Univerzitet u Beogradu
Matematički fakultet

Master rad

Asembliranje genoma na osnovu skupa kontiga

Student:
Darko Živanović

Mentor:
dr Nenad Mitić

Beograd, Oktobar 2011.

Mentor:

dr Nenad Mitić
Matematički fakultet Univerziteta u Beogradu

Članovi komisije:

dr Saša Malkov
Matematički fakultet Univerziteta u Beogradu

dr Miloš Beljanski
Institut za opštu i fizičku hemiju u Beogradu

Datum odbrane:

_____

Sadržaj

Spisak slika

Spisak tabela

Predgovor

Nasledna osnova svakog živog bića određena je njegovim genomom. On sadrži kompletan skup naslednih informacija nekog organizma. Kod prokariota genom se sastoji od jednog ili nekoliko kružnih molekula DNK, koji se često označavaju kao hromozomska DNK. Pored hromozomske DNK, prokarioti često nose jedan ili više manjih linearnih ili kružnih molekula DNK, označenih kao plazmidi. Oni su prisutni u većem broju kopija i nose gene koji u datim uslovima utiču na preživljavanje.

U okviru Instituta za molekularnu genetiku i genetičko inženjerstvo Univerziteta u Beogradu uspešno je sekvenciran celokupan genom bakterije *Lactobacillus paracasei subsp. paracasei BGSJ2-8* izolovane iz tradicionalnog domaćeg srpskog sira. U ovom radu je prikazan algoritam za asembliranje genoma navedene bakterije na osnovu skupa kontiga dobijenih procesom sekvenciranja. Algoritam je zasnovan na sličnosti bakterija iz roda *Lactobacillus*. Spajanje pojedinačnih kontiga je vršeno različitim tehnikama i alatima koji su kao ulaz koristili podatke dobijene blastovanjem celokupnog skupa kontiga. Izloženi su rezultati koji predstavljaju potencijalne superkontige na osnovu kojih eksperimentalnim metodama u potpunosti može da se sklopi genom pomenute bakterije.

Svi materijali i dobijeni rezultati se nalaze na kompakt disku koji je sastavni deo ovog rada.

Zahvalnica

Želim da se zahvalim brojnim ljudima koji su mi pomogli i omogućili da završim ovaj master rad. Pre svega se zahvaljujem svom mentoru, dr Nenadu Mitiću, profesoru Matematičkog fakulteta Univerziteta u Beogradu, bez čije svesrdne pomoći i vremena koje je nesebično delio sa mnom ovaj rad ne bi izgledao ovako. Zahvalnost dugujem i dr Milošu Beljanskom, višem naučnom saradniku Instituta za opštu i fizičku hemiju u Beogradu, čije su mi kritičke primedbe pomogle da sistematičnije pristupim ovoj materiji. Želim da izrazim i duboku zahvalnost dr Nataši Golić, dr Jeleni Begović i dr Branku Jovčiću, saradnicima sa Instituta za molekularnu genetiku i genetičko inženjerstvo Univerziteta u Beogradu na nizu korisnih saveta, stručnih smernica i iskrenih ohrabrenja tokom pisanja ovog rada.

Glava 1

Uvod

1.1 Šta je bioinformatika?

Tokom poslednjih nekoliko godina, veliki napredak u nauci i tehnologiji je omogućio uporednu analizu kompletnih genoma. Istraživanja koja zahtevaju analizu, skladištenje i obradu velikih količina podataka nisu bila moguća bez upotrebe računara, što je dovelo do stvaranja nove naučne discipline koju nazivamo **bioinformatika**. U najširem smislu, bioinformatika predstavlja svaku primenu računarskih nauka u biologiji. To je oblast koja spaja ispitivanje strukture i funkcije bioloških informacija sa teorijskim i praktičnim znanjima matematike i računarstva. Bioinformatika se definiše i kao primena informacionih tehnologija u analizi i obradi bioloških podataka.

Jedan od ciljeva bioinformatike je prevođenje velikog broja podataka o sekvenci u biološki značajne informacije, kao i razumevanje strukturnih, funkcionalnih i evolucionih podataka kodiranih u biološke sekvence. Problemi koje bioinformatika rešava uglavnom potiču iz genetike, ali ima problema i iz medicine i farmakologije (npr. predviđanje mutacije virusa u cilju pronalaženja lekova za određene bolesti), kao i iz drugih oblasti.

1.2 Osnovni biološki pojmovi

U ovom poglavlju su definisani osnovni biološki pojmovi neophodni za razumevanje problema koji se razmatra u radu. Od velikog broja molekula koji su zastupljeni u živim sistemima, za potrebe ovog rada su od posebnog značaja **nukleinske kiseline**, biološki makromolekuli koji predstavljaju nosioce naslednih informacija.

1.2.1 Hemijski sastav i građa nukleinskih kiselina

Nukelinske kiseline je prvi put izolovao **Fridrih Mišer** 1871. godine[13]. Naziv su dobile prema jedru (lat. *nucleus*) u kome su najviše zastupljene, mada ih ima i u citoplazmi. Njihova uloga u prenošenju naslednih informacija otkrivena je 1928. godine. U prirodi postoje dve vrste nukleinskih kiselina:

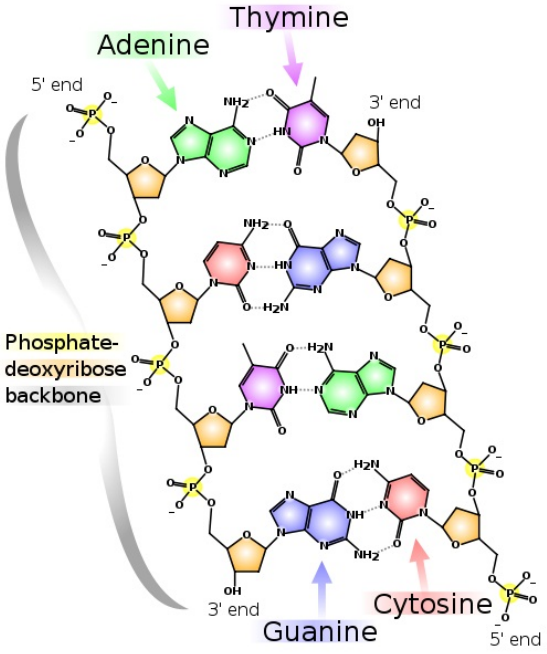
- dezoksiribonukleinska kiselina (**DNK**) i
- ribonukleinska kiselina (**RNK**).

Ovi makromolekuli su zastupljeni u svim vrstama organizama i veoma su značajni za održavanje života i evoluciju živog sveta.

Primarna struktura

Nukleinske kiseline su linearni polimerni makromolekuli čija je osnovna monomerna jedinica građe **nukleotid**. Svaki nukleotid se sastoji od jednog molekula heterociklične organske **azotne baze**, jednog molekula šećera **pentoze** u obliku furanoze, i jednog molekula **fosforne kiseline**. Nukleinske kiseline sadrže dva tipa azotnih baza: **purinske** i **pirimidinske**. Purinske baze (purini) koje ulaze u sastav nukleinskih kiselina su **adenin** (A) i **guanin** (G), a pirimidinske (pirimidini) **timin** (T), **citozin** (C) i **uracil** (U). U sastavu DNK mogu da se nađu adenin, guanin, timin i citozin, a u sastavu RNK adenin, guanin, uracil i citozin. Pentoza koja ulazi u sastav DNK je **2'-dezoksiriboza**, dok se u RNK nalazi **riboza**.

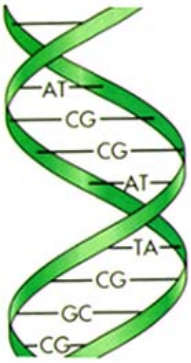
Nukleotidi su međusobno povezani fosfodiestarskim vezama između $C_{3'}$ pentoze jednog nukleotida i $C_{5'}$ pentoze narednog nukleotida u nizu. Okosnicu (kičmu) molekula nukleinskih kiselina čine pentoze i fosfatne grupe. Na jednom kraju molekula ostaje slobodna $C_{3'}$ hidroksilna grupa i taj kraj se naziva **3' kraj**, a na drugom $C_{5'}$ fosfatna grupa i taj kraj se naziva **5' kraj** molekula. Vrsta i redosled nukleotida DNK predstavljaju njenu **primarnu strukturu** i označeni su početnim slovima njihovih imena.



Slika 1.1: Hemijska struktura DNK-a

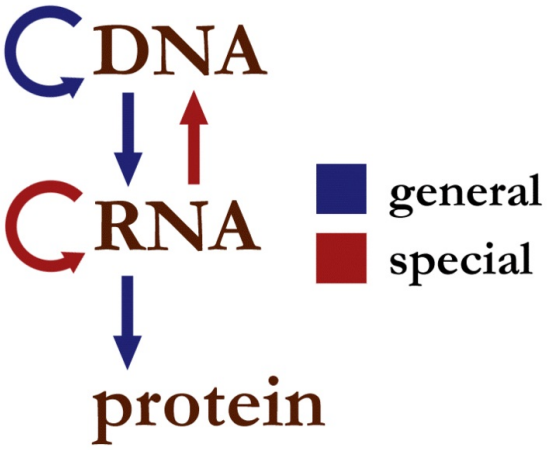
Sekundarna struktura

Sekundarnu strukturu DNK uspeli su da odgonetnu **Džejms Votson** i **Frensis Krik** 1953. godine. Osnovu te strukture čini dvolančana zavojnica (slika 1.2). Dva polinukleotidna lanca koja čine ovu zavojnicu su antiparalelna, što znači da se naspram 5' kraja jednog lanca nalazi 3' kraj drugog, i obrnuto. Adenin jednog lanca je uvek u paru sa timinom naspramnog lanca, dok je guanin uvek u paru sa citozinom. Važi i obrnuto. Fosfatne grupe su okrenute prema spoljašnjoj strani i zajedno sa pentozama čine skelet zavojnice.



Slika 1.2: Sekundarna struktura DNK-a

Nedugo nakon otkrića sekundarne strukture DNK postavljena je centralna dogma (slika 1.3) po kojoj se informacije sa DNK u procesu transkripcije (prepisivanja) prenose na RNK. Dalje se u procesu biosinteze proteina, tj. translacije (prevođenja), redosled nukleotida prevodi u redosled aminokiselina tako što svaka tri nukleotida (tripleti) kodiraju jednu aminokiselinu. Najkraće rečeno, DNK upravlja sintezom RNK, a RNK sintezom proteina.



Slika 1.3: Centralna dogma molekularne biologije

Biološke informacije se čuvaju u DNK niski, čiji su pojedini delovi (podniske) označeni kao geni koji nose šifre za proteine.

1.2.2 Gen i genom

Gen kao faktor koji ne menja svoje karakteristike pri prenošenju kroz generacije definisao je **Gregor Mendel** 1865. godine. Pojam gena je 1909. godine uveo danski botaničar **Vilhelm Johansen**[14]. Sa aspekta molekularne biologije, gen je segment molekula DNK koji sadrži instrukcije za sintezu proteina ili RNK. Gen može da se posmatra i kao redosled nukleotida u molekulu DNK koji određuje hemijsku strukturu specifičnog polipeptida ili molekula RNK.

Nasledna osnova svakog živog bića definisana je njegovim **genomom**. On se sastoji od duge sekvence nukleinskih kiselina koja obezbeđuje informaciju neophodnu da se organizam izgradi i funkcioniše. Termin **informacija** se koristi jer genom sam po sebi nema aktivnu ulogu u izgradnji i funkcionisanju organizma, već redosled nukleotida u molekulima DNK određuje nasledne osobine. Redosled nukleotida u molekulima DNK se koristi za stvaranje RNK i proteina na odgovarajućem mestu i u odgovarajuće vreme. Nastali proteini učestvuju u izgradnji organizma ili u metaboličkim reakcijama neophodnim za život.

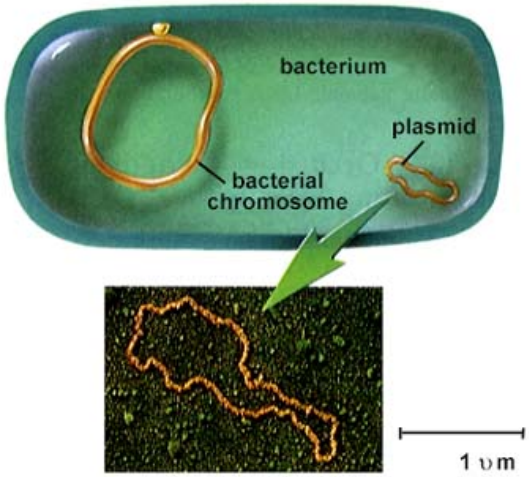
Genom sadrži kompletan skup naslednih informacija nekog organizma[1]. Genom se, fizički posmatrano, sastoji od jednog ili više molekula DNK. Funkcionalno posmatrano, genom se sastoji iz gena koji nose zapis za različite proteine i molekule RNK. Broj gena u genomima različitih organizama se veoma razlikuje. Tako, genom bakterije mikroplazme sadrži manje od 500 gena, dok ljudski genom sadrži između 30.000 i 40.000 gena. Krajnja definicija genoma podrazumeva određivanje redosleda nukleotida u svim molekulima DNK iz kojih se sastoji. Genomi bakterija sadrže i do nekoliko miliona nukleotida.

Fizička organizacija genoma

Prema složenosti građe ćelije svi ćelijski organizmi se dele na prokariote i eukariote. Prokariotama pripadaju bakterije i arhee, dok su eukariote svi ostali jednoćelijski i višećelijski organizmi. Prokariotske ćelije su male i jednostavne građe, opkoljene ćelijskim zidom i membranom, a nemaju jedro niti ćelijske organele, osim ribozoma.

Fizička organizacija genoma prokariota i eukariota se bitno razlikuje. Kod najvećeg broja prokariota genom je predstavljen sa jednim ili nekoliko molekula DNK, koji su najčešće kružni, ali mogu biti i linearni. Takvi molekuli DNK se često označavaju kao **hromozomska DNK**. Pored hromozomske DNK, prokarioti često nose jedan ili više malih kružnih ili linearnih molekula DNK, označenih kao **plazmidi**. Plazmidi se javljaju u većem broju kopija i nose gene koji u određenim uslovima mogu da budu korisni za preživljavanje. S obzirom da nisu esencijalni za rast prokariota, njihov genom se obično definiše samo hromozomskom DNK,

dok se na plazmide gleda kao na pomoćne komponente genoma. Red veličine plazmida je do stotinu hiljada nukleotida. U ovom radu su izdvojene i kontige koje ulaze u sastav plazmida bakterije *Lactobacillus paracasei subsp. paracasei BGSJ2-8*.



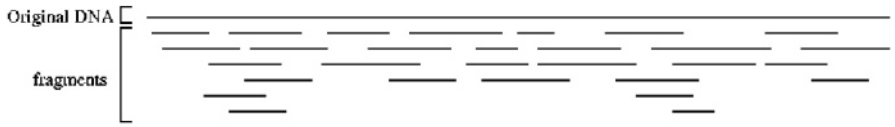
Slika 1.4: Fizička organizacija genoma prokariota

1.3 Sekvenciranje genoma

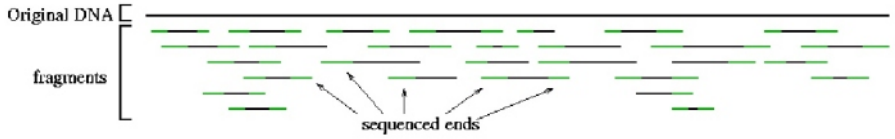
Sekvenciranje genoma predstavlja više različitih metoda za utvrđivanje redosleda nukleotida u molekulu DNK. 1977. godine je kompletno sekvenciran prvi genom, genom bakteriofaga $\phi X174$. Njegova veličina je 5.368 nukleotida. Automatizovano DNK sekvenciranje je omogućilo analitičke i uporedne studije genoma, dozvoljavajući naučnicima njihovo celokupno dešifrovanje. Iako veličina genoma varira od nekoliko miliona nukleotida kod bakterija, pa do nekoliko milijardi nukleotida kod ljudi i većine biljaka i životinja, hemijske reakcije koje naučnici koriste za dešifrovanje DNK baznih parova precizne su za samo oko 600 do 700 nukleotida u isto vreme[2]. Ovo predstavlja značajno ograničenje s obzirom da čak i najjednostavniji virusi sadrže desetine hiljada baznih parova, bakterije milione, a genomi sisara milijarde baznih parova.

Proces sekvenciranja započinje fizičkom podelom DNK na milione slučajnih delova, čiji se krajevi zatim **čitaju** DNK mašinom za sekvenciranje. Nakon toga, računarski program (poznatiji kao **asembler**) spaja preklapajuća čitanja i rekonstruiše originalnu sekvencu. Ovu opštu tehniku, koja je nazvana ***shotgun sequencing***, prvi je definisao **Fred Sanger** 1982. godine.

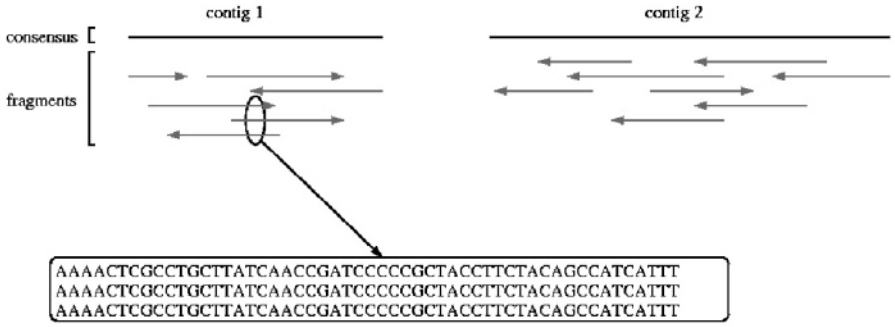
Kao i delovi slagalice, DNK čitanja koje *shotgun* sekvenciranje proizvodi moraju da se sklope u celokupan genom. Međutim, ovaj naizgled jednostavan proces sadrži niz tehničkih izazova. Tu se pre svega misli na greške u podacima,



Slika 1.5: Podela originalne DNK na skup fragmenata



Slika 1.6: Sekvenciranje krajeva fragmenata



Slika 1.7: Asembliranje preklapajućih čitanja na osnovu sličnosti sekvenci

pri čemu su neke od njih nastale zbog ograničenja tehnologije sekvenciranja, a neke ljudskom greškom tokom laboratorijskog rada. Čak i kada ne bi bilo grešaka, DNK sekvence imaju neke karakteristike poput ponavljajućih delova koji komplikuju proces asembliranja. Ljudski genom, na primer, sadrži ponavljanja koja se javljaju u više od 100.000 kopija. Čitanja koja pripadaju ponavljajućim sekvencama se teško pozicioniraju na korektan način. Sve to dovodi do velikih poteškoća prilikom asembliranja pojedinih genoma i rezultuje prazninama u njihovoj pokrivenosti.

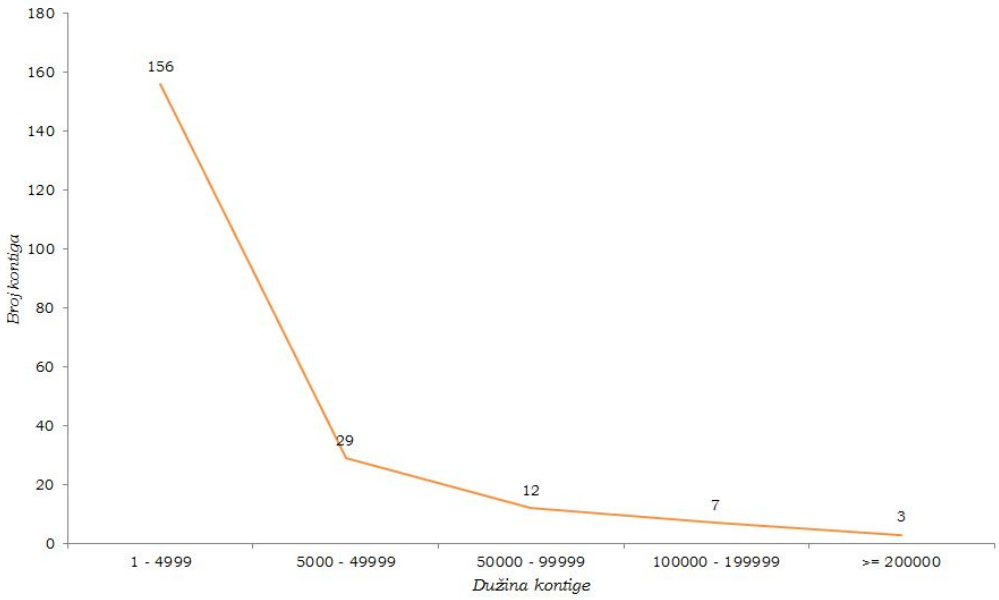
Rešavanje ovih problema zahteva dodatnu završnu fazu u celokupnom procesu, koja uključuje brojne intervencije od strane ljudi, ali je taj završetak jako skup i zahteva specijalizovane laboratorijske tehnike i visoko obučeno osoblje. Programi za asembliranje mogu drastično da smanje troškove koristeći dodatne informacije dobijene u završnoj fazi, ali većina današnjih asemblera ne uzima u obzir ove informacije i generiše genom na osnovu početnih *shotgun* čitanja.

1.3.1 Asembliranje genoma

U zavisnosti od tehnologije koju koriste, postojeće mašine za sekvenciranje odjednom mogu da proizvedu samo kratke sekvence na krajevima čitanja čije su dužine između 50 i 500 nukleotida. Da bi se sklopio celokupan genom, kratke sekvence se spajaju u duže nakon uklanjanja preklapajućih delova. Ovi duži, spojeni fragmenti se nazivaju **kontige**[1] i obično su dužina između 5.000 i 10.000 baznih parova[3]. Jedan deo preklapajućih kontiga se dalje spaja u **superkontige**, koje se na kraju povezuju i grade mapu celokupnog genoma. Istraživanja različitih algoritama za asembliranje su od vitalnog značaja za bioinformatiku.

1.3.2 Sekvenciranje genoma *Lactobacillus paracasei subsp. paracasei BGSJ2-8*

Institut za molekularnu genetiku i genetičko inženjerstvo Univerziteta u Beogradu uspešno je sekvencirao genom *Lactobacillus paracasei subsp. paracasei BGSJ2-8*[2] koji je izolovan iz tradicionalnog domaćeg srpskog sira. U tu svrhu je korišćen metod skeniranja oba kraja (eng. *pair-end sequencing*) sa dužinom sekvenciranih krajeva čitanja od 76 nukleotida. Na taj način je skenirano 26.576.780 parova krajeva čitanja. Prosečna dužina dobijenih kontiga je 14.994 nukleotida.



Slika 1.8: Broj kontiga prema njihovoj veličini

U nastavku je prikazan algoritam za anotaciju genoma pomenute bakterije na

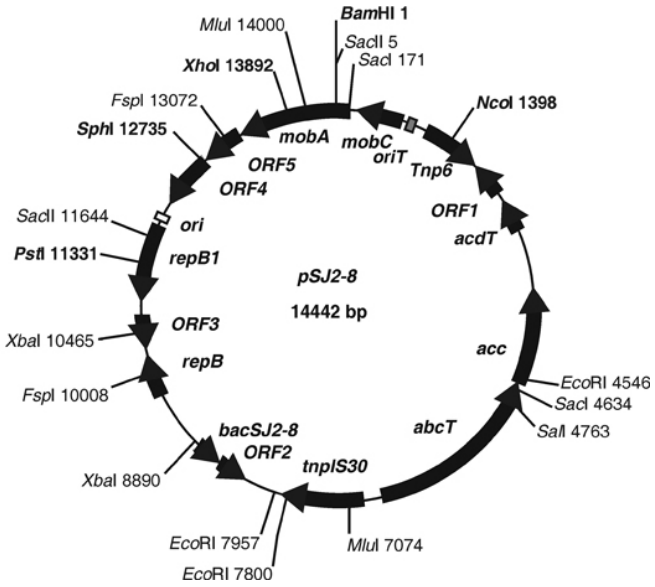
[1]Kontiga predstavlja neprekidni skup preklapajućih DNK sekvenci.

[2]*Lactobacillus paracasei subsp. paracasei BGSJ2-8* pripada diviziji *Firmicutes*, porodici *Lactobacillaceae*, rodu *Lactobacillus* i vrsti *Lactobacillus paracasei*.

osnovu skupa od 207 kontiga dobijenih procesom sekvenciranja. Algoritam je zasnovan na sličnosti bakterije *Lactobacillus paracasei subsp. paracasei BGSJ2-8* sa drugim bakterijama iz roda *Lactobacillus*. Izloženi su i rezultati koji predstavljaju potencijalne superkontige, a na osnovu kojih eksperimentalnim metodama može da se u potpunosti sklopi genom pomenute bakterije.

Slika 1.9: *Lactobacillus paracasei subsp. paracasei*

Nukleotidna sekvenca plazmida *pSJ2-8* izolovanog iz soja *Lactobacillus paracasei subsp. paracasei BGSJ2-8* može da se vidi u ***EMBL GenBank*** pod brojem ***FM246455***.



Slika 1.10: Restrikciona mapa plazmida *pSJ2-8* izolovanog iz soja *Lactobacillus paracasei subsp. paracasei BGSJ2-8*

Glava 2

Postojeće metode za asembliranje genoma

Programi za asembliranje genoma kao ulaz koriste redosled i prirodu (A, T, G, C) svakog nukleotida u čitanju. Ovi programi pridružuju meru kvaliteta svakom nukleotidu na svakoj poziciji svakog čitanja. Oni uključuju i sistem za upravljanje podacima, kao i alate za vizuelni prikaz i interaktivno uređivanje. Zbog njihove složenosti, većina softverskih paketa za asembliranje genoma nije dostupna na *web* serverima, već je neophodna instalacija na lokalnim (dosta jakim) mašinama.

Najčešće korišćeni softverski paketi predstavljeni su sledećim dvema tabelama.

Staden Package	http://staden.sourceforge.net/
CAP3 Sequence Assembly Program	http://seq.cs.iastate.edu/
Phred, Phrap, Consed	http://www.phrap.org/
TIGR Assembler	http://www.tigr.org/software/

Tabela 2.1: Slobodni softverski paketi.

Sequencher	http://www.genecodes.com/
CLC Main Workbench	http://www.clcbio.com/

Tabela 2.2: Komercijalni softverski paketi.

Za asembliranje genoma je potreban program koji je u stanju da prepozna značajna preklapanja između delova i izvrši njihovo asembliranje u pojedinačne sekvence, tj. kontige. U sledećem odeljku su detaljnije opisani programi za asembliranje genoma.

2.1 Staden Package

Staden Package je skup alata otvorenog koda za asembliranje i analizu DNK sekvenci. Razvijan je u Laboratoriji za molekularnu biologiju Univerziteta u

Kembridžu od 1977. godine. Program je bio dostupan bez naknade studentima širom sveta do 2003. godine, kada su novčana sredstva za dalji razvoj ukinuta. Od tada je *Staden Package* otvorenog koda.

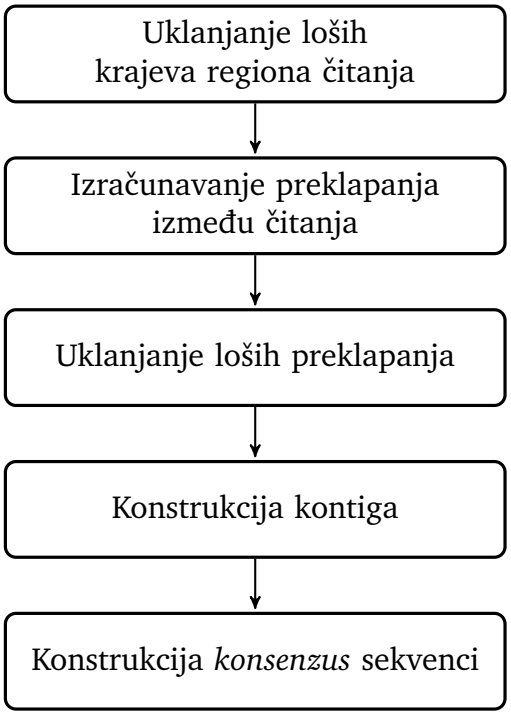
Staden Package se sastoji od velikog broja različitih alata. Glavni delovi su:

- **pregap4** - obezbeđuje grafički korisnički interfejs u kome se vrši priprema podataka za analizu ili asembliranje
- **trev** - brz i fleksibilan pregledač i uređivač *ABI*, *ALF*, *SCF* i *ZTR trace* datoteka
- **gap4** - obavlja asembliranje sekvenci, pozicioniranje i spajanje kontiga, proveru korektnosti spajanja, pretragu ponavljajućih regiona...
- **spin** - vrši analizu nukleotida sekvence u cilju pronalaženja gena i motiva. Može da obavlja translacije, pronalazi otvorene okvire čitanja, kodone itd. Većina rezultata je prikazana grafički. Ovaj program vrši i poređenje parova sekvenci.

Program je dostupan na adresi: http://staden.sourceforge.net/.

2.2 CAP3

CAP3 je jedan od najčešće korišćenih programa za asembliranje genoma zbog svoje jednostavnosti i efikasnih algoritama za pronalaženje značajnih preklapanja između fragmenata. Dostupan je za akademsku upotrebu bez naknade. Ulaz u program je skup sekvenci u *FASTA* formatu. Svaka pojedinačna sekvenca mora da ima svoje zaglavlje. Osnovni koraci algoritma za asembliranje koji je implementiran u *CAP3* su prikazani na narednom dijagramu.



Program je dostupan na adresi: http://pbil.univ-lyon1.fr/cap3.php/.

2.3 Phrap

Phrap je UNIX program za asembliranje DNK sekvenci. Deo je **Phred-Phrap-Consed** paketa. Razvio ga je **Fil Grin** u cilju asembliranja hibridnih plazmida dobijenih sekvenciranjem ljudskog genoma. Program se širom sveta koristi za različite projekte asembliranja sekvenci, između ostalog i bakterijskih genoma.

Kao ulaz *Phred* uzima *base-call* datoteke i poravnava pojedinačne fragmente koristeći **Smit-Votermanov** algoritam. Prilikom poravnavanja delova koriste se informacije o kvalitetu baza. Nakon pronalaženja sličnosti između sekvenci, program vrši asembliranje otklanjajući preklapajuće regione. Izlaz iz programa su kontige dobijene spajanjem svih preklapajućih čitanja.

Program je dostupan na adresi: http://www.phrap.org/.

2.4 TIGR Assembler

TIGR Assembler predstavlja nešto drugačijii pristup asembliranju velikih projekata *shotgun* sekvenciranja. Program rešava nekoliko glavnih problema u asembliranju takvih projekata:

- veliki proj poređenja baznih parova
- prisustvo ponavljajućih regiona i
- greške u sekvenciranju.

Početno poređenje fragmenata zasnovano na *oligonucleotide content* eliminiše potrebu za detaljnijim poređenjem većine baznih parova fragmenata, što značajno smanjuje računarsko vreme pretrage. Potencijalni ponavljajući regioni se prepoznaju utvrđivanjem mogućih preklapanja fragmenata. Problem ponavljajućih regiona se rešava postavljanjem jačih kriterijuma i njihovim asembliranjem na kraju procesa, kako bi se maksimalno iskoristile informacije iz neponavljajućih regiona.

Program je dostupan na adresi: http://www.jcvi.org/cms/research/software/.

2.5 Nedostaci postojećih metoda

Proces asembliranja genoma sadrži niz tehničkih izazova. Glavni problemi koji se javljaju su:

- greške u podacima
- ponavljajuće sekvence i
- nedostatak računarskih resursa.

Čak i kada ne bi bilo grešaka nastalih zbog ograničenja tehnologije sekvenciranja ili ljudskih propusta tokom laboratorijskog rada, DNK sekvence imaju neke karakteristike poput ponavljajućih delova koje komplikuju proces asembliranja. Takve sekvence se teško pozicioniraju na korektan način. Rešavanje ovih problema zahteva dodatnu završnu fazu koja je dosta skupa i zahteva ljudsku intervenciju, specijalizovane laboratorijske tehnike i visoko obučeno osoblje. Većina današnjih programa za asembliranje sklapa genom na osnovu početnih *shotgun* čitanja, čime se ne rešavaju navedeni problemi.

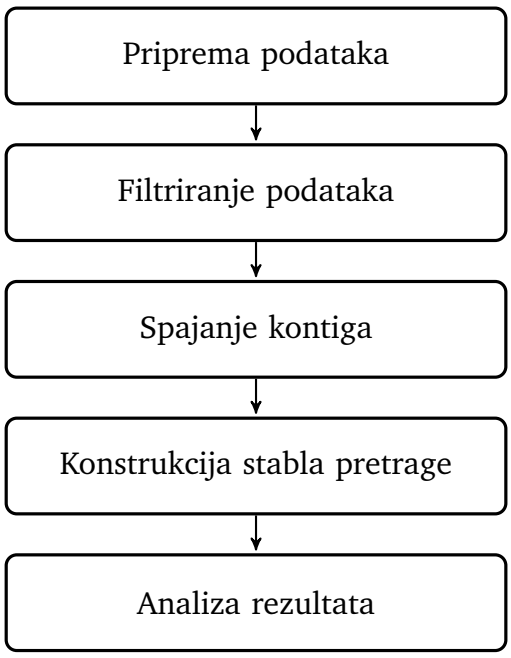
Najveći deo testiranih postojećih softverskih rešenja nije mogao da obradi celokupan skup kontiga na dostupnim računarima zbog nerealnih zahteva za računarskim resursima. *Staden Package* je jedini program koji je dao izlaz, ali je pokazano da dobijeni rezultati nisu dobri i da ne mogu da budu iskorišćeni. To je dovelo do potrebe za razvojem algoritma asembliranja zasnovanog na sličnosti bakterija iz roda *Lactobacillus*. Algoritam je razvijan na osnovu instrukcija dobijenih od saradnika sa Instituta za molekularnu genetiku i genetičko inženjerstvo Univerziteta u Beogradu.

Glava 3

Algoritam asembliranja zasnovan na sličnosti bakterija iz roda *Lactobacillus*

Kako gotova softverska rešenja nisu dala zadovoljavajuće rezultate, za potrebe ovog rada konstruisan je algoritam za asembliranje genoma na osnovu skupa kontiga zasnovan na sličnosti bakterije *Lactobacillus paracasei subsp. paracasei BGSJ2-8* sa drugim bakterijama iz roda *Lactobacillus*. Sve kontige su propuštene kroz ***BLAST*** program u cilju pronalaženja sličnosti pojedinačnih kontiga sa DNK sekvencama odgovarajućih organizama u ***NCBI*** nukleotidnoj bazi podataka. Na taj način je dobijen mogući redosled kontiga u genomu bakterije *Lactobacillus paracasei subsp. paracasei BGSJ2-8*.

U ovoj glavi su detaljno opisani pojedinačni koraci algoritma razvijenog od strane autora. Struktura algoritma je prikazana sledećim dijagramom.



3.1 Priprema podataka

Skup od 207 kontiga bakterije *Lactobacillus paracasei subsp. paracasei BGSJ2-8* predstavlja ulazni skup podataka razvijenog algoritma. Svaka od kontiga prolazi kroz procese **blastovanja** i **GC analize**[1], kako bi se pronašle njihove sličnosti sa DNK sekvencama bakterija iz roda *Lactobacillus* i moguće pozicije u genomu. Izdvojene informacije se čuvaju u relacionoj bazi podataka i koriste u daljim koracima algoritma.

3.1.1 Blastovanje kontiga

Bioinformatički centar NCBI

Najpoznatije javno dostupne baze podataka sekvenci DNK i proteina osnovane su i održavane u bioinformatičkim centrima kao što su ***EBI*** (eng. *European Bioinformatics Institute*), ***NCBI*** (eng. *National Center for Biotechnology Information*) i ***GenomeNet***. U ovom radu su korišćeni programi i podaci iz *NCBI* baze. ***NCBI*** je osnovan 1988. godine. Misija centra jeste razvoj novih informacionih tehnologija za razumevanje osnovnih molekularnih i genetičkih procesa koji kontrolišu zdravlje ili razvoj bolesti. ***NCBI*** održava automatizovane sisteme za čuvanje i analizu podataka o molekularnoj biologiji, biohemiji i genetici. Pored toga, centar istražuje napredne metode za obradu informacija putem računara kako bi se analizirala struktura i funkcija biološki značajnih molekula.

Na internet stranicama ***NCBI*** centra se nalaze računarski programi i baze podataka vezane za oblast bioinformatike. Najznačajnija od tih baza podataka je ***GenBank***, koja sadrži sekvence dobijene razmenom podataka sa drugim međunarodnim bazama podataka (***EMBL*** i ***DDBJ***), kao i od strane pojedinačnih laboratorija. Programski alati koje korisnik može da pronađe na stranicama centra su:

- ***BLAST***, program za traženje poklapanja između DNK ili proteinskih sekvenci
- ***ORF Finder***, program za nalaženje otvorenih okvira čitanja
- ***Electronic PCR***, program za identifikovanje *sequence tagged sites (STSs)* unutar DNK sekvenci i
- ***Sequin*** i ***BankIt***, programi za skladištenje sekvenci u baze podataka.

Bioinformatički program *BLAST*

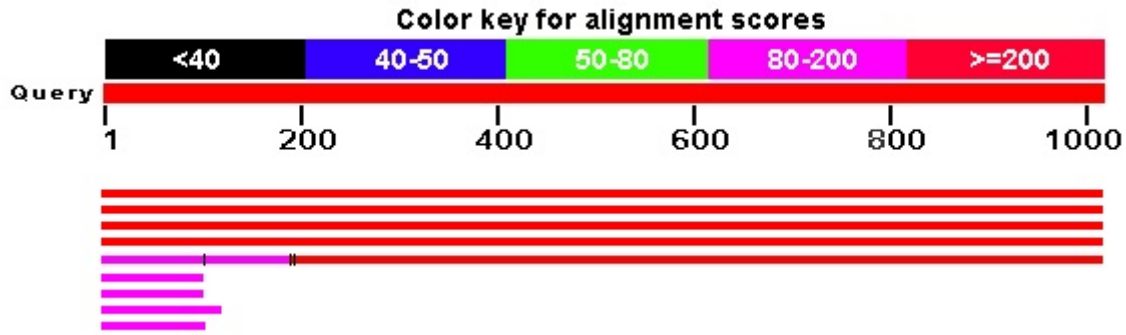
Poređenjem nove, nepoznate sekvence, dobijene sekvenciranjem genoma, sa anotiranim sekvencama moguće je da se predvidi njena biološka struktura, a samim tim i funkcija. Kako je rastao broj dostupnih sekvenci DNK i proteina, raslo

[1] Ako je kontiga mala, onda ne prolazi kroz proces GC analize zbog nemogućnosti podele sekvence na prozore odgovarajućih veličina.

je i interesovanje za razvoj računarskih programa za analizu istih. U početku su korišćene metode grube sile koje nisu mogle da odrede sličnost između sekvenci koje sadrže delove koji se ne poklapaju. Takođe je bilo onemogućeno pronalaženje poravnanja sekvenci koje nisu dovoljno slične ili su vrlo dugačke. Zbog toga su razvijane alternativne, brže metode koje upotrebljavaju drugačije algoritme.

BLAST (eng. *Basic Local Alignment Search Tool*) je računarski program koji se danas najčešće koristi za pronalaženje sličnih sekvenci. Nepoznatu sekvencu *BLAST* pokušava da poravna sa svakom sekvencom koja se nalazi u bazi podataka sekvenci DNK ili proteina. On traži delove sekvenci koji pokazuju najveći stepen sličnosti. Izlaz iz programa su one sekvence koje sadrže deo DNK (ili proteina) dovoljno sličan (u smislu statističke značajnosti) ulaznoj sekvenci. Najčešće korišćene verzije računarskog programa *BLAST* su ***blastn*** (namenjen poravnanju DNK sekvenci) i ***blastp*** (namenjen poravnanju sekvenci proteina). Za potrebe ovog rada korišćen je program ***blastn***.

Uz pomoć navedenog programa je izvršeno poređenje svih 207 kontiga sa DNK sekvencama u nukleotidnoj bazi podataka na ***NCBI*** serveru (`http://blast.ncbi.nlm.nih.gov`). Blastovani su krajevi svake kontige, pri čemu je kao reprezentativan uzorak uzimano po 1000 nukleotida sa oba kraja. Ako je dužina kontige manja od 2000 nukleotida, onda je vršeno blastovanje cele kontige. Rezultat ovog procesa je sličnost kontiga sa drugim organizmima. Međutim, ako blastovanje nije dalo rezultate za početak ili kraj kontige, onda je uzorak proširen na oko 1000 nukleotida nakon ili pre prvobitnog uzorka.



Slika 3.1: Rezultat blastovanja prvih 1.000 nukleotida kontige 125

Horizontalna osa (1-1000) na slici 3.1 odgovara unetoj sekvenci. Crvena boja ukazuje na vrlo dobar kvalitet poklapanja sa odgovarajućim organizmima.

Description	Max score	Total score	Query coverage	E value	Max ident
Lactobacillus casei BD-II, complete genome	1873	6553	100%	0.0	99%
Lactobacillus casei LC2W, complete genome	1873	6553	100%	0.0	99%
Lactobacillus casei BL23 complete genome, strain BL23	1873	6553	100%	0.0	99%
Lactobacillus casei str. Zhang, complete genome	1823	7226	100%	0.0	99%
Lactobacillus casei ATCC 334, complete genome	1399	6198	100%	0.0	100%

Slika 3.2: Lista poklapanja kontige 125 i DNK sekvenci odgovarajućih organizama

Prva kolona na slici 3.2 prikazuje organizme koji pokazuju sličnosti sa prvih 1000 nukleotida kontige 125. Preostale kolone opisuju kvalitet poklapanja.

```
 Score = 1873 bits (1014),  Expect = 0.0
 Identities = 1018/1020 (99%), Gaps = 0/1020 (0%)
 Strand=Plus/Plus

Query  1        CGCTGATGTTGTTAGGCGACTTCGAGACCGCGGGCTTTGCGGGCTCGAAGCGCCGTCACA  60
                ||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||
Sbjct  2925817  CGCTGATGTTGTTAGGCGACTTCGAGACCGCGGGCTTTGCGGGCTCGCAGCGCCGTCACA  2925876

Query  61       GCGTTCCAGCCCCAGCTGGCCGGAGATTGCGGAGTTTGGCACGGCAAGAAAGTTTTAGCC  120
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2925877  GCGTTCCAGCCCCAGCTGGCCGGAGATTGCGGAGTTTGGCACGGCAAGAAAGTTTTAGCC  2925936

Query  121      TAGCATCTTCCTGCGATTATTCAAAATTTAGTGAATCACAAGTCCGATGTGAGAATTATG  180
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2925937  TAGCATCTTCCTGCGATTATTCAAAATTTAGTGAATCACAAGTCCGATGTGAGAATTATG  2925996
```

Slika 3.3: Poklapanje kontige 125 i DNK sekvence odgovarajućeg organizma iz roda *Lactobacillus*

Lista poklapanja, kao i detalji o svakom poravnanju izdvojeni su iz izvornih *HTML* datoteka pomoću programa[2] napisanog u programskom jeziku ***JAVA*** i iskorišćeni za formiranje tabele u relacionoj bazi podataka, čija je struktura prikazana u sledećoj tabeli:

Key	Name	Data type	Length	Nullable
🔑	KONTIGA	CHARACTER	11	No
🔑	POC_KRAJ	CHARACTER	1	No
🔑	GENOM	VARCHAR	15	No
	ORGANIZAM	VARCHAR	1000	Yes
	UPARENA_STRUKTURA	VARCHAR	1000	Yes
	SKOR	REAL	4	Yes
	OCEKIVANO	DOUBLE	8	Yes
	PROCENAT	SMALLINT	2	Yes
	PRAZNINA	SMALLINT	2	Yes
	LANAC	VARCHAR	10	Yes
🔑	POCETAK_UPARENOG	SMALLINT	2	No
	KRAJ_UPARENOG	SMALLINT	2	Yes
🔑	POCETAK_U_GENOMU	INTEGER	4	No
🔑	KRAJ_U_GENOMU	INTEGER	4	No

Slika 3.4: Struktura tabele *BLAST_REZULTATI* u relacionoj bazi podataka

[2]Izvorni kod programa se nalazi na kompakt disku (*CD-R*) koji prati ovaj rad.

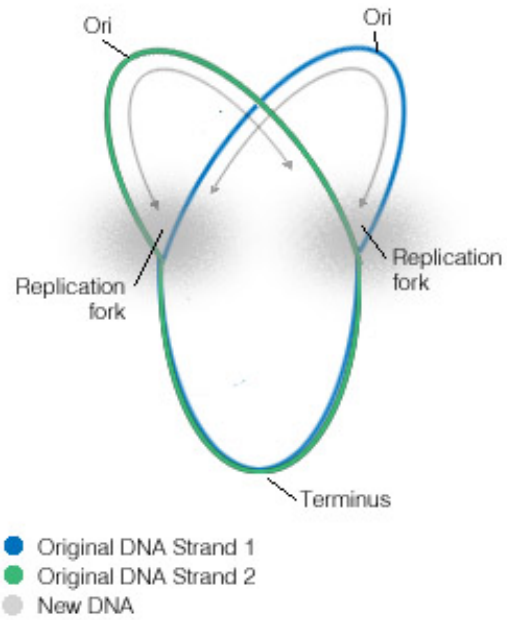
Tabela pored informacija o poklapanju kontiga i DNK sekvenci sličnih organizama sadrži i detalje o kraju kontige koji je blastovan i genu koji sadrži.

Proces blastovanja svih kontiga je ponovljen nedugo nakon prve iteracije. Razlog je pojavljivanje dve nove bakterije koje su pokazale visok stepen poklapanja sa bakterijom *Lactobacillus paracasei subsp. paracasei BGSJ2-8*, a koje se nisu nalazile u nukleotidnoj bazi podataka u trenutku kada je blastovanje vršeno prvi put. Nakon prvog punjenja relacione baze podataka, tabela **BLAST_REZULTATI** je sadržala 7.154 sloga, a nakon drugog 9.173 sloga.

Pokazano je i da pored kontiga koje pokazuju sličnosti sa drugim organizmima postoje i one koje ili nemaju poklapanja sa drugim organizmima, ili ta poklapanja nisu interesantna za ovaj rad. Korektno pozicioniranje takvih kontiga u genomu je znatno otežano i zahteva neke druge metode eksperimentalnog tipa.

3.1.2 GC analiza

GC analiza predstavlja metodu kumulativnih dijagrama koja pokazuje da se sastav nukleotida genoma menja u dve tačke razdvojene okvirno polovinom njegove dužine. Ove tačke se poklapaju sa mestima početka i kraja replikacije DNK, odnosno ***origin***-a i ***terminus***-a za sve bakterije u kojima takva mesta postoje.



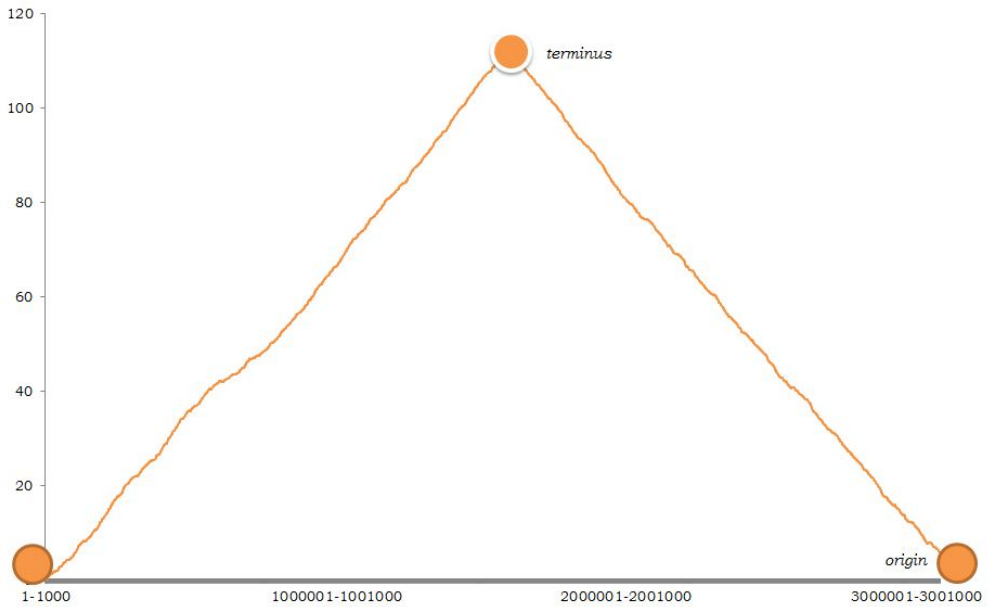
Slika 3.5: Početak i kraj replikacije DNK kod kružnog bakterijskog hromozoma

Činjenica da vodeći lanac sadrži više guanina od citozina se koristi za predviđanje lokacija *origin*-a i *terminus*-a u drugim bakterijskim genomima,

pa može da se iskoristi i za približno određivanje pozicija kontiga u genomu analizirane bakterije. U tu svrhu je korišćena mera sumiranja $\frac{(G-C)}{(G+C)}$ u susednim *prozorima* celokupnih genoma bakterija iz roda *Lactobacillus*. Ona dostiže svoj maksimum u ***terminus***-u, a minimum u ***origin***-u. Kako bi grafici dobijeni uz pomoć ove mere bili ilustrativni, potrebno je da se vodi računa o veličini prozora. Zbog toga je ispitivanje vršeno na prozorima različitih veličina.

Primenom GC analize na skup kontiga bakterije *Lactobacillus paracasei subsp. paracasei BGSJ2-8* ispitano je u kojoj *polovini* genoma se nalaze pojedinačne kontige, što je značajno olakšalo proces asembliranja celokupnog genoma. Te informacije su unete u relacionu bazu podataka i imaju važnu ulogu u daljim koracima algoritma.

Na GC kumulativnom dijagramu genoma bakterije *Lactobacillus casei LC2W* prikazanom na slici ispod lako se uočavaju ***origin*** i ***terminus***.



Slika 3.6: GC dijagram bakterije *Lactobacillus casei LC2W* sa veličinom prozora od 1000 nukleotida

Mera sumiranja $\frac{(G-C)}{(G+C)}$ je primenjena i na pojedinačne kontige koje sadrže bar 10.000 nukleotida[3], čime je pokazano da se u kontigi 125 nalazi tačka *origin*, odnosno spoj kraja i početka genoma bakterije *Lactobacillus paracasei subsp. paracasei BGSJ2-8*.

[3]Jer je kao veličina prozora uzeto 5.000, a zbog kumulativnosti mere zahteva se postojanje bar dva prozora.

Slika 3.7: GC dijagram kontige 125 sa veličinom prozora od 1000 nukleotida

3.2 Filtriranje podataka

Kako bi se izdvojili relevantni podaci iz skupa podataka dobijenog u prvom koraku algoritma, potrebno je da se definišu granične vrednosti za neke od atributa. U dogovoru sa saradnicima iz Instituta za molekularnu genetiku i genetičko inženjerstvo Univerziteta u Beogradu, posmatrani su samo slogovi u kojima kontige pokazuju sličnost veću od 90% sa organizmima iz roda *Lactobacillus*. U procesu asembliranja genoma nisu učestvovale kontige koje pripadaju plazmidu bakterije, već su one analizirane zasebno.

U ovom koraku su izdvojeni i najsličniji organizmi bakteriji *Lactobacillus paracasei subsp. paracasei BGSJ2-8*. Na osnovu zajedničkih karakteristika DNK sekvenci bakterija iz roda *Lactobacillus* u daljim koracima algoritma je vršeno spajanje kontiga.

Naziv organizma	**Broj slogova**
Lactobacillus casei ATCC 334	722
Lactobacillus casei BL23	627
Lactobacillus casei LC2W	620
Lactobacillus casei BD-II	614
Lactobacillus casei str. Zhang	524

Tabela 3.1: Najsličniji organizmi bakteriji *Lactobacillus paracasei subsp. paracasei BGSJ2-8* i broj slogova u relacionoj bazi podataka

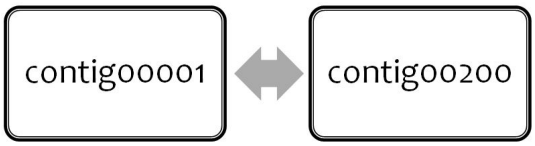
3.3 Spajanje kontiga

Dve kontige mogu da se spoje u potencijalnu superkontigu ako i samo ako su zadovoljeni sledeći uslovi:

- obe kontige pokazuju sličnost sa istim organizmom
- početak druge kontige u sličnom organizmu se nalazi na udaljenosti manjoj od 1000 nukleotida od kraja prve kontige
- orijentacije oba lanca moraju da se poklope i
- GC analiza obe pojedinačne kontige mora da pokazuje isti rast vrednosti mere sumiranja[4].

Na taj način su formirani parovi spojenih kontiga na osnovu kojih se dalje konstruišu stabla pretrage i posmatraju mogući kandidati za superkontige.

3.3.1 Primer spajanja kontiga



Slika 3.8: Primer spajanja dve kontige

Kraj prve i početak druge kontige su pokazali sličnost $\geq 90\%$ sa sledećim organizmima:

- *Lactobacillus casei BD-II*
- *Lactobacillus casei LC2W*
- *Lactobacillus casei ATCC 334* i
- *Lactobacillus casei BL23*.

Početak kontige 200 u bilo kom od navedenih organizama se nalazi neposredno iza kraja kontige 1. Sličnost obe kontige je uočena na *plus* lancima organizama. Na spoju dve kontige se nalazi hipotetički protein.

[4]Uslov ne mora da bude zadovoljen ako se radi o maloj kontigi na koju GC analiza ne može da se primeni.

Dužina kontige 1	969 nukleotida
Dužina kontige 200	512 nukleotida
Dužina potencijalne superkontige	1481 nukleotida

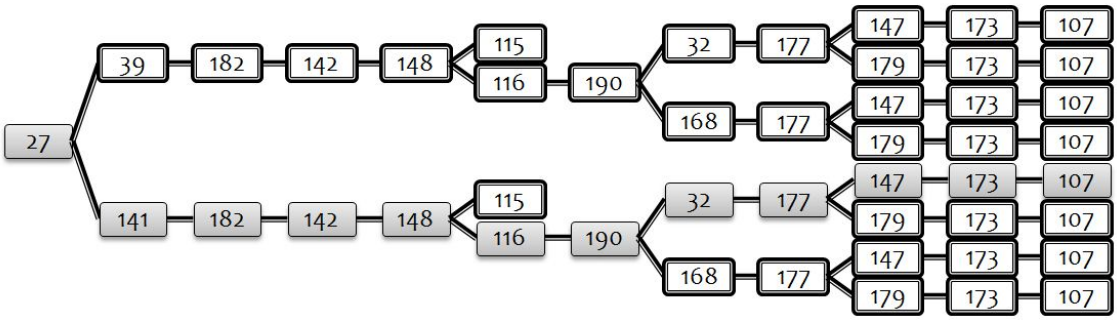
Tabela 3.2: Dužina potencijalne superkontige

Kombinacija pokazuje sličnost cele sekvence od 99% sa gore navedenim organizmima.

3.4 Konstrukcija stabala pretrage

Na osnovu izdvojenih parova, za svaku kontigu se posmatraju najsličniji organizmi i formiraju stabla koja predstavljaju sve moguće kombinacije za spajanje dve ili više kontiga. Formiranje stabala pretrage može da bude otežano zbog prirode nekih kontiga. Na primer, kontiga 114 je dužine 3193 nukleotida i blastovana je dva puta (početak i kraj). Specifično za ovu kontigu je da njen početak daje 100% sličnosti sa jednim organizmom, a njen kraj ne sadrži zadovoljavajuće poklapanje. Svi putevi od korena do listova formiranih stabala pretrage predstavljaju potencijalne superkontige.

3.4.1 Primer



Slika 3.9: Primer stabla pretrage

Svi putevi od korena do listova stabla pretrage sa slike 3.8 predstavljaju potencijalne *mini* superkontige. Svaka od kombinacija pokazuje sličnost cele sekvence (*query coverage* $\geq 99\%$, *max ident* $\geq 95\%$) sa organizmima iz roda *Lactobacillus*, što se objašnjava činjenicom da su kontige 32 i 168, kao i 147 i 179, identične do na par različitih nukleotida.

Glava 4

Analiza rezultata

Sve moguće kombinacije dobijene konstrukcijom stabala pretrage prolaze kroz ponovni proces blastovanja. Svaka kombinacija koja pokaže sličnost $\geq 90\%$ sa odgovarajućim organizmom, na nivou cele sekvence, ulazi u uži izbor. Međutim, pored uspešno formiranih superkontiga, postoje i one kontige koje nisu spojene ni sa jednom drugom iz jednog od sledeća dva razloga:

- ne postoji sličnost sa drugim organizmima u nukleotidnoj bazi podataka na NCBI serveru, ili
- ne postoji sličnost sa bakterijama iz roda *Lactobacillus*.

Kontige koje ne pokazuju sličnost ni sa jednim organizmom su predstavljene sledećom tabelom:

Kontiga	**Dužina kontige**
contig00012	1290
contig00016	553
contig00055	152
contig00064	265
contig00071	333
contig00090	467
contig00104	1905
contig00119	751
contig00153	1641
contig00175	241
contig00205	274
Ukupna dužina	**7872 nukleotida**

Tabela 4.1: Kontige koje ne pokazuju sličnost sa drugim organizmima u nukleotidnoj bazi podataka na NCBI serveru

Kontige koje ne pokazuju sličnost sa organizmima iz roda *Lactobacillus*:

Kontiga	**Dužina kontige**
contig00059	331
contig00061	420
contig00067	256
contig00069	330
contig00073	409
contig00076	289
contig00077	505
contig00079	547
Ukupna dužina	**3087 nukleotida**

Tabela 4.2: Kontige koje ne pokazuju sličnost sa organizmima iz roda *Lactobacillus*

U okviru Instituta za molekularnu genetiku i genetičko inženjerstvo Univerziteta u Beogradu izolovana je nukleotidna sekvenca plazmida *pSJ2-8* iz soja *Lactobacillus paracasei subsp. paracasei BGSJ2-8*. U njen sastav ulaze kontige 30, 42, 114, 144, 165 i 184.

4.1 Primeri superkontiga

U daljem tekstu su prikazane neke od superkontiga dobijene spajanjem pojedinačnih kontiga. One su zbog svojih dužina iskorišćene za ilustraciju rekonstrucije celokupnog genoma u sledećem odeljku.

Superkontiga #1

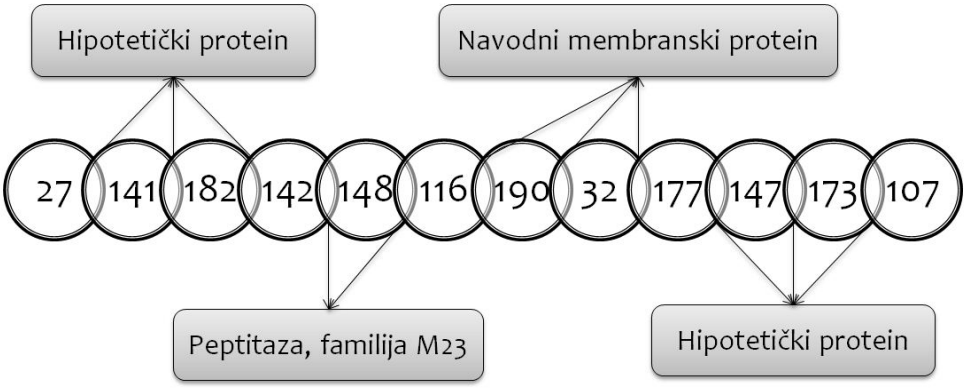
Superkontiga je dobijena spajanjem kontiga 27, 141, 182, 142, 148, 116, 190, 32, 177, 147, 173 i 107.

Slika 4.1: Superkontiga #1

Dužina superkontige je 8.291 nukleotida. Kombinacija pokazuje sličnost cele sekcence (*query coverage* $= 99\%$, *max ident* $= 95\%$) sa organizmima:

- *Lactobacillus casei BD-II*
- *Lactobacillus casei LC2W*
- *Lactobacillus casei BL23* i
- *Lactobacillus casei ATCC 334*.

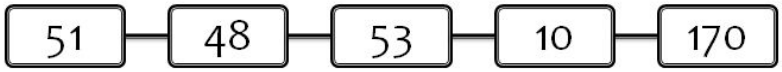
Geni na spojevima pojedinačnih kontiga koji kodiraju odgovarajuće proteine su prikazani na sledećoj slici. Činjenica da se na svakom spoju kontiga nalazi odgovarajući gen daje potvrdu da je spajanje pojedinačnih kontiga izvršeno na korektan način.



Slika 4.2: Geni na spojevima pojedinačnih kontiga

Superkontiga #2

Superkontiga je dobijena spajanjem kontiga 51, 48, 53, 10 i 170.



Slika 4.3: Superkontiga #2

Dužina superkontige je 537.778 nukleotida. Kombinacija pokazuje zadovoljavajuću sličnost na spojevima kontiga sa organizmima:

- *Lactobacillus casei BD-II*
- *Lactobacillus casei LC2W*
- *Lactobacillus casei BL23*
- *Lactobacillus casei ATCC 334* i
- *Lactobacillus casei str. Zhang*.

Na spoju kontiga 54 i 48 se nalazi *Inorganic pyrophosphatase*, dok se na spojevima kontiga 48 i 53, kao i 53 i 10, nalaze hipotetički proteini. Na spoju kontiga 10 i 170 nema ničega.

Superkontiga #3

Superkontiga je dobijena spajanjem kontiga 162 i 96.

Slika 4.4: Superkontiga #3

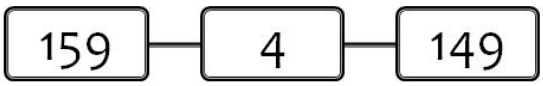
Dužina superkontige je 231.033 nukleotida. Kombinacija pokazuje zadovoljavajuću sličnost na spoju kontiga sa organizmima:

- *Lactobacillus casei BL23* i
- *Lactobacillus casei str. Zhang*.

Na spoju kontiga se nalazi hipotetički protein.

Superkontiga #4

Superkontiga je dobijena spajanjem kontiga 159, 4 i 149.



Slika 4.5: Superkontiga #4

Dužina superkontige je 59.971 nukleotida. Kombinacija pokazuje zadovoljavajuću sličnost na spojevima kontiga sa organizmima:

- *Lactobacillus casei str. Zhang* i
- *Lactobacillus casei ATCC 334*.

Na spoju kontiga 4 i 149 se nalazi *Conserved protein*.

Superkontiga #5

Superkontiga je dobijena spajanjem kontiga 103 i 6.

Slika 4.6: Superkontiga #5

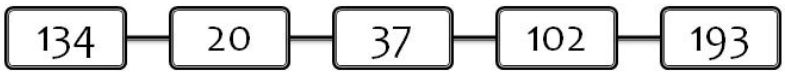
Dužina superkontige je 206.454 nukleotida. Kombinacija pokazuje zadovoljavajuću sličnost na spoju kontiga sa organizmima:

- *Lactobacillus casei BD-II*
- *Lactobacillus casei LC2W* i
- *Lactobacillus casei BL23*.

Na spoju kontiga 103 i 6 se nalazi *Hypothetical cytosolic protein*.

Superkontiga #6

Superkontiga je dobijena spajanjem kontiga 134, 20, 37, 102 i 193.



Slika 4.7: Superkontiga #6

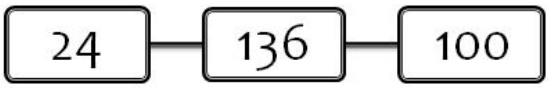
Dužina superkontige je 191.897 nukleotida. Kombinacija pokazuje zadovoljavajuću sličnost na spojevima kontiga sa organizmima:

- *Lactobacillus casei BD-II*
- *Lactobacillus casei LC2W*
- *Lactobacillus casei BL23* i
- *Lactobacillus casei ATCC 334*.

Na spoju kontiga 144 i 20 nema ničega, dok se na spoju kontiga 20 i 37 nalazi *Methyltransferase*. Na spoju kontiga 37 i 102 je uočena *Oxygen-sensitive ribonucleoside-triphosphate reductase*, dok je hipotetički protein pronađen na spoju kontiga 102 i 193.

Superkontiga #7

Superkontiga je dobijena spajanjem kontiga 24, 136 i 100.



Slika 4.8: Superkontiga #7

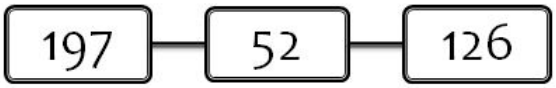
Dužina superkontige je 135.574 nukleotida. Kombinacija pokazuje zadovoljavajuću sličnost na spojevima kontiga sa organizmima:

- *Lactobacillus casei BD-II*
- *Lactobacillus casei LC2W*
- *Lactobacillus casei BL23*
- *Lactobacillus casei str. Zhang*
- *Lactobacillus rhamnosus ATCC 53103 DNA* i
- *Lactobacillus rhamnosus GG*.

Na spojevima kontiga 24 i 136, kao i 136 i 100, nalaze se hipotetički proteini.

Superkontiga #8

Superkontiga je dobijena spajanjem kontiga 197, 52 i 126.



Slika 4.9: Superkontiga #8

Dužina superkontige je 161.219 nukleotida. Kombinacija pokazuje zadovoljavajuću sličnost na spojevima kontiga sa organizmima:

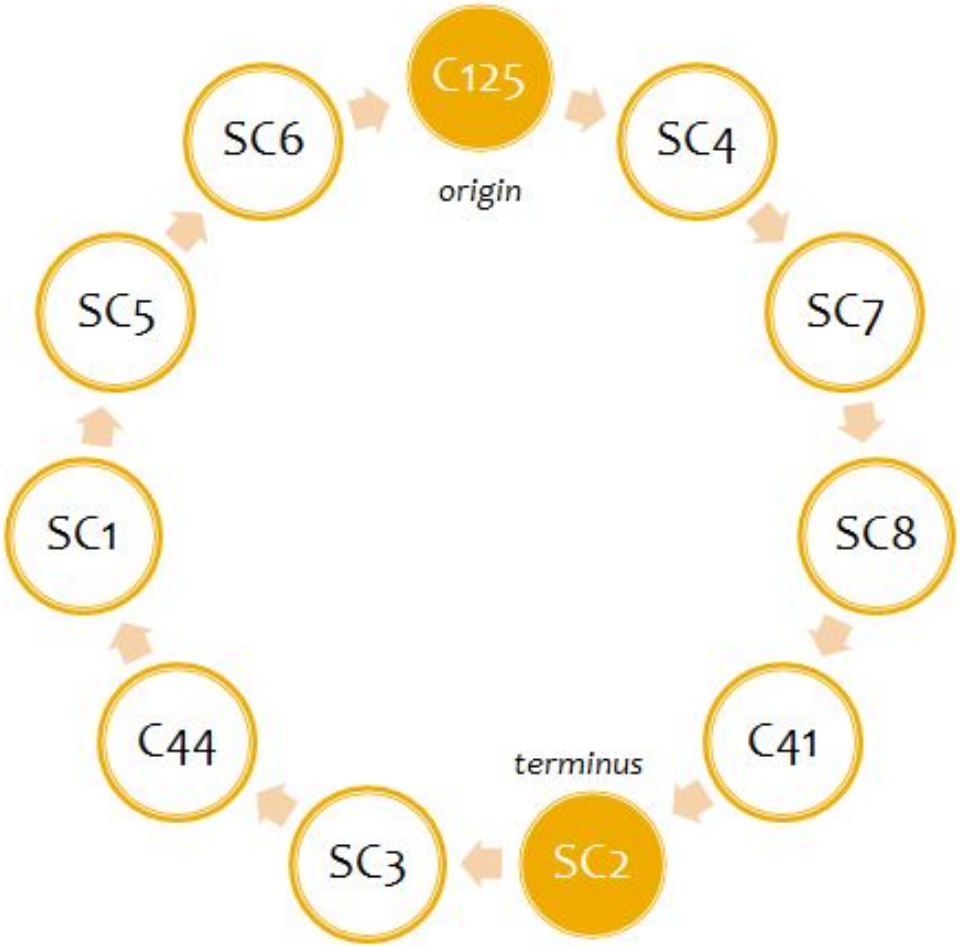
- *Lactobacillus casei BD-II*
- *Lactobacillus casei LC2W*
- *Lactobacillus casei BL23*
- *Lactobacillus casei str. Zhang* i
- *Lactobacillus casei ATCC 334*.

Na spoju kontiga 197 i 52 se nalazi hipotetički protein, dok na spoju kontiga 52 i 126 nema ničega.

4.2 Rekonstrukcija genoma

Koristeći najduže pojedinačne kontige, kao i superkontige formirane ovom metodom, genom većim delom može da se rekonstruiše. Na slici 4.10 je ilustrovana rekonstrukcija genoma na osnovu superkontiga navedenih u prethodnom odeljku, kao i tri kontige najveće dužine koje nisu spojene ovom metodom ni sa jednom drugom kontigom. Već je rečeno da se u kontigi 125 nalazi spoj kraja i početka kružnog molekula DNK. Na okvirno polovini genoma se nalazi superkontiga 2, jer je GC analizom utvrđeno da ona sadrži tačku *terminus*. Superkontige 4, 7 i 8, kao i kontiga 41, pokazuju rast vrednosti mere sumiranja

prilikom GC analize, što potvrđuje korektnost njihovog položaja u prvoj polovini genoma. Na osnovu sličnosti pojedinih fragmenata navedenih superkontiga i kontige sa sličnim organizmima iz roda *Lactobacillus*, utvrđen je njihov redosled u genomu. Na sličan način je predviđen položaj superkontiga 1, 3, 5 i 6, kao i kontige 44. Njihovom GC analizom je pokazano da imaju pad vrednosti kumulativne mere sumiranja, što potvrđuje njihov položaj u drugoj polovini genoma. Dužina ovako rekonstruisanog genoma iznosi 1.946.889 nukleotida.



Slika 4.10: Ilustracija rekonstrukcije genoma

Ostatak genoma može na sličan način u potpunosti da se rekonstruiše, ali je potrebna eksperimentalna provera pretpostavljenih spojeva, kao i utvrđivanje pozicije onih kontiga koje nisu pokazale sličnost sa organizmima interesantnim za ovaj rad.

Glava 5

Zaključak

U ovom radu je prikazan algoritam za asembliranje genoma bakterije *Lactobacillus paracasei subsp. paracasei BGSJ2-8* izolovane iz tradicionalnog domaćeg srpskog sira. Algoritam je zasnovan na sličnosti skupa od 207 kontiga dobijenih sekvenciranjem pomenute bakterije sa drugim sojevima bakterija iz roda *Lactobacillus*.

Sve kontige su propuštene kroz ***BLAST*** program u cilju pronalaženja sličnih organizama na osnovu čijih fragmenata DNK bi moglo da se izvrši spajanje pojedinačnih kontiga posmatrane bakterije. Sve veće kontige su analizirane u smislu određivanja tačaka u kojima se menja sastav nukleotida hromozoma kako bi se odredile okvirne pozicije kontiga u genomu, kao i mesto spajanja kraja i početka kružnog molekula DNK. Iz tabele u relacionoj bazi podataka gde su čuvane informacije o listama poklapanja, kao i detalji o svakom poravnanju dobijeni blastovanjem pojedinačnih kontiga, izdvojeni su samo oni slogovi u kojima su kontige pokazale sličnost $\geq 90\%$ sa organizmima iz roda *Lactobacillus*. Na osnovu prethodno izdvojenih podataka je dalje vršeno spajanje kontiga koje su pokazale sličnost sa istim organizmom i zadovoljile dodatne uslove koji se tiču orijentacije DNK lanaca i pozicija kontiga u sličnim organizmima. Zatim se za svaku pojedinačnu kontigu, koristeći prethodno izdvojene parove i njihove sličnosti sa sojevima bakterija iz identičnog roda, formiraju stabla pretrage koja predstavljaju sve moguće kombinacije za spajanje dve ili više kontiga.

Sve moguće kombinacije puteva od korena do listova stabala pretrage predstavljaju potencijalne supekontige koje prolaze kroz ponovni proces blastovanja. Svaka kombinacija koja pokaže sličnost $\geq 90\%$ sa odgovarajućim organizmom, na nivou cele ulazne sekvence, ulazi u uži izbor. Na taj način je dobijen veliki broj superkontiga izgrađenih spajanjem 3 ili više pojedinačnih kontiga.

Ova metoda nije uspela da predvidi pozicije 11 kontiga koje nisu pokazale sličnost ni sa jednim drugim organizmom iz nukleotidne baze podataka na *NCBI* serveru. Njihova ukupna dužina je 7.872 nukleotida i ostaje saradnicima sa Institutu za molekularnu genetiku i genetičko inženjerstvo Univerziteta u Beogradu da eksperimentalnim metoda utvrde njihove tačne položaje. Slična

situacija je i sa 8 kontiga koje nisu pokazale poklapanja sa organizmima iz roda *Lactobacillus*.

Da bi se rekonstruisala mapa celokupnog genoma potrebno je da se sve potencijalne superkontige eksperimentalnim metodama provere na spojevima pojedinačnih kontiga. Rezultati ovog rada su značajno smanjili skup od 207 kontiga, što direktno uslovljava jeftinije eksperimentalno utvrđivanje spojeva.

U radu je poseban akcenat stavljen na korišćenje GC analize prilikom asembliranja genoma, jer taj metod u svetu nije često korišćen u ove svrhe. Međutim, pozicioniranje kontiga u genomu koristeći ovu metodu nije pouzdano jer se ne zna na kom lancu molekula DNK se one nalaze. Prikazan algoritam asembliranja može da se primeni i na različite probleme asembliranja pojedinačnih čitanja ili kontiga organizama iz bilo kog roda.

Literatura

[1] Andreas D. Baxevanis, B. F. Francis Ouellette: *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, John Wiley & Sons, Inc., ISBN 0-471-38391-0, 2001.

[2] Jean-Michel Claverie, Cedric Notredame: *Bioinformatics for Dummies*, Wiley Publishing, Inc., ISBN 0-470-08985-7, 2007.

[3] David W. Mount: *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, ISBN 0-87969-608-7, 2001.

[4] Hans-Joachim Bockenhauer, Dirk Bongartz: *Algorithmic Aspects of Bioinformatics*, Springer Berlin Heidelberg New York, ISBN 978-3-540-71912-0, 2007.

[5] Peter Clote, Rolf Backofen: *Computaional Molecular Biology: An Introduction*, John Wiley & Sons, Inc., ISBN 0-471-87251-2, 2000.

[6] Nello Cristianini, Matthew W. Hahn: *Introduction to Computational Genomics*, Cambridge University Press, ISBN 0-521-67191-4, 2006.

[7] Andrzej Polanski, Marek Kimmel: *Bioinformatics*, Springer Berlin Heidelberg New York, ISBN 978-3-540-24166-9, 2007.

[8] Alexander Isaev: *Introduction to Mathematical Methods in Bioinformatics*, Springer Berlin Heidelberg New York, ISBN 3-540-21973-0, 2006.

[9] Jin Xiong: *Essential Bioinformatics*, Cambridge University Press, ISBN 0-521-60082-0, 2006.

[10] Andrei Grigoriev: *Analyzing Genomes with Cumulative Skew Diagrams*, Nucleic Acids Research, Vol. 26, No. 10: 2286-2290, 1998.

[11] Granger G. Sutton, Owen White, Mark D. Adams, Anthony R. Kerlavage: *TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects*, Genome Science & Technology, Vol. 1, No. 1, 1995.

[12] Xiaoqiu Huang, Anup Madan: *CAP3: A DNA Sequence Assembly Program*, 1999.

[13] R. Dahm: *Discovering DNA: Friedrich Miescher and the early years of nucleic acid research*, Human genetics, 122 (6): 56581, 2008.

[14] Mark B. Gerstein et al.: *What is a gene, post-ENCODE? History and updated definition*, Genome Research, 17 (6): 669-681, 2007.