. Druga standardna metrika je ispitivanje kontinualnosti sklapanja pomoću N50 statistike koja predstavlja statistiku skupa kontiga. N50 definišemo kao najveći broj x za koji važi da ukupan zbir dužina kontiga koje su duže od x ili jednake, čini bar 50% ukupne dužine svih kontiga. N50 statistika se može posmatrati kao srednja vrednost dužine sklapanih kontiga pri čemu su duže kontige važnije i dobijaju veću težinu. N50 statistika opisuje samo neprekidnost i ne sadrži informacije o tačnosti sklapanja.

Nakon sklapanja ponovo se poravnavaju uparena očitavanja i partner-uparena očitavanja sa kontigama gde se mogu detektovati različiti indikatori grešaka:

1. regioni u kojima je dubina pokrivenosti niska ili u kojima uparena očitavanja nisu u istom smeru poravnata ukazuju na greške u sklapanju

2. regioni koji su pokriveni očitavanjima u većoj meri nego ostali delovi ukazuju da su u rekonstruisanoj niski ponovci savršeni, dok u realnosti nisu [5], i to su regioni u kojima se gusto pojavljuju jednonukleotidni polimorfizmi (*SNP*)

3. ako imamo očitavanja koja su poravnata sa savršenim ponovcima i ako su među njima većina očitavanja identična, a neka se vrlo malo razlikuju, ovaj fenomen takođe ukazuje na nesavršenost ponovaka u realnosti

Spoljašnji alati mogu da budu od pomoći da pronađu kontaminirane delove posle sklapanja u formi zasebnih kontiga ili delova kontiga. Ovi alati pomoću lokalnog poravnanja (npr BLAST) pretražuju postojeće baze sklopljenih genoma drugih organizama za pronalaženje kontaminiranih regiona [5]. BLAST [12] pretraga omogućava poređenje niza sekvenci sa bibliotekama ili bazama podataka sekvenci. Ovi alati nisu savršeni, rezultati ne pokazuju nužno na greške, čak i ispravno sklapanje može da bude poravnato pored drugog organizma, ako je u pitanju organizam koji je srodan sa originalnim organizmom. Takođe, neke manje kontaminacije mogu da ostanu neprimećene u slučaju kada su drugi ispravni delovi kontaminirane kontige značajni pogoci za originalan organizam.

U slučaju kontaminacije ljudskog porekla (npr. kontaminacija zbog rukovanja uzorkom) posebno je problematično odvojiti kontige koje su ispravne od ostalih [5]. Delovi uzorka genoma sisara u slučaju *de novo* sekvenciranja će najbolje da se poravna sa genomom čoveka ili miša.

Poglavlje 2

Integratori

2.1 Ideja integrisanja

Alati za *de novo* sklapanje genoma su različiti po brzini, prilagodljivosti i kvalitetu rezultata algoritama sklapanja. Svako sklapanje je jedinstveno zbog stukture genoma, podataka i različitosti genoma. Ne zna se unapred koji algoritam sklapanja odgovara najviše nekim podacima. Razlike između dva genoma različitih organizama mogu biti u veličini, kompoziciji baznih parova, segmentima koji se ponavljaju i u stepenu polimorfizma. Nekim algoritmima je fokus na minimizaciji grešaka pri sklapanju, a drugima je na tome da poboljšaju neprekidnost, čak iako kontige sadrže neke greške [2]. Neki asembleri bolje rade na nekim specifičnim regionima genoma, a drugi na drugim regionima.

Uobičajeno je da se korak sklapanja uradi više puta sa različitim asemblerima, a nakon toga se izabere najbolji rezultat po određenim kriterijumima ili parametrima. Kod izbora algoritma za sklapanje moraju se uzeti u obzir karakteristike DNK sekvenci i željeni kvalitet sklapanja. Zbog različitosti DNK sekvenci kod različitih organizama, za konkretan uzorak neki algoritmi su bolji, a neki lošiji. Svaki skup podataka ima svoje karakteristike. Heuristike pojedinačnih asemblera obično samo delimično rešavaju predstavljene probleme. Neke metode za sklapanje nadmaše druge, ali trenutno je vrlo teško predvideti koji alat bi bio najbolji u datoj situaciji. Da bi mogli da poboljšaju *de novo* sklapanje, zbog pomenutih teškoća uvedeni su novi algoritmi. Jedna strategija za poboljšanje *de novo* sklapanja se zove usaglašavanje sklapanja (*assembly reconciliation*) [2]. Ova strategija podrazumeva upoređivanje rezultata iz više različitih asemblera i njihovo integrisanje (*merging*) sa ciljem dobijanja kombinacije koja je potencijalno bolja od originalnih sklapanja, čime se poboljšavaju neprekidnost i tačnost završnog rezultata. Alati koji vrše integrisanje se nazivaju integratori.

Za usaglašavanje sklapanja postoje razni alati. Reconciliator [13] je jedan od prvih integratora koji se sastoji od dve faze, prve faze koju čine naizmenične iteracije pronalaženja grešaka i njihovog ispravljanja i druge faze integrisanja. Koristi globalno poravnanje između sklapanja. Integrator GAM [3] umesto globalnog poravnanja pomoću lokalnog poravnanja i pomoćnog grafa opisuje veze između dva sklapanja. Za oba integratora su potrebne datoteke u kojima su navedena, uz svako sklapanje, i sva očitavanja koja su bila korišćena (takozvane afg datoteke). Samo mali broj asemblera nove

generacije pravi ove datoteke (na primer Velvet[9], Ray i SUTTA). Drugo ograničenje ovih integratora je što njihova ulazna sklapanja moraju da budu napravljena pomoću istog skupa očitavanja. Integrator GAA [14] globalnim poravnanjem između dva sklapanja napravi graf, koji posle koristi za integrisanje sklapanja. Glavna mana GAA integratora je obavezna faza globalnog poravnanja, koja je računarski zahtevni korak i može da uvede neispravne veze u slučaju da ima paralog sekvence u DNK-u. Integrator GAM-NGS [2] je napravljen nalik GAM integratoru i njegov glavni cilj je da učešljava dva sklapanja i dobije korektniji rezultat sa većom kontinualnošću. U ovom radu će biti korišćen GAM-NGS i o njemu će biti više u narednom potpoglavlju.

2.2 GAM-NGS (*Genomic assemblies merger for next generation sequencing*)

GAM-NGS je jedan od najbržih alata za usaglašavanje sklapanja i pogodan je za rad sa ogromnim skupovima podataka. Dok drugi alati za integrisanje obično koriste globalno poravnanje, bitna karakteristika GAM-NGS-a je da koristi lokalno poravnanje, što ga čini efikasnijim [2]. Pre integrisanja se vrši poravnanje očitavanja sa rezultatom asemblera. Poravnanje očitavanje je proces određivanja odakle su potekla očitavanja iz sklapanog genoma. S obzirom da je dužina očitavanja nekoliko desetina ili stotina nukleotida a dužina genoma nekoliko miliona nukleotida, ovo predstavlja kompleksan problem, ne samo zato što je genom veliki, već zato što je i traženje značajno manjih sekvenci u velikoj sekvenci težak zadatak. Drugi otežavajući faktor je da ne tražimo identičnu sekvencu u genomu, nego samo dovoljno sličnu datom očitavanju. Poravnanje očitavanja omogućava pronalaženje regiona u genomu koji su bili sastavljeni od istih očitavanja, i samim tim su dobri kandidati za predstavljanje istog lokusa[7].

Ako se rezultati sklapanja različitih alata razlikuju na istim pozicijama, kažemo da je na tim pozicijama došlo do konflikta. GAM-NGS u fazi integrisanja rešava konflikte između različitih sklapanja koristeći jednu specifičnu strukturu podataka, takozvani graf sklapanja (*assemblies graph - AG*). Graf sklapanja predstavlja težinski graf, to jest graf gde je svaka grana otežana verovatnoćom da je ona deo prave putanje, odnosno deo rezultata usaglašavanja sklapanja. Pretpostavka je da delovi koji su sagrađeni od istih očitavanja najverovatnije predstavljaju isti genomski lokus. Ideja za usaglašavanje sklapanja podrazumeva pronalaženje u velikoj meri sličnih delova između sklapanja, odnosno regiona kojima je pridružen veliki broj istih očitavanja. Pomoću grafa sklapanja, u ovim sličnim regionima, možemo da identifikujemo delove u kojima su asembleri u kontradikciji, na primer, u obliku balončića i petlje. Težine se iskorišćavaju za razrešavanje konflikata.

[7] specifična lokacija gena ili DNK sekvence na hromozomu

2.2.1 Ulazni fajlovi

Za GAM-NGS ulazni fajlovi su dva rezultata sklapanja i SAM fajlovi za svaku ulaznu biblioteku. SAM fajl (*Sequence Alignment Map*) predstavlja format za čuvanje bioloških sekvenci poravnatih sa referentnom sekvencom, nalik mapi poravnanja sekvenci. SAM fajl je tekstualni format koji podržava i kratka i duga očitavanja. Njemu ekvivalentan je binarni fajl BAM, koji čuva iste podatke u kompresovanoj binarnoj reprezentaciji. Ovi fajlovi se mogu analizirati i editovati korišćenjem SAMtools softvera [15].

Veliki broj asemblera ne izdaje datoteku u kojoj su navedena, uz dobijeno sklapanje, i sva očitavanja koja su bila korišćena (takozvana *afg* datoteka). Da bi premostio ovaj nedostatak, GAM-NGS aproksimira ove podatke koristeći samo očitavanja koja su uspešno poravnata sa dobijenim sklapanjem. Ovakva aproksimacija može da dovede do greške u slučaju kada ima vrlo sličnih regiona u genomu na više mesta. U takvim slučajevima algoritam nasumično izabere mesto gde će pridružiti očitavanje. Kako bi držao ovakve probleme pod kontrolom, GAM-NGS koristi samo očitavanja koja su poravnata sa jednom pozicijom i pritom odbacuje sva očitavanja koja imaju dva ili više poravnanja sa visokim skorom (tzv. dvosmislena poravnanja).

2.2.2 Pripreme pre konstrukcije grafa sklapanja

Na početku svakog integrisanja korisnik izabere koje će biti glavno (*master*) a koje sporedno (*slave*) sklapanje. Asembleri implementiraju različite heurističke algoritme, što dovodi do toga da se rezultati ponekad ne slažu u potpunosti[8]. Za identifikaciju i rešenje ovakvih situacija se izgradi težinski graf sklapanja, u čemu je sačuvan najverovatniji redosled regiona koji su izgrađeni od istih očitavanja. Problematični regioni u grafu sklapanja se uglavnom javljaju u obliku podgrafova, koji imaju neku specifičnu strukturu. Ovi lokalni problemi rešavaju se izborom najbolje putanje u grafu, koja maksimizuje određene parametre na lokalnom nivou. Program prilikom odlučivanja koja je bolja putanja koristi težine u grafu, informacije iz uparenih očitavanja i partner-uparenih očitavanja, ako ih ima. Ako nema dovoljno informacija da odluči koje sklapanje je ispravno u određenom regionu gde postoji konflikt, npr. težine u grafu su vrlo slične, tada merdžer uvek izabere verziju iz glavnog sklapanja. Posle razrešavanja kontradikcija između sklapanja, GAM-NGS učešljava kontige iz pojednostavljenog grafa i tako pronalazi konsenzus nisku. Tako je dobijeno poboljšano sklapanje, odnosno usaglašavanje sklapanja.

Okvir za određenu kontigu C u sklapanju A je definisan kao niz susednih očitavanja koja su joj pridružena u sklapanju A. Ako posmatramo dva različita sklapanja, sklapanje M (glavno (*master*)) i sklapanje S (sporedno (*slave*)), onda možemo da definišemo blok B kao par okvira nad M i S sastavljenih od istog niza očitavanja. Veličina bloka predstavlja broj očitavanja koja ga čine. Prilikom konstrukcije blokova posmatramo samo očitavanja koja su

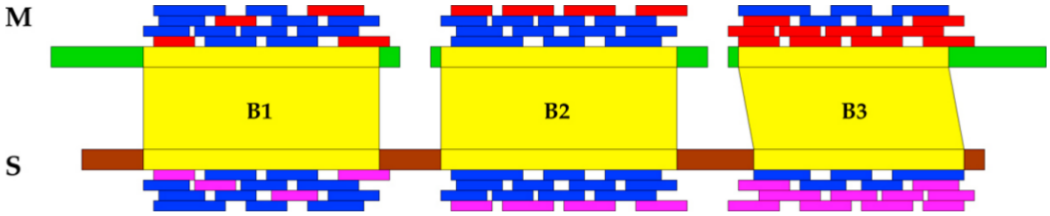
[8]Primer kontradikcije može da bude kada je redosled nekih delova u sklapanim sekvencama različit.

jedinstveno poravnata uz oba sklapanja. Za svako takvo očitavanje, ako je susedno sa nekim prethodno napravljenim blokom onda taj blok produžimo sa ovim očitavanjem, inače počinjemo novi blok.

Graf sklapanja se konstruiše pomoću blokova. Da bi se smanjila kompleksnost grafa sklapanja, uvedena su dva dodatna filtera pre konstukcije grafa, jedan na osnovu provere dužine blokova i drugi na osnovu analize dubine pokrivanja (*depth-of-coverage*).

Za filtriranje proverom dužine blokova ideja je da eliminišemo blokove koji su koristili mali broj očitavanja u odnosu na broj očitavanja koja su pridružena okvirima u oba sklapanja, odvojeno gledano. Maksimum između dva odnosa se uzima u obzir da ne bismo eliminisali blokove koji odgovaraju heterozigotnim regionima (može da se desi da jedan asembler vraća oba alela, dok drugi vraćaju samo jedan od njih).

Blokovi predstavljaju regione koji pripadaju sklapanju M i sklapanju S, koji dele relativno veliki broj poravnatih očitavanja. Na slici 2.1 plava očitavanja su skupovi susednih očitavanja, koja su jedinstveno poravnata sa istom kontigom u oba sklapanja. Štaviše, GAM-NGS odbacuje blokove kao što je B_3, koji sadrži malu količinu zajedničkih očitavanja u odnosu na broj očitavanja poravnatih u istim regionima. Zadržavanje bloka B_3 bi moglo da napravi pogrešnu vezu među kontigama.



SLIKA 2.1: Filtriranje blokova na osnovu dubine pokrivenosti

Izvor: [2] str. 5, slika 1

Drugi filter na osnovu analize dubine pokrivenosti uvodi dve vrste pokrivenosti okvira: pokrivenost blokovima (*block coverage* (*BC*)) i globalnu pokrivenost (*global coverage* (*GC*)). Neka je blok B sa okvirima F_M na glavnom sklapanju (*master*) i F_S na sporednom pokrivanju (*slave*). GAM-NGS izračuna za svaki okvir obe vrste pokrivenosti. Neka je $\mathcal{R}_{F_M}$ skup svih jedinstveno poravnatih očitavanja na okviru F_M, a $\mathcal{R}_{B_M}$ je skup svih jedinstveno poravnatih očitavanja koji su bili korišćeni za pravljenje bloka B. ($\mathcal{R}_{B_M} \subseteq \mathcal{R}_{F_M}$) Definišemo ih na sledeći način:

$$BC_{F_M} = \frac{\sum_{r \in \mathcal{R}_{B_M}} |r|}{|F_M|}$$

$$GC_{F_M} = \frac{\sum_{r \in \mathcal{R}_{F_M}} |r|}{|F_M|}$$

GAM-NGS eliminiše blokove koje ne ispunjavaju sledeći uslov za unapred definisan prag T_C od strane korisnika:

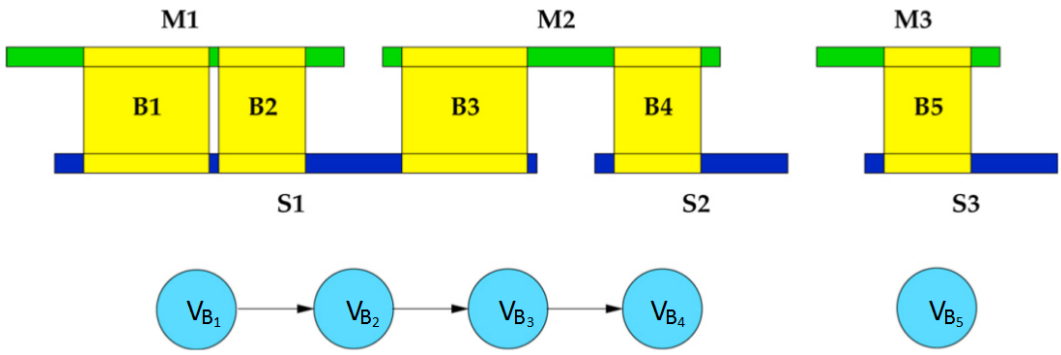
$$max\left\{\frac{BC_{F_M}}{GC_{F_M}}, \frac{BC_{F_S}}{GC_{F_S}}\right\} \geq T_C$$

Ideja je da se eliminišu blokovi koji su napravljeni pomoću malog broja očitavanja u odnosu na broj poravnatih očitavanja na oba okvira. Koristi se maksimum između dva razlomka da ne bi eleminisali blokove koji odgovaraju regionima sa heterozigotnim pozicijama.

U posebnom koraku definiše se redosled blokova u odnosu na okvire koji su nad njima.

2.2.3 Konstrukcija grafa sklapanja

Za svaki blok B definišemo čvor V_B u grafu sklapanja. Dva čvora su povezana granom, ako odgovarajući blokovi dele bar jedan okvir na istoj kontigi iznad njih. Detaljnije, ako je u prethodno definisanom redosledu blokova B_1 pre B_2, onda je usmerenje grane od V_{B1} do V_{B2}, kao na slici 2.2.



SLIKA 2.2: Konstrukcija grafa sklapanja

Izvor: [2] str. 6, slika 2

Svakoj grani pridružimo težinu koju računamo u odnosu na više faktora vezanih za regione blokova čije čvorove povezuje baš ta grana. Neki od faktora su pokrivenost očitavanjima i broj ispravno i neispravno orijentisanih uparenih očitavanja među blokovima. Izračunata težina predstavlja verovatnoću da je grana deo ispravne putanje u grafu sklapanja.

Svaka putanja u grafu sklapanja odgovara jednom nizu blokova, tako da svaki par susednih blokova u nizu leži na istoj kontigi bar u jednom sklapanju. Na osnovu toga, graf sklapanja se može iskoristiti za produženje kontiga ili njihovo spajanje.

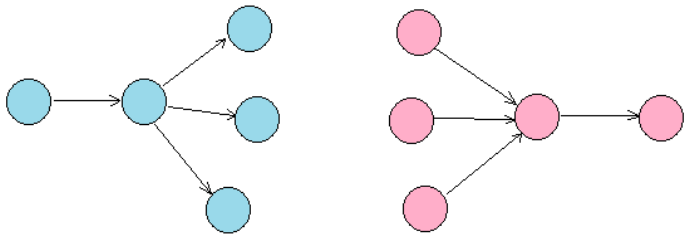
2.2.4 Rešavanje problematičnih regiona

Redosledi blokova koji slede iz dva sklapanja mogu da budu u kontradikciji. Na primer, ako dva bloka leže na jednoj istoj kontigi u oba sklapanja,

samo sa suprotnim smerom, onda se formira ciklus u grafu sklapanja. Jako povezane komponente (*strongly connected components, (SCC)*) koje sadrže bar dva čvora označavaju situaciju gde se sklapanja M i S ne slažu u redosledu nekih blokova. Za pronalaženje ovakvih regiona koristi se Tarjanov algoritam [16], koji u linearnom vremenu pronalazi jako povezane komponente u grafu sklapanja.

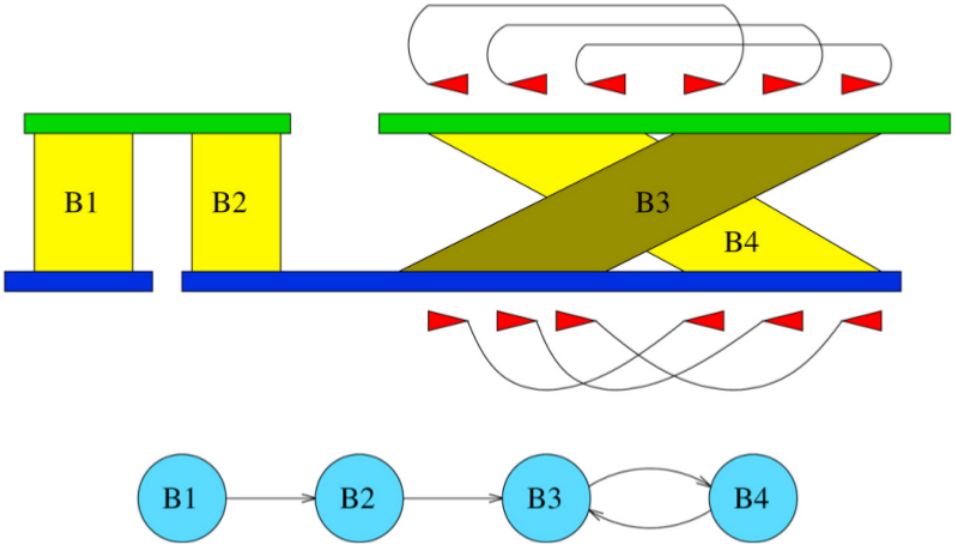
Drugi mogući problem se javlja u divergentnim putanjama koje mogu da ukazuju na situacije gde se sklapanja ponašaju drugačije na lokalnom nivou: jedno sklapanje je proširilo sekvencu na drugi način u odnosu na drugo sklapanje. Izborom lokalnog optimalnog rešenja pomoću težine grana, u grafu sklapanja se razrešavaju konflikti (npr. u slučaju grananja (*bifurcation*) putanja koja minimizuje broj grešaka u sklapanjima će biti izabrana). Kada su težine slične i ne može se jasno izabrati najbolji, onda se vraća na glavno sklapanje, a kontige iz njega se uzimaju kao rešenje integrisanja.

Između raznih struktura generisanih u grafu, zbog protivrečnih delova sklapanja, pojavljuju se ciklusi i grananja. Grananja se sastoje od putanja koje samo divergiraju ili konvergiraju. Na slici 2.3 podgraf koji je označen plavom bojom predstavlja primer divergentne putanje sa najviše jednom ulaznom granom i više izlaznih, a roze predstavlja konvergentnu putanju.



SLIKA 2.3: Divergentne i konvergentne putanje

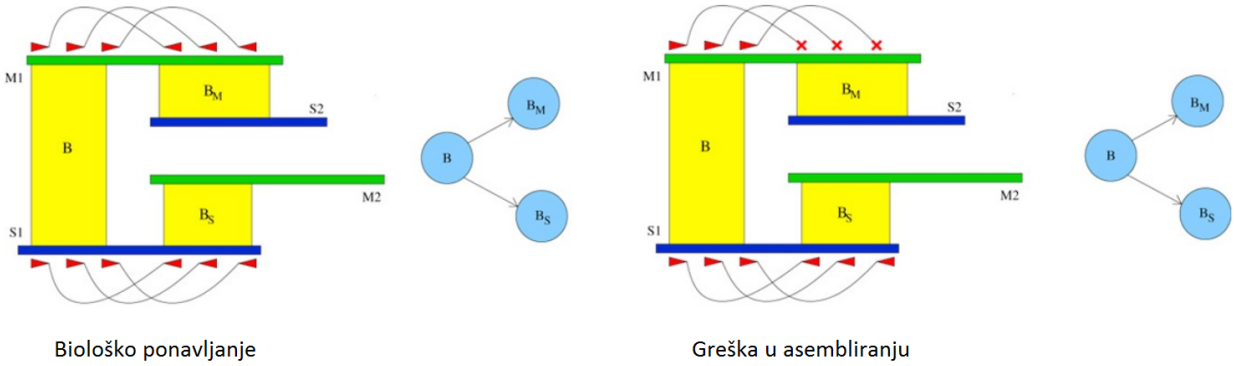
Ciklusi koji sadrže tačno dva čvora ukazuju na inverzije u istoj kontigi u sklapanjima. Za razrešavanje ovakvih ciklusa iskorišćavaju se partner-uparena očitavanja i uparena očitavanja. Ako je ciklus zaista rezultat invertovanih blokova, onda će partner-uparena očitavanja i uparena očitavanja biti samo u jednom sklapanju poravnata sa ispravnom orijentacijom. (na slici 2.4 donje sklapanje obeleženo plavom bojom)



SLIKA 2.4: Ciklus sa tačno dva čvora u grafu sklapanja kao rezultat jedne inverzije uz jednu kontigu u sklapanju *M* i skalapanju *S*.

Izvor: [2] str. 7, slika 3

Partner-uparena očitavanja su poravnata sa pogrešnom orijentacijom u jednom sklapanju. Ispravno sklapanje je plave boje u ovom slučaju. Dakle, ako se ukazuje da je nekoliko očitavanja poravnato ispravno u jednom sklapanju, a sa pogrešnom orijentacijom u drugom sklapanju, onda se u rezultatu integrisanja koristi ispravna sekvenca iz sklapanja sa dobrom orijentacijom. U suprotnom se koristi glavno sklapanje u problematičnom regionu.



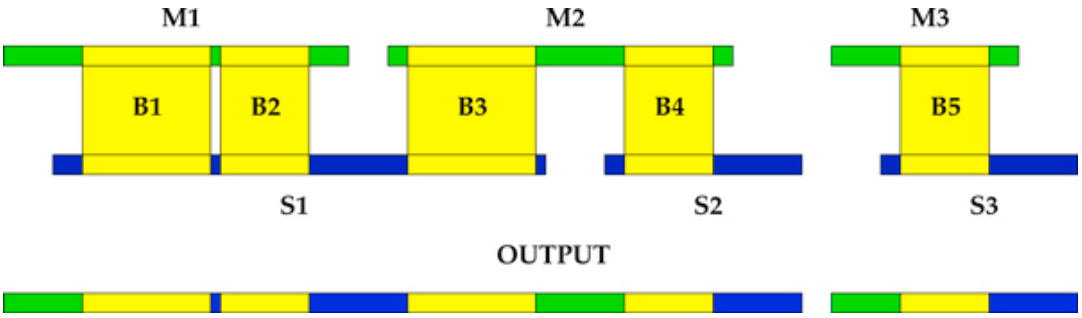
SLIKA 2.5: Grananja u grafu sklapanja

Izvor: [2] str. 8, slika 4

Podgrafovi koji sadrže grananje (*bifurcation*) mogu da predstavljaju ponovke u niski ili grešku, što je prikazano u primeru na slici 2.5. Oni mogu da znače

biološko ponavljanje ili grešku u asembliranju. U slučaju biološkog ponavljanja uparena očitavanja su ispravno poravnata. U slučaju greške u asembliranju uparenih očitavanja na kontigi $M1$ mogu da pomognu u razrešavanju grananja (*fork*) u sklapanju.

Integrisanje se vrši posle razrešavanja konflikata u grafu sklapanja. Pronalaze se razdvojene putanje (koje ne sadrže istu granu) u grafu, da bi se pomoću njih napravio nacrt poravnanja kontiga koje pripadaju različitim sklapanjima. Primenjuje se jedan algoritam poluglobalnog poravnanja, kako bi bilo utvrđeno da kontige imaju visoku sličnost (bar 95% identičnost) i da se može izvesti integrisanje. Rezultat je sekvenca koja pripada sklapanju koje se lokalno pokazuje kao najbolje. Izlaz za finalno poboljšano sklapanje posle integrisanja je skup prethodno unapređenih kontiga kao i kontige iz glavnog sklapanja koje nisu bile uključene u integrisanju. (slika 2.6)



SLIKA 2.6: GAM-NGS faza integrisanja

Izvor: [2] str. 9, slika 5

Poglavlje 3

Konkretnije o izabranim asemblerima

3.1 Asembler ABySS (*Assembly by Short Sequences*)

Nove tehnologije DNK sekvenciranja proizvode ogromne količine kratkih očitavanja dužine od 25 do 500 baznih parova visokog kvaliteta. Ova očitavanja su znatno kraća nego očitavanja drugih tehnologija, ali je ukupan broj sekvenciranih baznih parova za red veličine veći. Nedostatak nekih asemblera je nemogućnost efikasnog sklapanja velike količine podataka generisanih iz sekvenciranja velikih genoma. Većina drugih asemblera su jednonitne aplikacije dizajnirane da rade na jednom procesoru. Za razliku od njih, ABySS je softver koji paralelno vrši sklapanje kratkih očitavanja.

Glavna inovacija u ABySS-u je distribuirana reprezentacija De Brojnovog grafa, koji omogućava paralelno sklapanje milijardi kratkih očitavanja kroz računarske mreže [6]. ABySS je posebno koristan kod *de novo* asembliranja DNK sekvenci organizama za koje ne postoji referentna sekvenca.

3.1.1 Algoritam

Algoritam ABySS programa se sastoji od dve faze.

1. U prvoj fazi su generisani svi mogući k-grami pomoću raspoloživih očitavanja. U skupu k-grama uklone se greške i naprave se inicijalne kontige. Zbog nesavršene pokrivenosti očitavanja, bez korišćenja informacija iz uparenih očitavanja i partner-uparenih očitavanja, kontige se produžavaju duž putanja u De Brojnovom grafu, dok ne stignu do grananja na putanji, ili do kraja putanje.

2. U drugoj fazi partner-uparena očitavanja i uparena očitavanja se koriste za produženje kontiga sa razrešavanjem višesmislenih putanja u De Brojnovom grafu.

3.1.2 Prva faza: uklanjanje grešaka i pravljenje inicijalnih kontiga

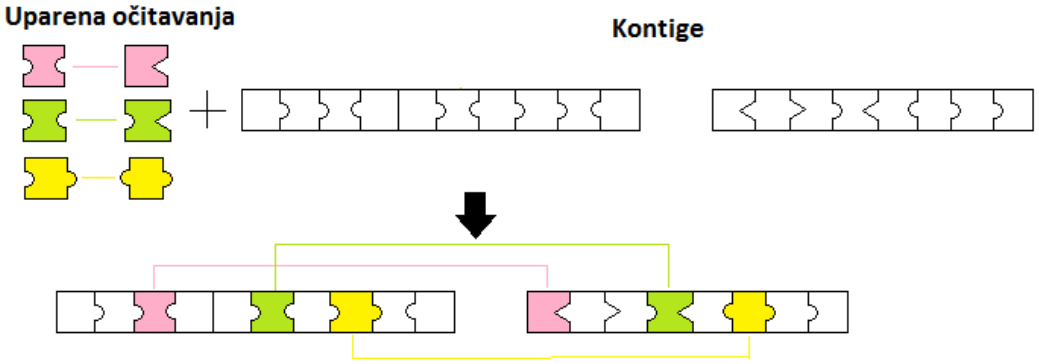
Očekuje se da se za vreme spajanja putanja u De Brojnovom grafu, prilikom formiranja kontiga, pojave greške u njihovom sklapanju. U eksperimentalnim podacima nisu svi fragmenti fiksne dužine a kod partner-uparenih očitavanja procena udaljenosti nije savršena, zbog čega dolazi do grešaka pri sklapanju kontiga. Ove greške mogu da budu manje, kao što je jedan indel zbog pogrešne procene broja ponavljanja neke sekvence, ili veće, gde su se daleki delovi DNK sekvenci pogrešno spojili.

ABySS paralelno rešava ovaj problem uklanjanjem onih čvorova i grana u De Brojnovom grafu koji su rezultati grešaka u sekvenciranju DNK sekvenci, pri čemu se ovo uklanjanje vrši pre sklapanja [6]. Slepe putanje (*"dead-end" branches, tips*) su sastavljene od mešavine ispravnih i neispravnih k-grama. Ispravni k-grami vezuju niz neispravnih k-grama u grafu. Kako su očitavanja sa greškom jedinstvena, odgovarajući k-grami sa greškom prave putanje bez nastavka. Prvo se eliminišu ove slepe putanje, ako su kraće od nekog unapred određenog praga. Posle se razrešavaju i eliminišu balončići. Za brisanje balončića pronalazi se svaki čvor divergencije[9] u grafu. Od čvora divergencije prati se svaka putanja, dok se ne spoje posle n koraka, gde je $k \leq n \leq 2k$. Ako se putanje spajaju, onda se putanja sa nižom pokrivenošću očitavanja eliminiše. Ova dva koraka eliminisanja se iteriraju za uklanjanje manjih grešaka. Poslednji korak u ovoj fazi je formiranje inicijalnih kontiga spajanjem putanje u De Brojnovom grafu pomoću nedvosmislenih grana (one grane čiji početni čvor ima samo njih kao izlazne grane).

3.1.3 Druga faza: produženje kontiga pomoću partner-uparenih očitavanja

U drugoj fazi ovog algoritma sklapanja razrešavaju se dvosmislenosti između kontiga ako su na raspolaganju partner-uparena očitavanja ili uparena očitavanja. Identifikuju se kontige koje mogu da se povežu.

[9] Na osnovu greške u očitavanjima može da se formuliše neispravna putanja, koja je povezana sa pravilnim De Brojnovim grafom na oba svoja kraja. Početni čvor ove putanje naziva se čvorom divergencije.



SLIKA 3.1: Kontige povezane sa 3 uparena očitavanja

Očitavanja se poravnaju sa inicijalnim kontigama i formiraju se skupovi povezanih kontiga. Veze napravljene na osnovu grešaka i veze napravljene sa pogrešno uparenim očitavanjima se brišu. Kontige su povezane ako ih bar p uparenih očitavanja povezuje (podrazumevano $p = 5$). Za svaku kontigu C skup kontiga S je generisan iz lista kontiga koje su uparene sa kontigom C. U De Brojnovom grafu se traži jedna jedinstvena putanja iz kontige C koja posećuje svaku kontigu iz skupa P. Uklanjanjem veza koje ne odgovaraju kriterijumima spajaju se inicijalne kontige na konzistentnoj putanji i dobijaju se finalne kontige.

3.1.4 Distribuirani De Brojnov graf

ABySS-ova inovacija je da koristi distribuiranu reprezentaciju De Brojnovog grafa, koji omogućava paralelno sklapanje kroz računarske mreže. Jedinstvena reprezentacija De Brojnovog grafa koji omogućava da susedne sekvence ne moraju da budu fizički na istom računaru je distribuirani De Brojnov graf. Ova osobina omogućava da podelimo sekvence na više računara [6].

Pronalaženje nekog k-grama je deterministički proces koji koristi samo nisku karaktera k-grama. Izračunavanje lokacije, odnosno indeksa, određenog k-grama se vrši funkcijom heširanja k-grama. Dobijeni indeks je isti i za k-gram i za njegov obrnuti komplement iz razloga što je nebitno sa koje strane lanca DNK molekula se vrši sekvenciranje. Pomoću ovog indeksa moguće je pristupiti k-gramu kroz mrežu računara (samim tim i indeks za obrnuto komplementarni k-gram).

Veze između k-grama se čuvaju na način koji je nezavisan od fizičke lokacije k-grama. Jedan k-gram, odnosno, jedan čvor u De Brojnovom grafu, može da ima najviše osam grana, po jednu za svaku moguću ekstenziju {A, C, G, T} u oba smera (slika 3.2). Ova informacija se čuva u 8 bita po k-gramu, kao niska koja označava da li grana postoji (vrednost 1 na odgovarajućem bitu), ili nije prisutna u De Brojnovom grafu (vrednost 0 na odgovarajućem bitu). Susedni k-grami se lako generišu pomoću ove informacije.

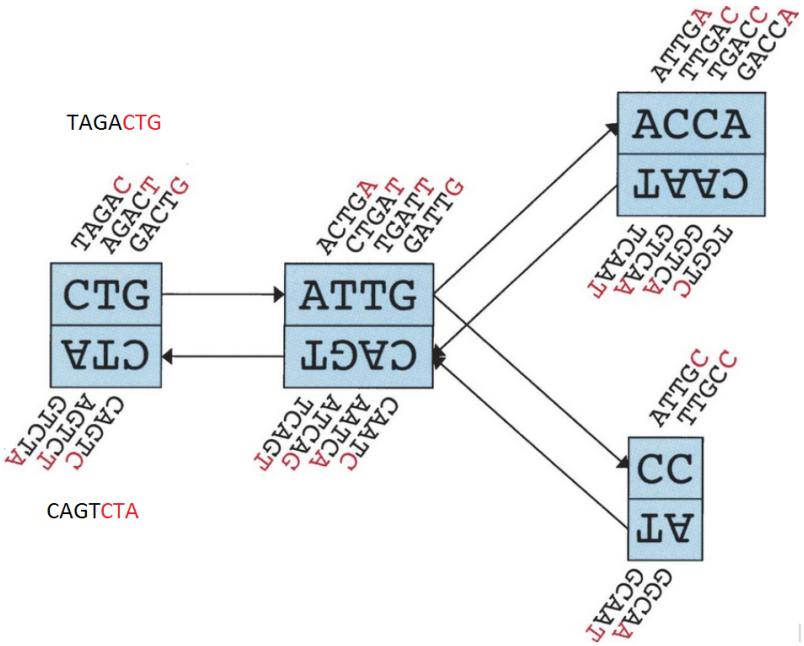
SLIKA 3.2: Jedan čvor u De Brojnovom grafu može da ima najviše osam grana

3.2 Asembler Velvet

Velvet je asembler koji nad ulaznim kratkim očitavanjima, u kombinaciji sa uparenim očitavanjima, vrši sklapanje. Ima četiri faze: heširanje očitavanja u k-gramima, konstrukcija De Brojnovog grafa, ispravljanje grešaka i razrešavanje ponovaka. Efikasno manipuliše De Brojnovim grafom prilikom uklanjanja grešaka iz grafa, i razrešavanja ponovaka [9]. Ova dva zadatka izvršava posebno: u prvom koraku koriguje greške i spaja odgovarajuće sekvence, a u drugom koraku razrešava ponovke tako što odvoji putanje koje dele lokalno preklapanje. Velvet predstavlja skup metoda za obradu kratkih očitavanja koje koriste strukturu De Brojnovog grafa i koje mogu da uklone greške. U slučaju da su na raspolaganju uparena očitavanja, Velvet razrešava veliki broj ponavljanja. U slučaju da nisu na raspolaganju, ponovci ne mogu da budu razrešeni kada su ponovljene niske duže od dužine k-grama. Velvet može da izvrši kvalitetno sklapanje sa kratkim očitavanjima kada je ukupna pokrivenost očitavanja visoka i bez referentnog genoma. Pomoću dodatnih informacija iz uparenih očitavanja Velvet razrešava većinu malih ponovaka. Osetljiv je na veličinu parametra k. Optimalna veličina zavisi od genoma, pokrivenosti očitavanja, kvaliteta očitavanja i od dužine očitavanja. Jedna mogućnost je isprobavanje raznih veličina za parametar k i izbor najbolje alternative.

3.2.1 Struktura i reprezentacija De Brojnovog grafa

U De Brojnovom grafu svaki čvor N predstavlja niz preklapajućih k-grama. Susedni k-grami se preklapaju na $(k-1)$ nukleotida. Informacija sačuvana u čvorovima je niska sastavljena od poslednjih nukleotida njihovih k-grama (slika 3.3). Ova niska karaktera predstavlja sekvencu čvora u oznaci $s(N)$.

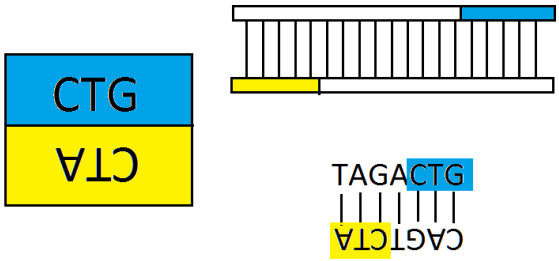


SLIKA 3.3: Primer reprezentacije De Brojnovog grafa

Izvor: [9] str. 3, slika 1

Čvorovi su prikazani kao pravougaonici i predstavljaju niz preklapajućih k-grama (na slici 3.3 je $k = 5$) pomoću njihovih poslednjih nukleotida. Čvorovi su dvostruki pravougaonici zato što predstavljaju i niz obrnuto komplementarnih k-grama. Povezani su usmerenim granama ako je sufiks poslednjeg k-grama ((*k-1*)-gram) izlaznog čvora identičan prefiksu prvog k-grama ulaznog čvora. Kako dva zalepljena čvora predstavljaju obrnuto komplementarne niske, svaka grana ima njemu simetričnu granu.

Svaki čvor N je povezan sa obrnuto komplementarnim čvorom od N, koji je zapravo obrnuta niska obrnuto komplementarnih k-grama. Vizualno objašnjenje je na slici 3.4. Ova reprezentacija garantuje da se preklapanja između očitavanja sa suprotnih lanaca DNK sekvenci uzimaju u obzir. Unija čvora i njemu simetričnog se zove blok. Kada se neka promena desi nekom čvoru, onda se ta promena uradi i simetrično na zalepljenom. Da bismo obezbedili da nijedan k-gram nije sam sebi obrnuto komplementaran, k mora da bude neparan broj (na primer AGCT je sam sebi obrnuto komplementaran) [9].

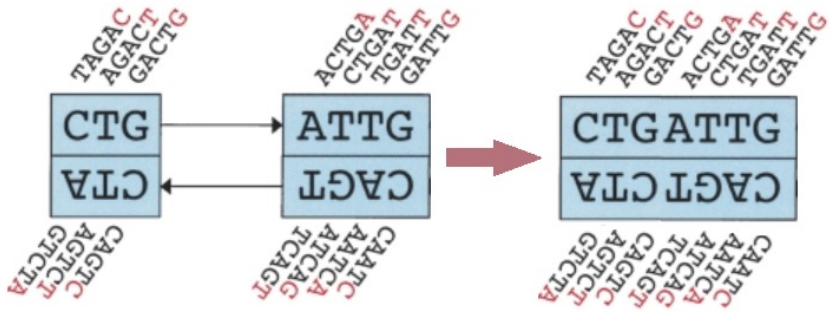


SLIKA 3.4: Vizualno objašnjenje bloka

Izvlačenje sekvence nukleotida sa putanje u De Brojnovom grafu je direktno čitanje. Ako je dat prvi k-gram prvog čvora na putanji, onda se nastavak dobija samo čitanjem sadržaja čvorova redom.

3.2.2 Pojednostavljivanje grafa i uklanjanje grešaka

Kad jedan čvor A ima tačno jednu izlaznu granu koja pokazuje na drugi čvor B, koji ima tačno jednu ulaznu granu (upravo ta grana), onda njihove blokove spajamo u jedan. Dva čvora na levom delu slike 3.3 primera grafa mogu da budu spojeni, zato što formiraju jedan niz preklapajućih k-grama, čime dobijamo jedan čvor CTGATTG (slika 3.5). Iterativno se niz čvorova koji odgovaraju prethodnom kriterijumu spajaju u jedan blok.



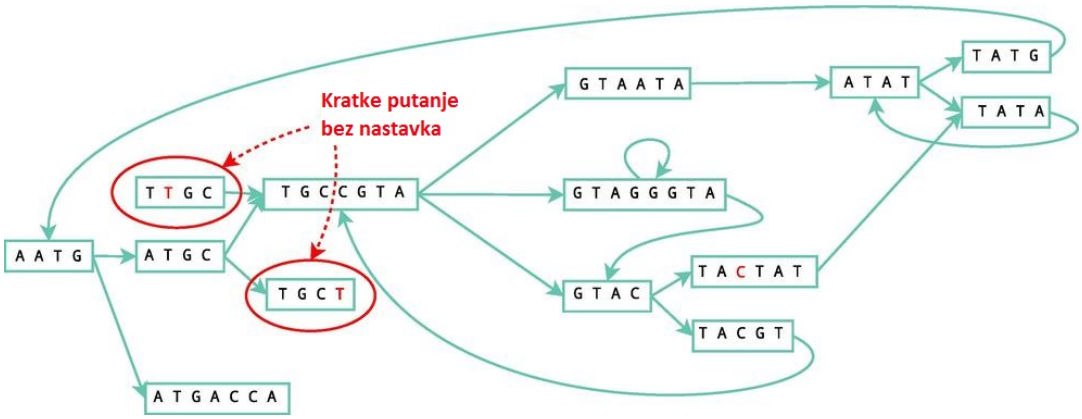
SLIKA 3.5: Spojeni čvorovi

Izvor: modifikacija slike [9] str. 3, slika 1

Prilikom uklanjanja grešaka Velvet se fokusira na topološke strukture, odnosno specifične podgrafove kao što su slepe putanje, balončići i nepravilne konekcije (*erroneous connection*) zbog grešaka u sekvenciranju ili zbog spojenih dalekih slepih putanja. Svaki od njih se uklanja iz grafa na sledeći način:

- Slepu putanju čini niz čvorova koji na jednom kraju nisu povezani ni sa čim, odnosno nemaju nijednu izlaznu ili nijednu ulaznu granu. Primer je na slici 3.6. Slepe putanje uklanjamo jedino ako su kraće od $2k$,

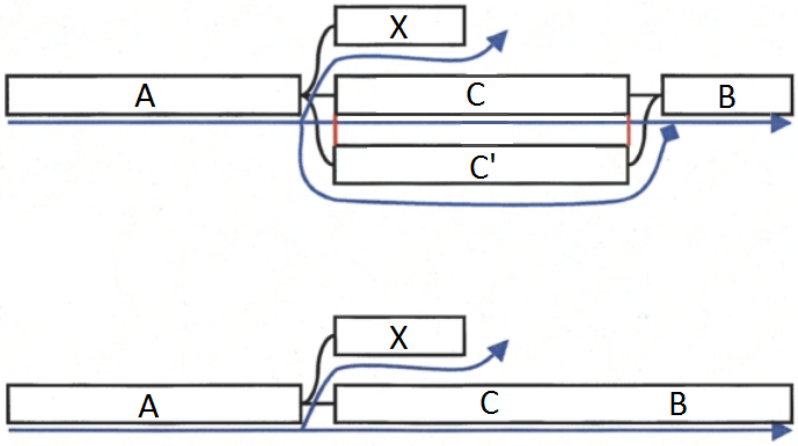
zato što one duže od $2k$ mogu da predstavljaju ispravnu sekvencu ili sekvencu sa nagomilanim greškama, a takve je vrlo teško razlikovati. Počevši od čvora gde počinje slepa putanja, ako postoji druga putanja koja je bolje pokrivena očitavanjima, onda slepu putanju uklanjamo. Ovaj proces izvodimo iterativno.



SLIKA 3.6: Uklanjanje slepih putanja

Izvor: https://en.wikipedia.org/wiki/Velvet_assembler, slika 3

- Uklanjanje balončića se vrši modifikacijom Dijkstrinog algoritma koji se zove *Tour Bus* algoritam. Slično kao i kod ABySS-a, ako su niske koje odgovaraju stranama balončića dovoljno slične, onda ih integrišemo. Brišemo jednu stranu balončića, ako važe uslovi da je izabrana strana manje pokrivena očitavanjima, kraća od unapred određenog praga. *Tour Bus* algoritam u nekim slučajevima slične putanje, odnosno nesavršene ponovke uglađuje (*smoothing out*) u identične ponovke. Putanje u balončićima se integrišu u slučaju da važe uslovi da obe putanja sadrže manje od 200 čvorova, sekvence koje odgovaraju putanjama su kraće od 100 baznih parova i sekvence su bar 80% identične. Na slici 3.7 C i C' su dovoljno slični. Kako je C pouzdanija, C' brišemo.



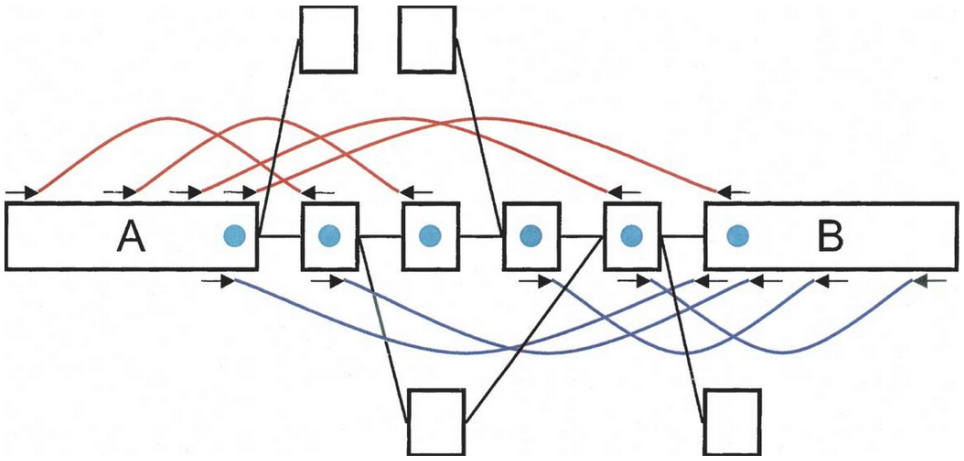
SLIKA 3.7: Uklanjanje balončića

Izvor: [9] str. 7, slika 2

- Uklanjanje nepravilnih konekcija (*erroneous connection*): ove nepoželjne konekcije ne prave podgrafove sa prepoznatljivim strukturama kao što su slepe putanje ili balončići. Velvet ove greške rešava tako što briše delove grafa koji nemaju određenu pokrivenost očitavanjima. Prag određuje korisnik na osnovu nacrta pokrivenosti posle uklanjanja balončića. Čvorovi koji su ostali posle *Tour Bus* algoritma a koji su sa niskim pokrivenostima očitavanja, sa velikom verovatnoćom sadrže greške, zbog lažnih preklapanja koji su kreirani usled eksperimentalnih grešaka.

3.2.3 Modul *Breadcrumb*

Ponovci istih fragmenata u strukturi genoma otežavaju sklapanje. Neophodno je razrešavanje ovih ponovaka u sekvenci da bismo ispravno produžili i povezali kontige kroz ponavljajuće regione koje generišu petlje u De Brojnovom grafu. Modul Velveta koji se zove *Breadcrumb* izvršava ovaj zadatak iskorišćavanjem informacija iz uparenih očitavanja.



SLIKA 3.8: Modul *Breadcrumb*

Izvor: [9] str. 13, slika 5

Ideja algoritma *Breadcrumb* modula je opisan na primeru. Posmatramo dve duge kontige A i B na slici 3.8, koje su napravljene posle uklanjanja grešaka, i povezane su sa nekoliko uparenih očitavanja (plave i crvene boje na slici). Mali pravougaonici su čvorovi na putanji između A i B ili su čvorovi iz De Brojnovog grafa koji su povezani sa prethodnim čvorovima. Putanja između dugih kontiga može da sadrži prekid iz više razloga. Naime, putanja nije jedinstvena zbog toga što međučvorovi predstavljaju ponavljajuće sekvence, i/ili zbog preklapanja sa nekim dalekim delom genoma i/ili zbog nekih nerešenih grešaka. Pronalaženje precizne putanje u grafu od čvora A do čvora B nije linearno zato što se posećuju i alternativne putanje. Označavamo čvorove koji sadrže jednu stranu uparenih očitavanja (na slici 3.8 označeno plavim krugovima) i putanju od čvora A do čvora B tražimo jedino među označenim čvorovima. Označavanjem dobijemo jedan jednostavniji podgraf u kojem je pronalaženje putanje u idealnom slučaju linearno.

Breadcrumb algoritam korišćenjem uparenih očitavanja uparuje čvorove čije su oznake duže od VUF-a, u nastavku dugački čvorovi. Za svaki dugačak čvor, *Breadcrumb* označava čvorove koje sadrže drugi deo para očitavanja, čiji je prvi deo u spomenutom čvoru. Proširuje se kontiga koja odgovara dugačkom čvoru kontigama koje su sadržile drugi deo para očitavanja, dok je proširenje moguće na jedinstven način, odnosno, dokle god nema više mogućih opcija za naredni čvor u nizu za proširenje. Rezultat celokupnog Velvet sklapanja su kontige koje dobijemo posle izvršenja *Breadcrumb* algoritma.

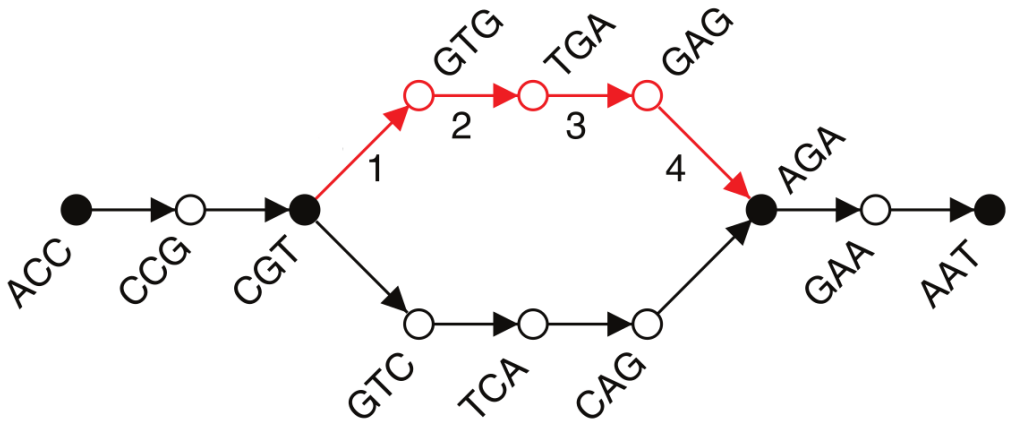
3.3 Asembler SPAdes (*St. Petersburg genome assembler*)

Asembler SPAdes može da se koristi i za standardno (višećelijsko) sklapanje i za sklapanje jedne ćelije (*single-cell assembly*)[10] [8]. SPAdes koristi modifikovanu verziju De Brojnovog grafa. Da bismo prikazali kako je SPAdes implementiran, potrebno je da uvedemo nekoliko novih pojmova:

1. *H-očitavanje*
   Kažemo da je usmerena putanja u De Brojnovom grafu *h-putanja* ako su njeni početni i završni čvorovi čvorovi divergencije (uvedeni u potpoglavlju 3.1.2), a ostali čvorovi putanje nisu. Svaka grana u grafu pripada tačno jednoj h-putanji. Svakoj h-putanji odgovara jedno *h-očitavanje*, odnosno niska koja je predstavljena tom putanjom.

   Na slici 3.9 je primer dekompozicije De Brojnovog grafa u h-putanje. Na ovoj slici je prikazan De Brojnov graf nad očitavanjima $ACCGTCAGAAT$ i $ACCGTGAGAAT$ sa parametrom $k = 4$, u kom su grane označene tetragramima a čvorovi trigramima. Čvorovi divergencije su predstavljeni punim krugovima a ostali čvorovi praznim krugovima. H-putanja u crvenoj boji CGT / GTG / TGA / GAG / AGA definiše h-očitavanje $CGTGAGA$. Ostale h-putanje sa slike 3.9 su ACC / CCG / CGT, CGT / GTC / TCA / CAG / AGA i AGA / GAA / AAT i odgovaraju h-očitavanjima $ACCGT$, $CGTCAGA$ i $AGAAT$.

[10] Sekvenciranje jedne ćelije (*Single cell sequencing (SCS)*) je moćan skup tehnologija za proučavanje retkih ćelija i opisivanje kompleksnih populacija. Umesto analiziranja uzoraka tkiva, koji su kompozicije više miliona ćelija, detaljno i obimno se analizira individualna ćelija [17]. Ovakvo optimizovano sekvenciranje nove generacije pruža više informacija o razlikama između ćelija [18]. Većina bakterija iz okeana ili iz tela čoveka ne mogu da se kloniraju u laboratoriji, stoga ne mogu ni da se sekvenciraju koristeći metodu sekvenciranja nove generacije (*NGS technology*). Ove probleme rešava sekvenciranje jedne ćelije (*single cell sequencing (SCS)*)

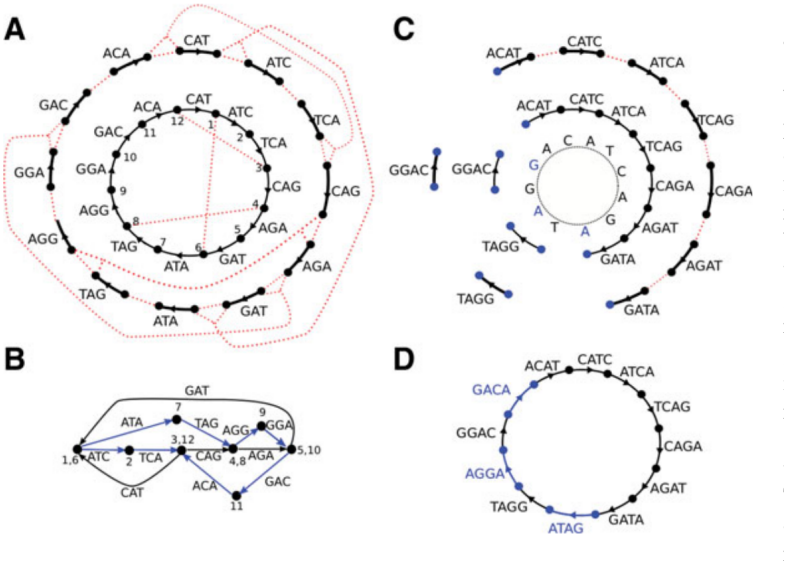


SLIKA 3.9: De Brojnov graf nad očitavanjima $ACCGTCAGAAT$ i $ACCGTGAGAAT$. H-putanja u crvenoj boji $CGT/GTG/TGA/GAG/AGA$ definiše h-očitavanje $CGTGAGA$.

Izvor: [8] str. 458, slika 1

2. *Fleksibilni De Brojnov graf*
   Izbor veličine k-grama utiče na konstukciju i na strukturu De Brojnovog grafa. Ako parametar k ima manju vrednost, graf će sadržati petlje, zato što ponovci u strukturi genoma koji su duži nego parametar k ne mogu da budu razrešeni. Ako parametar k ima veću vrednost, može se desiti da se preklapanja između očitavanja ne detektuju posebno u regionima sa niskim pokrivenošću. Kada je parametar k mali, De Brojnov graf je komplikovan, a u slučaju da je veliki, graf je fragmentisan. Fleksibilni De Brojnov graf rešava prethodni problem omogućavanjem korišćenja više različitih vrednosti parametra k u kombinaciji.

   Ako je R skup očitavanja onda definišemo $R_{k-d,k}$ kao skup svih h-putanja iz standardnih De Brojnovih grafova nad k-gramima, (k-1)-gramima, (k-2)-gramima do (k-d)-gramima. Fleksibilni De Brojnov graf sa oznakom $DB(R, k-d, k)$ je De Brojnov graf nad unijom dva skupa k-grama, gde prvi predstavlja sve k-grame iz skupa R (skup očitavanja) a drugi sve k-grame iz skupa h-očitavanja $R_{k-d,k}$ dobijenih iz raznih standardnih De Brojnovih grafova. Primer konstrukcije fleksibilnog De Brojnovog grafa je na slici 3.10.



SLIKA 3.10: Standardni i fleksibilni De Brojnov graf za kružni genom $CATCAGATAGGA$. Kružni genom je pokriven skupom očitavanja $R = \{ACAT, CATC, ATCA, TCAG, CAGA, AGAT, GATA, TAGG, GGAC\}$. Ovaj skup predstavlja devet tetragrama od mogućih dvanaest iz kružnog genoma, dok tri tetragrama nedostaju ($ATAG, AGGA, GACA$). Međutim, svaki mogući trigram iz kružnog genoma je pokriven očitavanjima iz skupa R uključujući i trigrame iz tetragrama koji nedostaju.

(A) Spoljašnji krug predstavlja trigrame iz skupa očitavanja R. Unutrašnji krug predstavlja rezultat nekoliko spajanja (ne svih). Crvene linije u unutrašnjem krugu su povezani čvorovi koji će biti spojeni.

(B) Standardan De Brojnov graf nad trigramima (komplikovan). H-putanje dužine 2 su predstavljene plavom bojom, odgovarajuća h-očitavanja su $R_{3,4} = \{ATAG, AGGA, GACA\}$. Unija $R_{3,4} \cup R$, to jest unija ovih h-očitavanja i skupa očitavanja je zapravo skup svih mogućih dvanaest tetragrama kružnog genoma.

(C) U spoljašnjem krugu za svaki od devet tetragrama iz skupa očitavanja postoji jedna posebna grana. Središnji krug predstavlja standardan De Brojnov graf nad tetragramima skupa R (fragmentisan). Unutrašnji krug je sam genom.

(D) Fleksibilni De Brojnov graf $DB(R, 3, 4)$, dobijen kao običan De Brojn graf nad tetragramima skupa $R \cup R_{3,4}$

Izvor: [8] str. 459, slika 2

Slika 3.10 pokazuje standardan De Brojnov graf nad trigramima (komplikovan), standardan De Brojnov graf nad tetragramima (fragmentisan) i fleksibilni De Brojnov graf $DB(R, 3, 4)$ koji nije ni komplikovan ni fragmentisan.

3. *K-bigram*
   K-bigram *(k-bimer)* definišemo kao trojku $(a|b,d)$ koji se sastoji od k-grama a i b i od celog broja d koji predstavlja procenu udaljenosti između specifičnih očitavanja a i b u genomu.

4. *Dvojna grana*
   Neka je C jedan (nepoznat) Ojlerov ciklus u fleksibilnom grafu G. Definišemo dvojnu granu (*biedge*) $(a|b,d)$, gde su a i b grane u grafu G i d je procena udaljenosti između grane a i b u ciklusu C. Dvojne grane odgovaraju k-bigramima, ali ova definicija ne zavisi od niski karaktera.

3.3.1 Faze sklapanja

SPAdes asembler se sastoji od četiri glavne faze:

- Faza 1: Konstrukcija grafa sklapanja.
  Za dati skup parova očitavanja napravi fleksibilni graf sklapanja i sačuva informacije gde se originalna očitavanja mapiraju na grafu sklapanja.
- Faza 2: Prilagođavanje uparenih k-grama.
  Izračunaju se h-dvojne grane[11] pomoću niza transformacija uparenih očitavanja.
- Faza 3: Konstrukcija uparenog grafa sklapanja pomoću koncepta uparenog De Brojnovog grafa.
  Koristeći dobijene h-dvojne grane kontruiše se upareni graf sklapanja.
- Faza 4: Konstrukcija kontige.
  Ispisuju se kontige iz uparenog grafa sklapanja.

U narednim poglavljima detaljnije opisujemo faze SPAdes asemblera.

Faza 1: Konstrukcija i pojednostavljenje grafa sklapanja.

Svaki NGS asembler se bavi konstrukcijom grafa sklapanja, obično i pojednostavljenjem De Brojnovog grafa (na primer uklanjanje balončića kod Velveta). SPAdes uvodi novi pristup za konstrukciju grafa sklapanja pomoću fleksibilnog De Brojnovog grafa. Za dati skup parova očitavanja konstuiše fleksibilni graf sklapanja i održava strukturu podataka za čuvanje informacija gde se mapiraju originalna očitavanja pored grafa sklapanja.

SPAdes je razvio jednu novu ideju za uklanjanje balončića i koristi modifikovanu verziju ideje iz rada Chitsaz-a [19] i iterativni De Brojnov graf (pristup iz rada Peng [20]). Za dati skup popravljenih uparenih očitavanja SPAdes prvo konstruiše fleksibilan De Brojnov graf i izračuna dubinu pokrivenosti očitavanjima za svaku h-putanju u grafu. Sledi transformacija parova učitavanja u dvojne grane i izrada pomoćnih histograma. Na kraju, SPAdes

[11] H-dvojna grana je dvojna grana $(a|b,d)$, gde su a i b početne grane nekim h-putanjama.

pojednostavljuje graf uklanjanjem balončića, slepih putanja i himeričnih očitavanja[12].

Fleksibilan De Brojnov graf je poboljšenje u odnosu na standardan De Brojnov graf nad očitavanjima bez greške, ali u slučaju da ima očitavanja sa greškama potrebne su neke modifikacije. Uklanjanje balončića i slepih putanja je neophodno.

Korekcija grešaka u grafu sklapanja zasnovana je na ideji da greške u očitavanjima obično naprave podgrafove sa sledećim specifičnim strukturama u De Brojnovom grafu:

1. Pogrešno pročitan nukleotid i indeli na sredini nekog očitavanja obično vode do balončića. Rezultat malih razlika između nesavršenih ponovaka takođe može da bude balončić.

2. Greške pri kraju očitavanja mogu da vode do slepih putanja: kratke h-putanje sa izlaznim stepenom nula.

3. Himerična očitavanja mogu da naprave pogrešne veze u grafu. Oni mogu da budu rezultat identičnih grešaka blizu kraja jednog očitavanja i blizu početka drugog očitavanja.

4. Ulazni podaci obično sadrže očitavanja niskog kvaliteta koja se ne mapiraju na genom, nego obično formiraju kratke i izolovane h-putanje sa niskom pokrivenošću. SPAdes ih uklanja nakon prethodnih pojednostavljenja grafa.

SPAdes čuva informacije o uklanjanju prethodnih struktura. Na ovaj način omogućava korišćenje metoda obrnute pretrage (*backtracking*) za praćanje kako se očitavanja poravnavaju sa grafom sklapanja pri pojednostavljenju grafa.

Faza 2: Iskorišćavanje uparenih očitavanja

SPAdes koristi informacije iz uparenih očitavanja prilikom regulisanja k-bigrama. Transformiše se skup k-bigrama dobijen u 1. fazi sa prilično netačnim udaljenostima u skup regulisanih k-bigrama sa tačnim ili skoro tačnim procenama udaljenosti. Regulacija k-bigrama podrazumeva zamenu originalnih k-bigrama sa virtualnim i regulisanim k-bigramima izvlačenjem tačne procene udaljenosti između k-grama u genomu pomoću zajedničke analize histograma udaljenosti i putanja u grafu sklapanja [8].

Udaljenost dva h-očitavanja u genomu može da se proceni pomoću k-bigrama koji ih povezuju (prvi k-gram iz k-bigrama može da se pridruži prvom h-očitavanju a drugi drugom).

[12]Himerična očitavanja su očitavanja koja sadrže greške kao rezultat sekvenciranja kontaminiranih uzoraka.

Faza 3: Konstrukcija uparenog grafa sklapanja pomoću koncepta uparenog De Brojnovog grafa

SPAdes konstruiše upareni graf sklapanja pomoću koncepta uparenog De Brojnovog grafa. Neka je C jedan (nepoznat) Ojlerov ciklus u fleksibilnom grafu G. Ciklus C je konzistentan sa dvojnom granom $(a|b,d)$ ako postoje primeri grana a i b u ciklusu C na udaljenosti d. Za dati skup DG dvojnih grana, ciklus C je DG-konzistentan ako je ciklus konzistentan sa svim dvojnim granama iz skupa DG.

Pomoću prethodno uvedenih koncepata, nalaženje Ojlerove putanje u standardnom De Brojnovom grafu svodimo na problem nalaženja Ojlerovog ciklusa koji je konzistentan dvojnim granama (*Biedge Consistent Eulerian Cycle (BCEC)*). Drugim rečima, za zadati Ojlerov fleksibilan graf G i skup dvojnih grana DG potrebno je naći DG-konzistentan Ojlerov ciklus u grafu G.

Faza 4: Konstukcija kontiga

SPAdes konstuiše kontige iz prethodno dobijenog uparenog grafa sklapanja svođenjem na problem nalaženja Ojlerovog cilusa. Štaviše, koristeći pomoćne strukture i metode obrnute pretrage nad njima računa gde su originalna očitavanja poravnata sa dobijenim kontigama.

Poglavlje 4

Rezultati

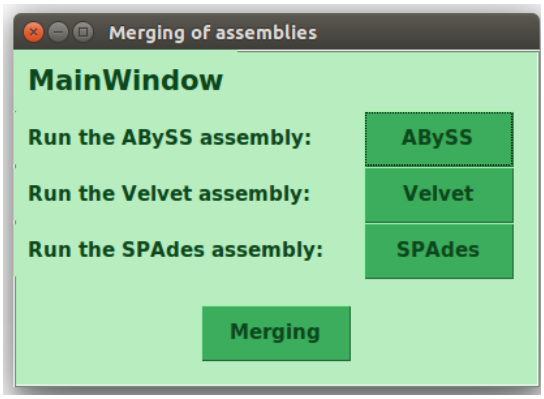
4.1 Integrisan program sklapanja i integrisanje sa grafičkim korisničkim interfejsom

Napravljen je GUI okruženje za postojeće asemblere ABySS, Velvet i Spades i integratora GAM-NGS. U okviru programa

4.1.1 Grafički korisnički interfejs

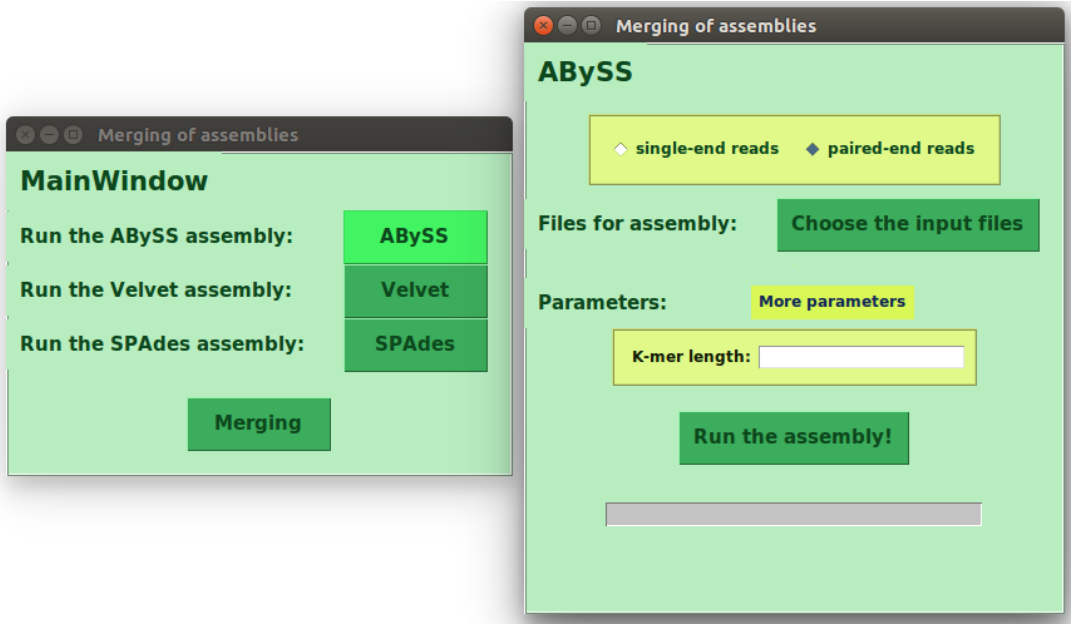
Program je namenjen naučnicima iz oblasti genetike sa ograničenim znanjem *Linux* sistema za sklapanje DNK sekvenci pomoću alata AbySS, Velvet i SPAdes i inegrisanje rezultata različitih asemblera pomoću GAM-NGS integratora putem korisničkog interfejsa. Korisnicima je omogućavano evaluiranje rezultata posle integrisanja komoću prethodno navedenih statistika (FRCi Quast). U nastavku je predstavljen izgled korisničkog interfejsa.

Pri pokretanju programa pojavljuje se glavni prozor koji je predstavljen na slici 4.1.



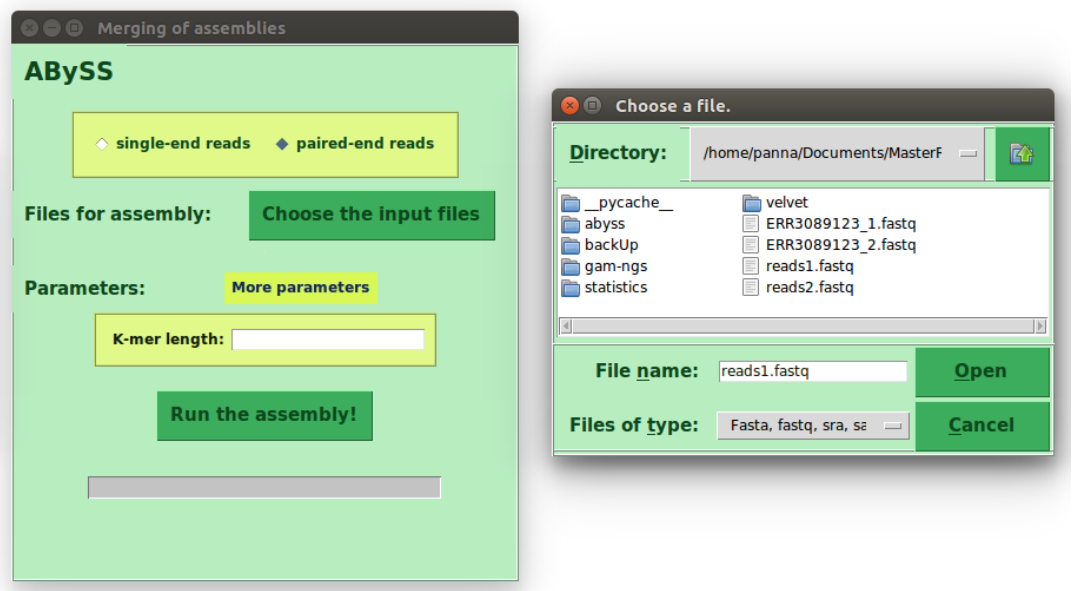
SLIKA 4.1: Glavni prozor

Izborom dugmeta sa oznakom asemblera koji korisnik želi da koristi otvara se prozor za zadavanje ulaza i parametara za taj asembler. Očitavanja za sklapanje su grupisana po tipu u odvojene datoteke. Obično ulazne datoteke koje sadrže očitavanja sadrže samo jednostrana očivanja (*single-end reads*) ili samo jednu stranu uparenih očitavanja (*paired-end reads*). Pre izbora ulaznih datoteka prvo izaberemo tip (slika 4.2).



SLIKA 4.2: Prozor za ABySS asembler

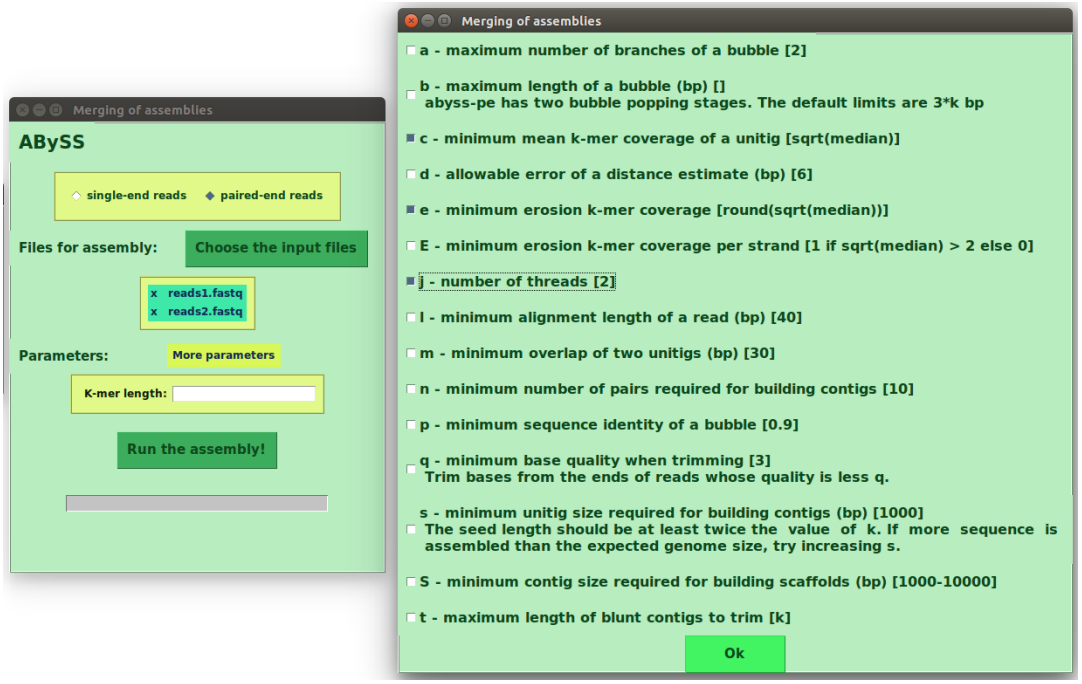
Nakon izbora vrste ulaznih datoteka otvara se prozor za njihov izbor klikom na dugme "Izaberi ulazne datoteke" (*Choose the input files*), kao što se vidi na slici 4.3. Ako su u pitanju datoteke koje sadrže uparena očitavanja, onda su po konvenciji leve strane očitavanja (*forward read*) su u datoteci sa sufiksom 1, a desne strane (*backward read*) sa sufiksom 2. Na primer na slici 4.3 *reads1.fastq* i *reads2.fastq*.



SLIKA 4.3: Prozor za izbor ulaznih datoteka. Leve strane uparenih očitavanja su po konvenciji u datoteci sa sufiksom 1, a desne strane sa sufiksom 2.

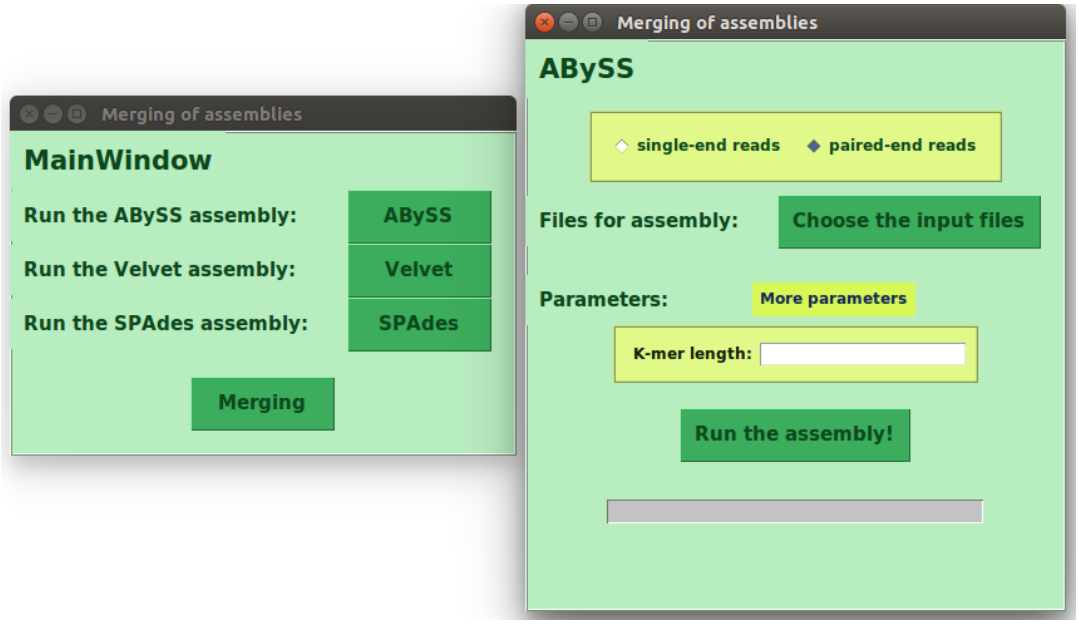
Posle izbora ulaznih datoteka korisnik može da odabere dodatne parametre za sklapanje klikom na dugme "Više parametara" (*More parameters*). Novi

prozor se pojavljuje sa listom mogućih dodatnih parametara, kao što se vidi na slici 4.4, na kojoj su navedeni: parametri, njihovi kratki opisi i podrazumevane vrednosti parametara u uglastim zagradama.



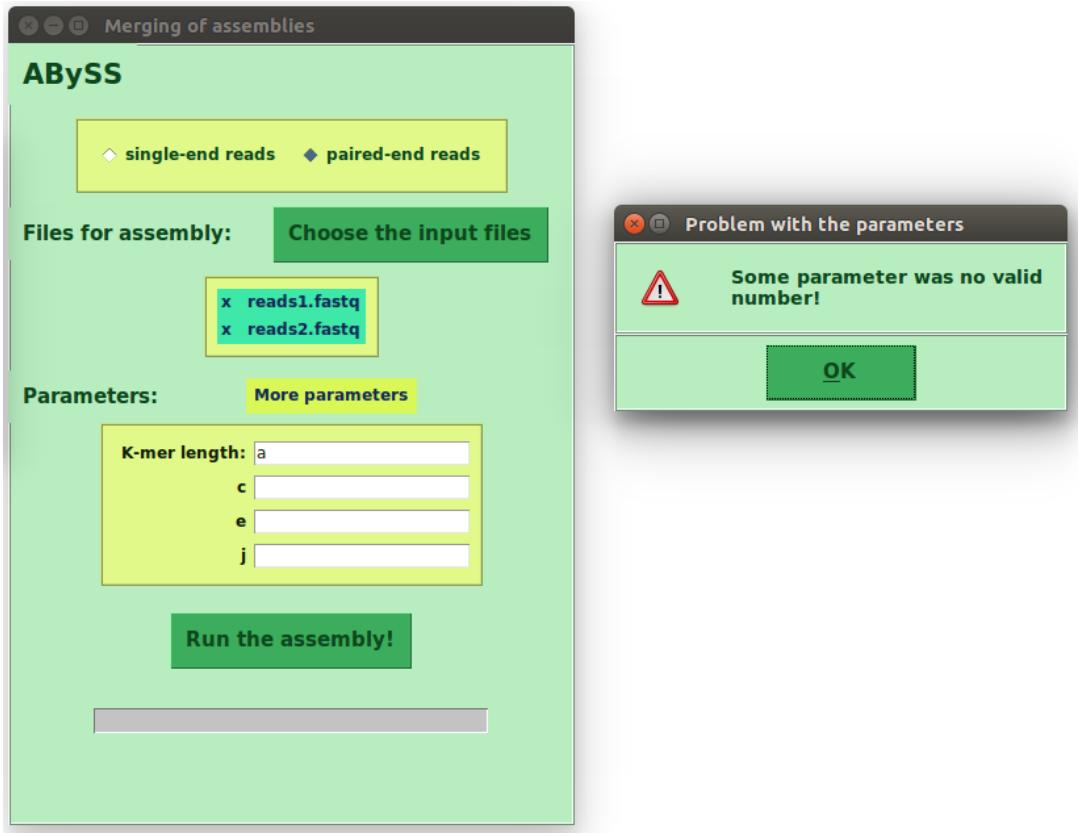
SLIKA 4.4: Prozor sa dodatnim parametrima za ABySS asembler. Naznačene su oznaka, kratak opis i u uglastim zagradama podrazumevana vrednost parametra.

Klikom na dugme "Ok" na dnu prozora pojavljuju se dodatni parametri u prethodnom prozoru pored obaveznog parametra za dužinu k-grama kao što se vidi na slici 4.5.



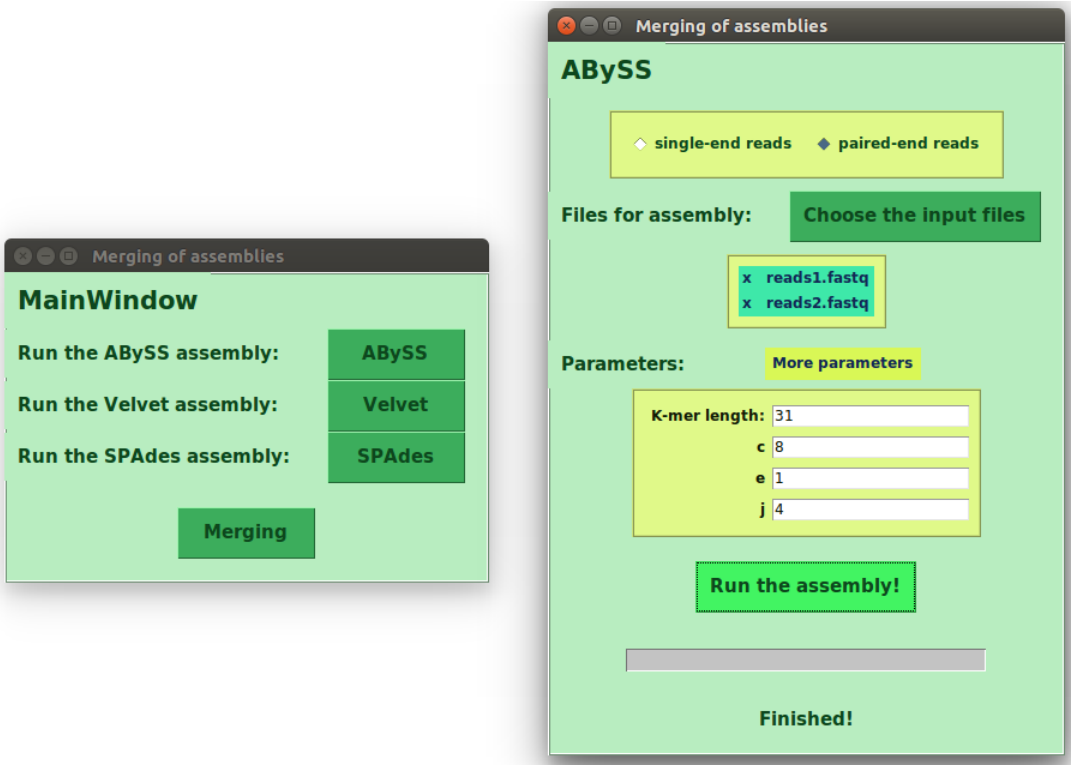
SLIKA 4.5: Izabrani dodatni parametri se pojavljuju u okviru za parametre

Sledeći korak je unošenje vrednosti za parametre i sklapanje. Pre pokretanja asemblera, proverava se validnost parametara. Ukoliko se naiđe na problem sa nekim parametrom pojavljuje se novi prozor, kao na slici 4.6, koji obaveštava o grešci. Program ne dozvoljava da se sklapanje pokrene dok nisu svi parametri ispravni. Na primer, dužina k-grama mora biti neparan ceo broj iz opsega [15, 64].



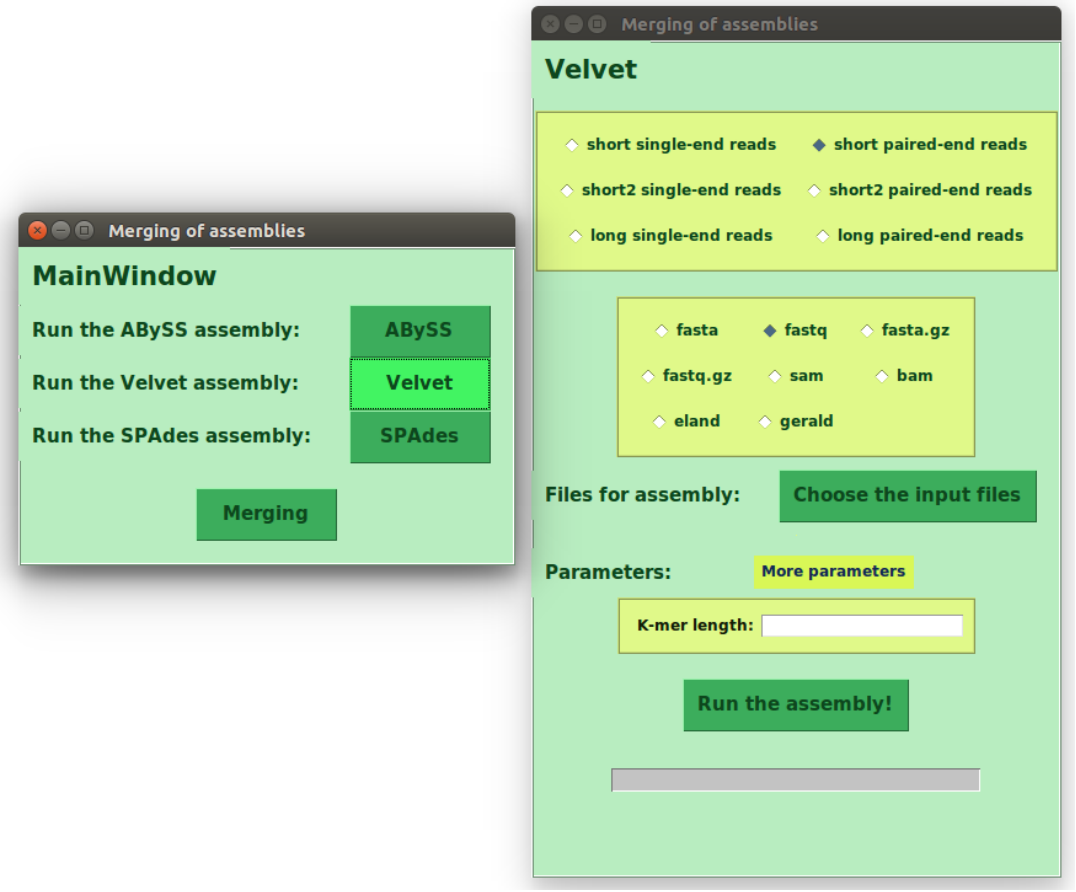
SLIKA 4.6: U slučaju primećene greške prilikom provere validnosti parametara pojavi se upozorenje sa opisom greške.

Po završetku sklapanja DNK sekvenci pomoću Abyss-a zaustavlja se linija napredovanja (*progress bar*) ispod dugmeta "Ok" i pojavljuje se poruka "Završeno!" (*Finished!*) (slika 4.7).



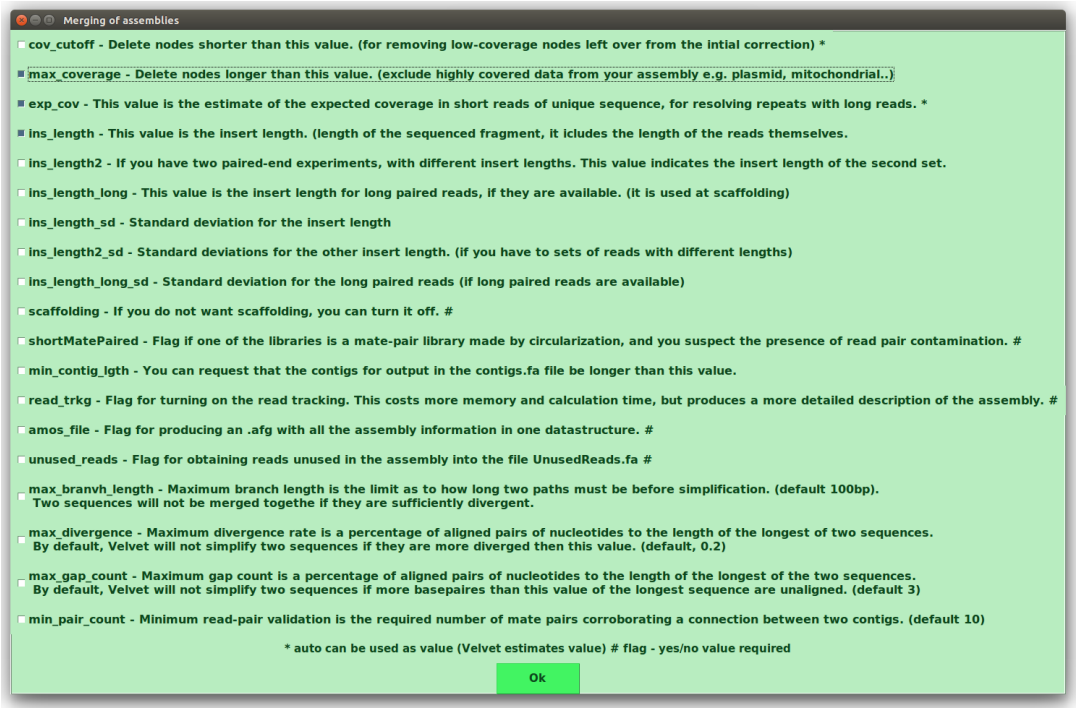
SLIKA 4.7: Na dnu prozora pojavi se poruka "Završeno!" (*Finished!*) po završetku sklapanja.

Rad sa asemblerom Velvet je veoma sličan. Razlika je u tome da treba odabrati ne samo tip, nego i ekstenziju ulaznih datoteka, pre izbora datoteke iz sistema (slika 4.8).



SLIKA 4.8: Prozor za Velvet asembler

U slučaju asemblera Velvet dodatni parametri su drugačiji u odnosu na Abyss. Za parametre koji su označeni zvezdom (*) vrednost može da bude broj ili ključna reč "auto". U drugom slučaju program će sam da proceni vrednosti parametra pri radu. U slučaju da je parametar označen oznakom taraba (#), u pitanju je fleg (*flag*) i vrednosti mogu da budu samo ključne reči **da** ili **ne** na engleskom jeziku (*yes/no*). Značenje ovih oznaka je na dnu prozora za dodatne parametre kao što se vidi na slici 4.9.



SLIKA 4.9: Prozor za dodatne parametre

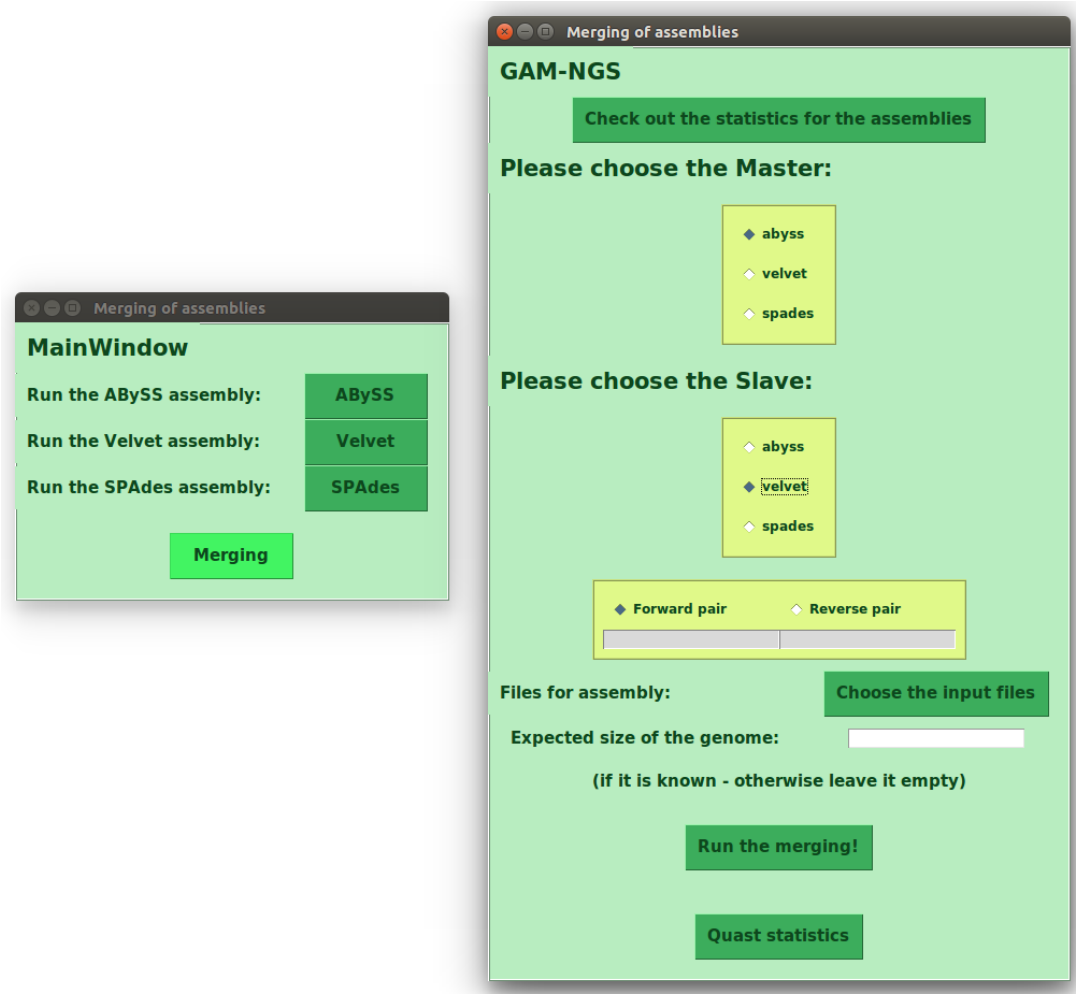
Treći asembler u okviru rada je SPAdes. Na slici 4.10 se vide dva prozora u GUI-ju koji odgovaraju ovom asembleru. Potrebno je odvojiti ulazne datoteke u skupove, takozvane biblioteke. Pre izbora svake ulazne datoteke prvo se izabere redni broj biblioteke (*library number*) kojoj pripada, nakon toga tip ulazne datoteke u gornjem žutom okviru i podtip u donjem okviru, ako izabranom tipu pripadaju podtipovi. Izbor opcija u donjem pravougaoniku je omogućen kada izabrani tip ima podtipove i neophodna je specifikacija, u ovom slučaju donji pravougaonik je prikazan u levom prozoru, odnosno izbor opcija u njemu je omogućen, a inače nije potreban (desni prozor). Na primer, ulazna datoteka koja sadrži uparena očitavanja može da sadrži samo leve parove, samo desne parove očitavanja ili kombinacije, pa je neophodna specifikacija, a ulazna datoteka koja sadrži jednostrana očitavanja nema ove podtipove (desni prozor).

SLIKA 4.10: Prozor za SPAdes asembler

Izbor dodatnih parametara je identičan kao u prethodna dva asemblera. Lista dodatnih parametara je prikazana na slici 4.11.

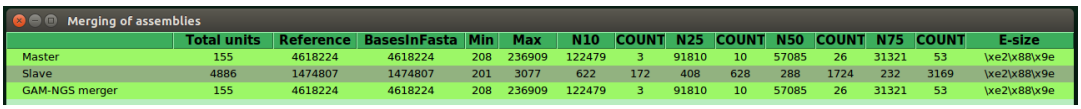
SLIKA 4.11: Prozor za dodatne parametre

Integrisanje rezultata sklapanja se vrši integratorom GAM-NGS. Prozor u koji korisnik može da unese odgovajuće parametre pojavljuje se nakon pritiska na dugme "Integrisanje" (*Merging*). Kao što se vidi na slici 4.12 korisnik treba da odabere koji asembler će biti glavni (*master*) a koji će biti sporedni (*slave*). U sledećem koraku se zadaju ulazne datoteke. Kao što je ranije naznačeno, u slučaju uparenih očitavanja odvojene datoteke sadrže dva dela očitavanja. Pri izboru ulaznih datoteka korisnik mora da označi da li odgovarajuća datoteka sadrži leve krajeve očitavanja uparenih krajeva ili desne.



SLIKA 4.12: Prozor za integrator GAM-NGS

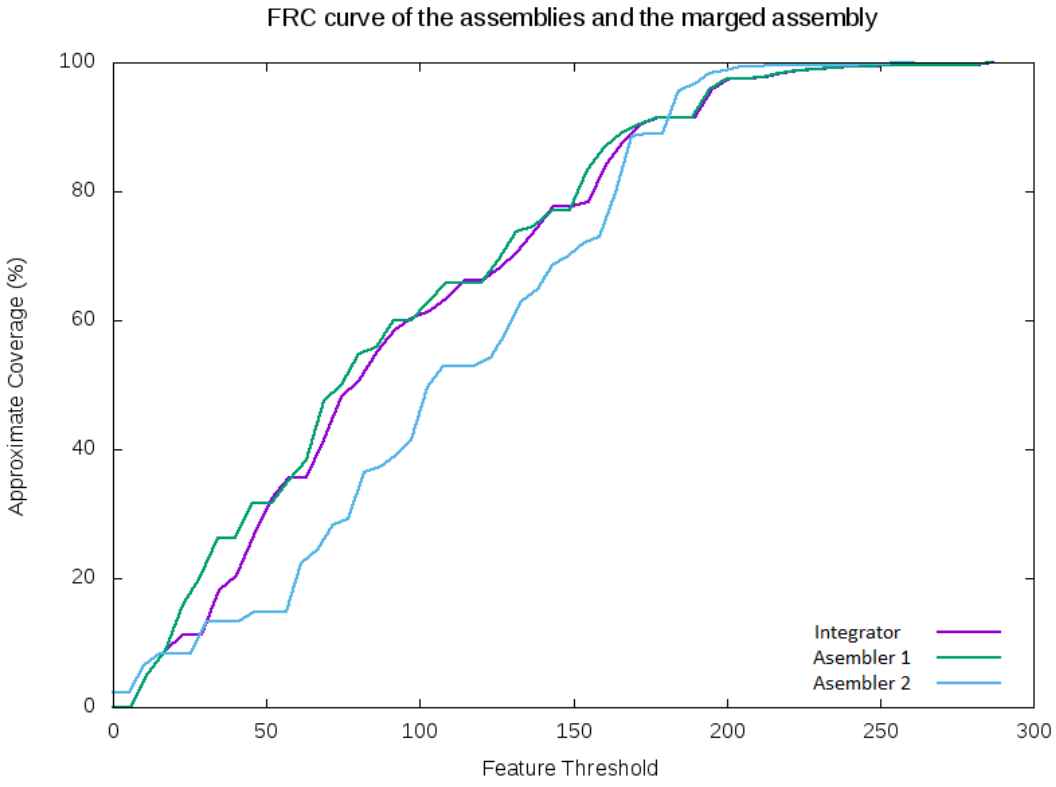
Ukoliko se zna procena očekivane veličine genoma, možemo da je zadamo u označeno ulazno polje zarad naprednije analize rezultata asemblera i integratora.



	Total units	Reference	BasesInFasta	Min	Max	N10	COUNT	N25	COUNT	N50	COUNT	N75	COUNT	E-size
Master	155	4618224	4618224	208	236909	122479	3	91810	10	57085	26	31321	53	\xe2\x88\x9e
Slave	4886	1474807	1474807	201	3077	622	172	408	628	288	1724	232	3169	\xe2\x88\x9e
GAM-NGS merger	155	4618224	4618224	208	236909	122479	3	91810	10	57085	26	31321	53	\xe2\x88\x9e

SLIKA 4.13: Rezultat analize integratora

Kada se završi proces integrisanja ocenjuju se rezultati pomoćnim alatima GAGE statistikom i FRC krivom. Rezultat se pojavljuje u novom prozoru. Deo tog prozora koji odgovara GAGE statistici ilustrovan je na slici 4.13, a deo koji ilustruje FRC krivu na slici 4.14. Koncepti i parametri koji se dobijaju objašnjeni su u sledećem poglavlju 4.2.



SLIKA 4.14: FRC krive rezultata asemblera i integratora

4.2 Testiranje i evaluacija

4.2.1 GAGE - (*Genome Assembly Gold-standard Evaluations*)

GAGE statistike (*Genome Assembly Gold-standard Evaluations*) predstavljaju metode koje opisuju performanse raznih asemblera i integratora u odnosu na dobijene kontige i skafolde odabranog genoma [21]. Služe za poređenje rezultata sklapanja i integrisanja na podacima dobijenim *large-scale* sekvenciranjem nove generacije.

Korišćene metrike su: broj, N50 veličina i E-veličina(*E-size, error-corrected sizes*) kontiga i skafolda.

4.2.2 N25, N50, N75 vrednosti

Postupak dobijanja statističke vrednosti N25, N50 i N70 je objašnjen na primeru vrednosti N50. Sve što sledi važi i za vrednosti N25 i N75 analogno.

N50 je mera koja opisuje kvalitet sklapanog genoma koji je predstavljen kontigama raznih dužina. N50 statistika je slična srednjoj vrednosti i medijani dužine kontiga pri čemu veću težinu dobijaju duže kontige. Ovu statistiku, koja je u širokoj upotrebi, možemo da opišemo kao težinsku srednju vrednost, tako da je 50% celog sklapanja sadržano u kontigama čija je dužina veća ili jednaka od ove vrednosti [22]. Broj kontiga čija je dužina veća od vrednosti N50 označava se sa L50.

N50 statistiku definišemo kao meru kvaliteta sklapanja u odnosu na neprekidnost (*contiguity*). Za dati skup kontiga, statistiku N50 definišemo kao maksimalnu dužinu najkraće kontige od svih podskupova kontiga čija suma dužina predstavlja 50% dužine celokupne dužine genoma. Ukupan broj baza iz svake kontige duže od N50 je približno jednak ukupnom broju baza iz svake kontige kraće od N50. Na primer, ako posmatramo 9 kontiga dužina $\{2, 3, 4, 5, 6, 7, 8, 9, 10\}$, njihova suma je 54, polovina sume je 27. Neka je veličina genoma 54, 50% ovog sklapanja bi bilo 27 (polovina dužine genoma), što se može dobiti na više načina. Jedan način da dobijemo 27 je zbir $10 + 9 + 8 = 27$. Kako su ovi brojevi najveći iz skupa dužine kontiga, ne možemo da nađemo drugi zbir tako da je najkraća kontiga duža od 8. Prema definiciji $N50 = 8$ je dužina najkraće kontige od svih onih, koji sadrže polovinu sekvence predstavljenog genoma. Dakle, kada upoređujemo N50 vrednosti raznih sklapanja, veličina sklapanja mora da bude ista, da bi upoređivanje bilo smisleno.

4.2.3 E-veličina (*E-size*)

E-veličina je statistika izračunata nad kontigama, koja je tako dizajnirana da odgovara na pitanje: ako nasumično izaberemo jednu lokaciju (jednu bazu u DNK) u referentnom genomu, koja je očekivana veličina kontige koja sadrži tu lokaciju [21]. E-veličinu računamo kao: $\Sigma G = \frac{L_C \times 2}{G}$, gde L_C je dužina kontige C, a G je dužina genoma procenjena kao suma dužina svih kontiga.

4.2.4 FRC kriva

Ideja konstruisanja FRC krive je inspirisana ROC krivama [23]. ROC kriva [24] grafički predstavlja jedan binarni klasifikacioni sistem kod kog prag diskriminacije varira (*discrimination threshold*). Prag diskriminacije je očekivana verovatnoća pozitivnog ishoda za binarni klasifikacioni sistem.

FRC kriva[25] (*feature-response curve*) opisuje osetljivost asemblera. Koristi pogrešno poravnata uparena očitavanja i partner-uparena očitavanja za identifikaciju spornih regiona, koje zovemo atributima (*feature*) [26]. Oslanja se na princip da može predvideti preciznost sklapanja pomoću identifikacije i prebrojavanja atributa na svakoj kontigi. U atribute spadaju:

- regioni sa niskom dubinom pokrivenosti
- regioni sa visokom dubinom pokrivenosti
- regioni sa velikim brojem uparenih očitavanja čije strane su poravnate u suprotnoj orijentaciji

- regioni u kojima je u velikom broju poravnata samo jedna strana uparenih očitavanja
- i druge

FRC kriva opisuje u kojoj meri sklapanje sadrži delove genoma (njegovu pokrivenost - *coverage*), kao funkciju u zavisnosti od broja atributa. Dakle, za svaku kontigu prebroje se svi atributi. Kontige se sortiraju po veličini od najveće do najmanje. Za svaki prag atributa (*feature threshold*) (tačka sa x-ose) samo se najduže kontige, čiji je ukupan broj atributa manji nego izabran prag, koriste za izračunavanje procena pokrivenosti genoma (tačka sa y-ose). Tako se dobijaju tačke FRC krive. Ilustracija FRC krive je prikazana na slici 4.14. Rezultujuća kriva je slična ROC krivi. Sklapanje predstavljeno najstrmijom krivom verovatno sadrži najmanji broj grešaka u sklapanju.

FRC može da se koristi kao metrika za upoređivanje kvaliteta raznih sklapanja više asemblera. FRC ne zahteva referentni genom za validaciju (osim procene veličine genoma, ako je poznata). Zbog ovog svojstva ovaj alat je veoma koristan kod projekata koji koriste *de novo* sklapanje.

4.2.5 QUAST (*Quality Assessment Tool*)

QUAST je skup metoda koji računa metrike nad sklapanim genomima i pogodan je za proveravanje kvaliteta dobijenih sklapanja i integrisanja. Ovaj program se može koristiti i sa i bez referentnog genoma [27]. Quast računa:

- Broj kontiga ($\geq x$ bp), za razne brojeve x.
- Ukupnu dužinu kontiga ($\geq x$ bp).
- Broj kontiga: ukupan broj kontiga u sklapanju/integrisanju.
- Ukupnu dužinu: ukupan broj baznih parova (*bp*) u sklapanju/integrisanju.
- Dužinu referentnog genoma.
- CG (%): Ukupan broj nukleotida G (guanin) i C (citozin) u sklapanju, podeljen sa ukupnom dužinom sklapanja.
- CG referentnog genoma.
- N50 broj.
- NG50: slično N50-u dužina kontiga, tako da kontige te dužine i duže predstavljaju 50% dužinu referentnog genoma (umesto sklapanog genoma).
- Procenat poravnatih očitavanja uz sklapanje/integrisanje (%).
- Procenat poravnatih očitavanja uz referentnog genoma (%).
- Srednju vrednost dubine pokrivenosti.

4.3 Rezultati

Program je testiran na sirovim podacima sekvenciranja organizama *Salmonella enterica* i *Staphylococcus aureus* i rezultati sklapanja su upoređeni sa referentnim genomima tog organizma kako bismo ispitali da li korišćenje integratora daje poboljšanje u odnosu na korišćenje pojedinačnih asemblera. Evaluacija rezultata izvedena je GAGE statistikom [28] i prikazana pomoću FRC krivih i pomoću alata Quast [29]. Rezultati na svim kombinacijama asemblera ($ABySS + Velvet$, $ABySS + SPAdes$, $SPAdes + Velvet$) prikazani u narednim tabelama 4.1, 4.2, 4.3, 4.4, 4.5 i 4.6, gde kolone predstavljaju podake iz alata Quast: N50 veličina, NG50 veličina, procenat genoma koji je rekonstuisan i druge metrike. Kolone u tabelama sa naslovom u oblika "GAM-NGS $X + Y$" predstavljaju rezultat integrisanja pomoću asemblera X i Y, gde je X glavno sklapanje a Y je sporedno sklapanje. $X, Y \in \{A, S, V\}$, gde su $A = ABySS$, $S = SPAdes$ i $V = Velvet$.

Prvi skup podataka se sastoji od uparenih očitavanja genoma organizma *Salmonella enterica*, dobijen Illumina MiSeq sekvenciranjem. Očitavanja su objavljena na zvaničnom sajtu NCBI-a [30] (*National Center for Biotechnology Information*) sa oznakom SRX4199189 [31] 12.06.2018-e godine. Referentna sekvenca *Salmonella enterice* je sa oznakom NC_003198.1 [32]. Quast tabele koje opisuju ovaj organizam su tabele 4.1, 4.2 i 4.3.

Sklapanje	ABySS	SPAdes	GAM-NGS A+S	GAM-NGS S+A
Kontige (≥ 10000 bp)	87	48	104	47
Kontige (≥ 25000 bp)	1	47	2	46
Kontige (≥ 50000 bp)	0	34	0	35
Ukupna dužina (≥ 10000 bp)	1212335	4552347	1483418	4507523
Ukupna dužina (≥ 25000 bp)	33957	4535325	67815	4490914
Ukupna dužina (≥ 50000 bp)	0	4082516	0	4114676
Broj kontiga	939	67	829	65
Ukupna dužina	4112479	4592117	4142874	4547069
Dužina reference	4809037	4809037	4809037	4809037
CG (%)	52.75	52.14	52.75	52.09
CG reference (%)	52.09	52.09	52.09	52.09
N50	6735	132266	7735	130812
NG50	5628	114825	6604	113871
Poravnata očitavanja (%)	96.6	99.54	96.89	99.47
Poravnata očitavanja uz referencu(%)	92.69	92.69	92.69	92.69
Srednja vrednost dubine pokrivenosti	80	77	80	77

TABELA 4.1: Sklapanje organizma *Salmonella enterica* kombinacijom ABySS + SPAdes

Sklapanje	Velvet	Spades	GAM-NGS S+V	GAM-NGS V+S
Kontige (≥ 10000 bp)	140	48	48	74
Kontige (≥ 25000 bp)	61	47	47	64
Kontige (≥ 50000 bp)	16	34	34	36
Ukupna dužina (≥ 10000 bp)	3974950	4552347	4552347	4461712
Ukupna dužina (≥ 25000 bp)	2704635	4535325	4535325	4306271
Ukupna dužina (≥ 50000 bp)	1166631	4082516	4082516	3280029
Broj kontiga	292	67	67	106
Ukupna dužina	4569472	4592117	4592117	4573612
Dužina reference	4809037	4809037	4809037	4809037
CG (%)	52.16	52.14	52.14	52.15
CG reference (%)	52.09	52.09	52.09	52.09
N50	29168	132266	132266	79716
NG50	28324	114825	114825	77409
Poravnata očitavanja (%)	99.07	99.54	99.54	99.27
Poravnata očitavanja uz referencu(%)	92.69	92.69	92.69	92.69
Srednja vrednost dubine pokrivenosti	77	77	77	77

TABELA 4.2: Sklapanje organizma *Salmonella enterica* kombinacijom SPAdes + Velvet

Sklapanje	ABySS	Velvet	GAM-NGS A+V	GAM-NGS V+A
Kontige (≥ 10000 bp)	87	140	97	104
Kontige (≥ 25000 bp)	1	61	2	2
Kontige (≥ 50000 bp)	0	16	0	0
Ukupna dužina (≥ 10000 bp)	1212335	3974950	1386850	1483418
Ukupna dužina (≥ 25000 bp)	33957	2704635	62207	67815
Ukupna dužina (≥ 50000 bp)	0	1166631	0	0
Broj kontiga	939	292	845	829
Ukupna dužina	4112479	4569472	4139910	4142874
Dužina reference	4809037	4809037	4809037	4809037
CG (%)	52.75	52.16	52.75	52.75
CG reference (%)	52.09	52.09	52.09	52.09
N50	6735	29168	7485	7735
NG50	5628	28324	6340	6604
Poravnata očitavanja (%)	96.6	99.07	96.84	96.89
Poravnata očitavanja uz referencu (%)	92.69	92.69	92.69	92.69
Srednja vrednost dubine pokrivenosti	80	77	80	80

TABELA 4.3: Sklapanje organizma *Salmonella enterica* kombinacijom ABySS + Velvet

Drugi skup podataka se sastoji od uparenih očitavanja genoma organizma *Staphylococcus aureus*-a, dobijen Illumina MiSeq sekvenciranjem. Očitavanja su objavljena na zvaničnom sajtu NCBI-a [30] sa oznakom ERX2074299 [33] 10.04.2018-e godine. Referentna sekvenca *Staphylococcus aureus*-a je sa oznakom NC_007795.1 [34]. Quast tabele koje opisuju ovaj organizam su tabele 4.4, 4.5 i 4.6.

Program je testiran na dve vrste bakterija. Testiranje programa na složenijim organizmima je bilo suviše vremenski i postorski zahtevno na kućnom računaru. (Konfiguracija računara: Memorija $3,8GiB$, Procesor Intel Core $i3-5005U$ CPU @$2.00GHz \times 4$, Disk $488,0GB$)

Sklapanje	Velvet	SPAdes	GAM-NGS S+V	GAM-NGS V+S
Kontige (≥ 10000 bp)	88	60	60	70
Kontige (≥ 25000 bp)	37	41	41	41
Kontige (≥ 50000 bp)	6	15	15	13
Ukupna dužina (≥ 10000 bp)	2235742	2672869	2672869	2458277
Ukupna dužina (≥ 25000 bp)	1409995	2329457	2329457	1955346
Ukupna dužina (≥ 50000 bp)	357946	1437945	1437945	992103
Broj kontiga	249	95	95	155
Ukupna dužina	2777345	2813453	2813453	2794097
Dužina reference	2821361	2821361	2821361	2821361
CG (%)	32.63	32.63	32.63	32.62
CG reference (%)	32.87	32.87	32.87	32.87
N50	25172	51979	51979	37136
NG50	24296	51979	51979	37136
Poravnata očitavanja (%)	98.92	99.8	99.8	99.46
Reference mapped (%)	87.89	87.89	87.89	87.89
Properly paired (%)	96.04	98.0	98.0	97.27
Poravnata očitavanja uz referencu (%)	87.33	87.33	87.33	87.33
Srednja vrednost dubine pokrivenosti	32	32	32	32

TABELA 4.4: Sklapanje organizma *Staphylococcus aureus* kombinacijom Velvet + SPAdes

Sklapanje	ABySS	SPAdes	GAM-NGS A+S	GAM-NGS S+A
Kontige (≥ 10000 bp)	79	60	18	59
Kontige (≥ 25000 bp)	39	41	1	41
Kontige (≥ 50000 bp)	9	15	0	16
Ukupna dužina (≥ 10000 bp)	2366425	2672869	232607	2680420
Ukupna dužina (≥ 25000 bp)	1717601	2329457	29418	2359090
Ukupna dužina (≥ 50000 bp)	657817	1437945	0	1494836
Broj kontiga	195	95	876	92
Ukupna dužina	2781159	2813453	2587943	2814077
Dužina reference	2821361	2821361	2821361	2821361
CG (%)	32.64	32.63	32.52	32.63
CG reference (%)	32.87	32.87	32.87	32.87
N50	31253	51979	4072	52877
NG50	30755	51979	3719	52877
Poravnata očitavanja (%)	97.48	99.8	91.32	99.8
Reference mapped (%)	87.89	87.89	87.89	87.89
Properly paired (%)	95.25	98.0	87.55	98.07
Poravnata očitavanja uz referencu (%)	87.33	87.33	87.33	87.33
Srednja vrednost dubine pokrivenosti	31	32	31	32

TABELA 4.5: Sklapanje organizma *Staphylococcus aureus* kombinacijom SPAdes + ABySS

4.4 Zaključak

Problem preciznog sklapanja genoma je jedan od najznačajnih u savremenoj genomici. Razvijen je veliki broj programa za sklapanje zasnovanih na De Brojnovom grafu koji posebno uspešno rešavaju neke specifične elemente ovog zadatka.

GAM-NGS je nedavno razvijen integrator koji se koristi preko komandne linije čime se relativno ograničava njegova šira primena. Cilj ovog rada je implementacija programa koji bi za učesljavanjenje koristio GAM-NGS koji je zasnovan na kombinacijama poznatih asemblera ABySS, Velvet i SPAdes

Sklapanje	ABySS	Velvet	GAM-NGS A+V	GAM-NGS V+A
Kontige ($\geq$ 10000 bp)	79	88	12	90
Kontige ($\geq$ 25000 bp)	39	37	0	37
Kontige ($\geq$ 50000 bp)	9	6	0	6
Ukupna dužina ($\geq$ 10000 bp)	2366425	2235742	146310	2266338
Ukupna dužina ($\geq$ 25000 bp)	1717601	1409995	0	1410691
Ukupna dužina ($\geq$ 50000 bp)	657817	357946	0	358713
Broj kontiga	195	249	925	244
Ukupna dužina	2781159	2777345	2577699	2781334
Dužina reference	2821361	2821361	2821361	2821361
CG (%)	32.64	32.63	32.53	32.63
CG reference (%)	32.87	32.87	32.87	32.87
N50	31253	25172	3866	25172
NG50	30755	24296	3568	25172
Poravnata očitavanja (%)	97.48	98.92	90.93	99.05
Reference mapped (%)	87.89	87.89	87.89	87.89
Properly paired (%)	95.25	96.04	86.72	96.24
Poravnata očitavanja uz referencu (%)	87.33	87.33	87.33	87.33
Srednja vrednost dubine pokrivenosti	31	32	31	32

TABELA 4.6: Sklapanje organizma *Staphylococcus aureus* kombinacijom ABySS + Velvet

kao i razvoj grafičkog korisničkog interfejsa koji bi omogućio ugodno i jednostavno korišćenje od strane korisnika bez značajne informatičke ekspertize.

Program je testiran na sirovim podacima sekvenciranja organizama *Salmonella enterica* i *Staphylococcus aureus* i rezultati sklapanja su upoređeni sa referentnim genomima tog organizma. Na osnovu rezultata N50 i NG50 u gorenavedenim tabelama vidi se da je značajno poboljšanje postignuto u kombinaciji ABySS i SPAdes u odnosu na sklapanje pojedinačnim asemblerima.

Program razvijen u okviru ovog istraživanja predstavlja značajan doprinos rešavanju problema sklapanja genoma i dostupan je na zahtev korisnika.

Literatura

[1] Students of University of Belgrade - Faculty of Mathematics. *Uvod u bioinformatiku - beleške sa predavanja*. 2018, pp. 47–68.

[2] Riccardo Vicedomini et al. "GAM-NGS: genomic assemblies merger for next generation sequencing". In: *BMC Bioinformatics* 14.7 (Apr. 2013), pp. 1–18. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3633056/.

[3] Alberto Casagrande et al. "GAM: Genomic Assemblies Merger: A Graph Based Method to Integrate Different Assemblies". In: *2009 IEEE International Conference on Bioinformatics and Biomedicine* (Nov. 2009), pp. 321–326. URL: https://ieeexplore.ieee.org/document/5341771.

[4] Turner and Frances. "Assessment of Insert Sizes and Adapter Content in Fastq Data from NexteraXT Libraries". In: *Frontiers in Genetics* 5.5 (Jan. 2014), p. 5. URL: https://www.researchgate.net/publication/260170597_Assessment_of_Insert_Sizes_and_Adapter_Content_in_Fastq_Data_from_NexteraXT_Libraries.

[5] Robert Ekblom and Jochen B W Wolf. "A field guide to whole-genome sequencing, assembly and annotation". In: *Evolutionary Applications* 7 (Nov. 2014), pp. 1026–1042. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4231593/.

[6] Jared T. Simpson et al. "ABySS: A parallel assembler for short read sequence data". In: *Genome Research* 19 (June 2009), pp. 1117–1123. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2694472/#.

[7] Nurk Sergey et al. "Assembling Single-Cell Genomes and Mini-Metagenomes From Chimeric MDA Products". In: *Research in Computational Molecular Biology* (2013), pp. 158–170. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3791033/.

[8] "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing". In: *Journal of computational biology* 19.5 (May 2012), pp. 455–477. URL: https://www.ncbi.nlm.nih.gov/pubmed/22506599.

[9] Daniel R. Zerbino and Ewan Birney. "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs". In: *Genome Research* 18.5 (May 2008), pp. 811–829. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2336801/.

[10] Zhenyu Li et al. "Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph". In: *Briefings in Functional Genomics* 11.1 (Jan. 2012), pp. 25–37. URL: https://academic.oup.com/bfg/article/11/1/25/191455.

[11] Phillip Compeau and Pavel Pevzner. *Bioinformatics Algorithms: An Active Learning Approach*. Vol. 1. 2. Active Learning Publishers, 2015, pp. 115–180. URL: http://bioinformaticsalgorithms.com/.

[12] *BLAST - Basic Local Alignment Search Tool*. 2019. URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi (visited on 06/05/2019).

[13] Zimin AV et al. "Assembly reconciliation". In: *Bioinformatics*. 24.1 (Jan. 2008), pp. 42–45. URL: https://www.ncbi.nlm.nih.gov/pubmed/18057021?dopt=Abstract.

[14] Guohui Yao et al. "Graph accordance of next-generation sequence assemblies". In: *Bioinformatics* 28.1 (Jan. 2012), pp. 13–16. URL: https://academic.oup.com/bioinformatics/article/28/1/13/218208.

[15] Li H et al. "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–2079. URL: https://www.ncbi.nlm.nih.gov/pubmed/19505943.

[16] Robert Tarjan. "Depth-First Search and Linear Graph Algorithms". In: *SIAM J. Comput* 1.2 (1972), pp. 146–160. URL: https://epubs.siam.org/doi/abs/10.1137/0201010.

[17] Yong Wang and Nicholas E. Navin. "Advances and Applications of Single Cell Sequencing Technologies". In: *Moll Cell.* 58.4 (May 2015), pp. 598–609. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC444195.

[18] James Eberwine et al. In: *Nature Methods* 11 (2014), pp. 25–27. URL: https://www.nature.com/articles/nmeth.2769.

[19] Chitsaz H et al. "Efficient de novo assembly of single-cell bacterial genomes from short-read data sets." In: *Nat Biotechnol.* 29.10 (Sept. 2011), pp. 915–921. URL: https://www.ncbi.nlm.nih.gov/pubmed/21926975.

[20] Peng Y. et al. "IDBA—a practical iterative de Bruijn graph de novo assembler." In: *Berger B. - Research in Computational Molecular Biology.* 6044 (2010). URL: https://link.springer.com/chapter/10.1007/978-3-642-12683-3_28.

[21] Salzberg SL et al. "GAGE: Acritical evaluation of genome assemblies and assembly algorithms." In: *Genome Research* (2012), pp. 557–567. URL: http://genome.cshlp.org/cgi/doi/10.1101/gr.131383.111.

[22] *Why is N50 used as an assembly metric:* 2019. URL: http://www.acgt.me/blog/2013/7/8/why-is-n50-used-as-an-assembly-metric.html (visited on 01/30/2019).

[23] *FRCurve*. 2019. URL: http://amos.sourceforge.net/wiki/index.php/FRCurve (visited on 01/30/2019).

[24] *Receiver operating characteristic*. 2019. URL: https://en.wikipedia.org/wiki/Receiver_operating_characteristic (visited on 02/12/2019).

[25] Francesco Vezzi, Giuseppe Narzisi, and Bud Mishra. "Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathons". In: *Plos one* 7.12 (Dec. 2012), pp. 1–11. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0052210.

[26] *FRC - A tool able to evaluate and rank de novo assemblies and assemblers. Feature Response Curve*. 2019. URL: https://opensource.scilifelab.se/projects/frc/ (visited on 01/30/2019).

[27] Alexey Gurevich et al. "QUAST: quality assessment tool for genome assemblies". In: *Bioinformatics* 29.8 (Apr. 2013), pp. 1072–1075. URL: https://www.ncbi.nlm.nih.gov/pubmed/23422339.

[28] *GAGE: generally applicable gene set enrichment for pathway analysis*. 2019. URL: http://gage.cbcb.umd.edu (visited on 01/30/2019).

[29] *QUAST - Quality Assessment Tool for Genome Assemblies*. 2019. URL: http://quast.sourceforge.net/quast (visited on 06/05/2019).

[30] *The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.* 2019. URL: https://www.ncbi.nlm.nih.gov/ (visited on 03/10/2019).

[31] *Whole genome Illumina MiSeq sequence of Salmonella enterica*. 2019. URL: https://www.ncbi.nlm.nih.gov/sra/?term=SRR7296632 (visited on 03/10/2019).

[32] *Salmonella enterica subsp. enterica serovar Typhi str. CT18, complete genome*. 2019. URL: https://www.ncbi.nlm.nih.gov/nuccore/NC_003198.1 (visited on 05/10/2019).

[33] *Illumina MiSeq paired end sequencing: Staphylococcus aureus*. 2019. URL: https://www.ncbi.nlm.nih.gov/sra/ERX2074299[accn] (visited on 03/10/2019).

[34] *Staphylococcus aureus subsp. aureus NCTC 8325 chromosome, complete genome*. 2019. URL: https://www.ncbi.nlm.nih.gov/nuccore/NC_007795.1 (visited on 05/10/2019).