



Univerzitet u Beogradu
Matematički fakultet

Master rad

**Određivanje pozicije *riboswitch* sekvenci u genomu
Lactobacillus paracasei subsp. paracasei BGSJ2-8**

Student:
Aleksandar Stefanović

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

1	Uvod	1
1.1	Šta je bioinformatika?	1
1.2	Nukleinske kiseline	2
1.2.1	Primarna i sekundarna struktura nukleinskih kiselina	2
1.3	Gen i genom	4
1.3.1	Organizacija genoma	5
1.4	Regulacija ekspresije gena kod prokariota	6
2	Riboswitch sekvence	9
2.1	Familije <i>riboswitch</i> sekvenci	10
2.1.1	<i>FMN riboswitch (RFN element)</i>	10
2.1.2	<i>TPP riboswitch (THI element)</i>	10
2.1.3	<i>T-box leader riboswitch</i>	11
2.1.4	<i>SAM riboswitch (S box leader)</i>	11
2.1.5	<i>Purine riboswitch</i>	12
2.1.6	<i>Lysine riboswitch</i>	13
2.2	Pronalaženje <i>riboswitch</i> sekvenci	13
3	Pronalaženje <i>riboswitch</i> sekvenci u genomu bakterije <i>Lactobacillus paracasei</i> subsp. <i>paracasei</i> BGSJ2-8	15
3.1	Rfam baza podataka	15
3.1.1	Primer	16
3.1.2	Rezultati	17
3.2	RibEx: veb server za određivanje pozicija <i>riboswitch</i> sekvenci i drugih konzerviranih bakterijskih regulatornih elemenata	17
3.2.1	Veb server	18
3.2.2	Primer	19
3.2.3	Rezultati	19
3.3	RiboSW: alat za pretraživanje <i>riboswitch</i> sekvenci	19
3.3.1	Primer	20
3.3.2	Rezultati	21
3.4	Uporedna analiza rezultata	21
4	Analiza rezultata i ispitivanje korektnosti potencijalnih <i>riboswitch</i> sekvenci	23
4.1	Otvoreni okviri čitanja	23

4.2	Ispitivanje postojanja palindroma u potencijalnim <i>riboswitch</i> sekvencama	25
4.3	Blastovanje mogućih <i>riboswitch</i> sekvenci	26
4.4	Primer korektne <i>riboswitch</i> sekvence	27
5	Zaključak	29
	Literatura	31
6	Dodatak	33
6.1	Pronađene <i>riboswitch</i> sekvence u genomu bakterije <i>Lactobacillus paracasei</i> subsp. <i>paracasei</i> BGSJ2-8	33

Spisak slika

1.1	Hemijska struktura molekula DNK	3
1.2	Sekundarna struktura molekula DNK	3
1.3	Redosled nukleotida u genu <i>kodira</i> protein	5
1.4	Struktura prokariotske ćelije	5
1.5	Struktura informacione RNK	6
1.6	Elektronska mikrografija procesa transkripcije i translacije	7
2.1	Sekundarna struktura <i>FMN riboswitch</i> sekvence	10
2.2	Sekundarna struktura <i>TPP riboswitch</i> sekvence	11
2.3	Sekundarna struktura <i>SAM riboswitch</i> sekvence	12
2.4	Sekundarna struktura <i>Purine riboswitch</i> sekvence	12
2.5	Sekundarna struktura <i>Lysine riboswitch</i> sekvence	13
3.1	Grafički korisnički interfejs programa za pretragu RNK familija u ulaznoj sekvenci na Rfam veb serveru	16
3.2	Izlaz iz programa za pretragu RNK familija u ulaznoj sekvenci na Rfam veb serveru	17
3.3	Grafički korisnički interfejs RibEX veb servera	18
3.4	RibEx pronalazi T-box <i>riboswitch</i>	19
3.5	RiboSW veb interfejs	20
3.6	Izlaz iz RiboSW programa	21
4.1	Struktura tabele RIBOSVICEVI u lokalnoj bazi podataka	23
4.2	Struktura tabele ORFOVI u lokalnoj bazi podataka	24
4.3	Dozvoljeni položaj ORF-ova u odnosu na <i>riboswitch</i> sekvencu	24
4.4	Dozvoljene orijentacije okolnih ORF-ova	25
4.5	Primer postojanja komplementarnog palindroma u mogućoj <i>riboswitch</i> sekvenci	26
4.6	Primer nekorektne <i>riboswitch</i> sekvence	26
4.7	Primer dobro orijentisanih ORF-ova kod <i>riboswitch</i> sekvence	27
4.8	Primer korektne regulacije nizvodnog gena kod <i>riboswitch</i> sekvence u sličnom organizmu	27

Spisak tabela

3.1	Broj pronađenih <i>riboswitch</i> sekvenci po familijama uz pomoć Rfam baze podataka	17
3.2	Broj pronađenih <i>riboswitch</i> sekvenci po familijama uz pomoć RibEX programa	19
3.3	Broj pronađenih <i>riboswitch</i> sekvenci po familijama uz pomoć RiboSW programa	21
3.4	Broj pronađenih <i>riboswitch</i> sekvenci po familijama koristeći RibEx, Rfam i RiboSW	22

Predgovor

U molekularnoj biologiji *riboswitch* sekvenca predstavlja deo molekula informacione RNK koja kontroliše proces u kome se nasledne informacije koje nosi gen koriste za sintezu proteina ili funkcionalne RNK. Cilj ovog rada je određivanje pozicija *riboswitch* sekvenci u skupu kontiga genoma *Lactobacillus paracasei subsp. paracasei* BGSJ2-8 sekvenciranog za potrebe Instituta za molekularnu genetiku i genetičko inženjerstvo Univerziteta u Beogradu (IMGGI). U okviru ovog rada je razmatrano 6 familija *riboswitch* sekvenci od ukupno 22 anotirane i dostupne na internet adresi: <http://rfam.janelia.org/>. Pronalaženje *riboswitch* sekvenci je vršeno uz pomoć postojećih softverskih rešenja. Izloženi su rezultati dobijeni primenom ovih alata. Izvršena je uporedna analiza rezultata i ispitivanje korektnosti potencijalnih *riboswitch* sekvenci, a one koje su zadovoljile definisane uslove su poslate IMGGI na eksperimentalnu proveru njihove korektnosti.

Svi materijali i dobijeni rezultati se nalaze na kompakt disku priloženom uz ovaj master rad.

Zahvalnica

Pre svega želim da se zahvalim svom mentoru, dr Nenadu Mitiću, profesoru Matematičkog fakulteta Univerziteta u Beogradu, bez čije pomoći i vremena koje je 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 priđem ovoj materiji. Zahvalnost dugujem i dr Nataši Golić, dr Jeleni Begović i dr Branku Jovčiću, saradnicima sa IMGGI u Beogradu na nizu korisnih saveta, stručnih smernica i iskrenih ohrabrenja tokom pisanja ovog rada.

Glava 1

Uvod

1.1 Šta je bioinformatika?

Velike količine podataka koje zahtevaju efikasnu obradu dovele su do saradnje naučnika iz oblasti računarstva i molekularne biologije. Posledica stapanja ova dva naučna područja dovela je do pojave nove naučne discipline koja se naziva **bioinformatika**. Bioinformatika može da se opiše i kao *naučna oblast u kojoj se primenom metoda računarskih i matematičkih nauka vrše istraživanja u domenu biomedicinskih nauka*.

Problemi koje bioinformatika rešava su najčešće iz oblasti molekularne biologije, ali postoji i veći broj problema iz drugih oblasti:

- genomika
- proteomika
- modeliranje bioloških sistema
- obrada biomedicinskih slika i signala
- uočavanje i modeliranje korelacija između makro svojstava genskog ili proteinskog sastava
- biohemija i hemija
- i drugo.

Bioinformatika se bavi algoritmima za generisanje novih znanja iz biologije i medicine, kao i za poboljšanje i otkrivanje novih modela izračunavanja. Osnovne karakteristike ove naučne discipline su relativno velika količina podataka, složena obrada podataka i velika dinamika priliva informacija i znanja. Primarni cilj bioinformatike je razumevanje bioloških procesa. Skladištenje, pretraživanje, obrada i analiza velikih količina podataka nemoguća je bez korišćenja posebno razvijenih računarskih programa.

1.2 Nukleinske kiseline

Nukleinske kiseline su složeni biološki makromolekuli.

Postoje dva tipa nukleinskih kiselina:

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

DNK je nosilac bioloških informacija u ćeliji, dok molekuli RNK učestvuju u prenošenju tih informacija i njihovom prevođenju u proteine. Nukleinske kiseline su makromolekuli čiju jedinicu građe predstavljaju nukleotidi.

U izgradnji nukleotida koji formiraju DNK učestvuju:

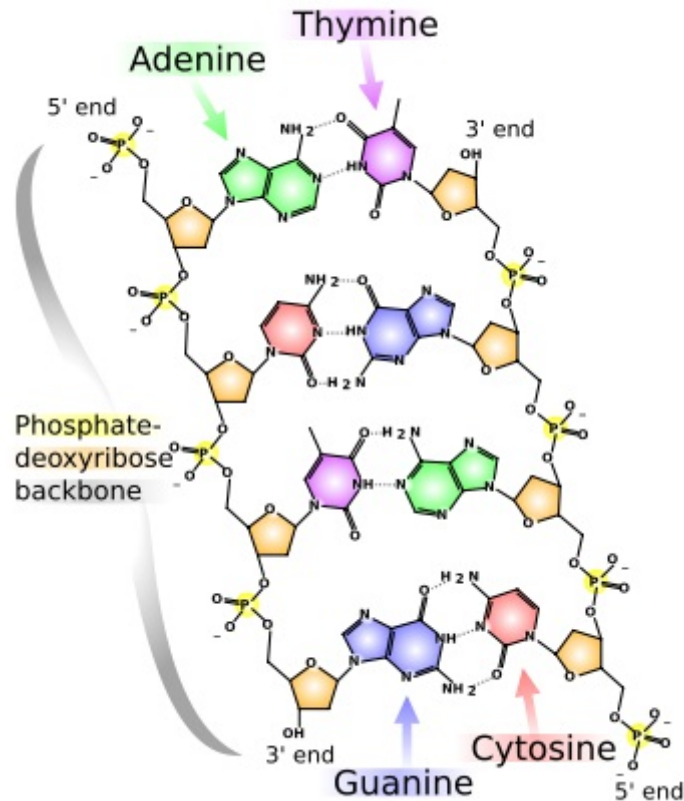
- pentozni šećer *dezoksiriboza*
- purinske baze *adenin* (A) i *guanin* (G), ili pirimidinske baze *citozin* (C) i *timin* (T) i
- kiselinski ostatak fosforne kiseline.

U izgradnji nukleotida koji formiraju RNK učestvuju:

- pentozni šećer *riboza*
- purinske baze *adenin* (A) i *guanin* (G), ili pirimidinske baze *uracil* (U) i *citozin* (C) i
- kiselinski ostatak fosforne kiseline.

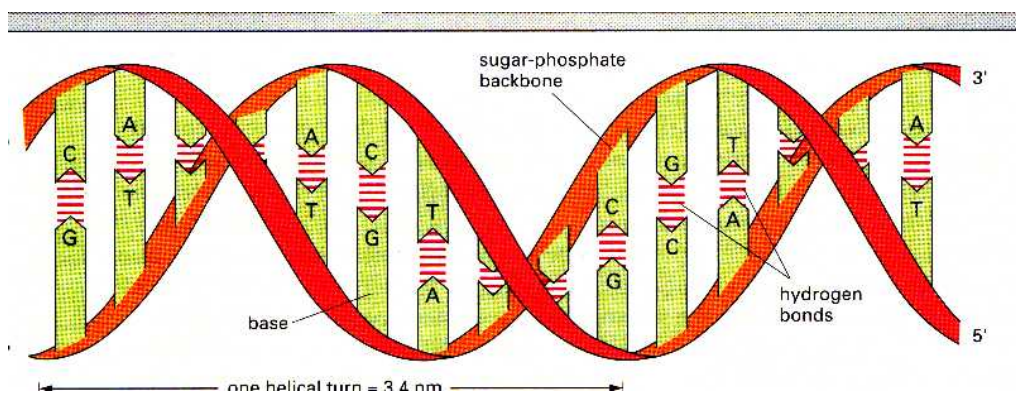
1.2.1 Primarna i sekundarna struktura nukleinskih kiselina

Primarna struktura nukleinskih kiselina predstavlja sadržaj i redosled nukleotida u polinukleotidnom lancu. Utvrđivanje redosleda vezivanja nukleotida u molekulu nukleinskih kiselina predstavlja njenu pravu karakterizaciju. Za određivanje primarne strukture razvijeno je nekoliko metoda.



Slika 1.1: Hemijska struktura molekula DNK

Sekundarna struktura DNK predstavlja prostornu organizaciju polinukleotidnih lanaca u molekulu DNK. Sekundarna struktura DNK je prvi put objavljena 1953. godine od strane *Votsona* i *Krika*. Molekul DNK se sastoji iz dva polinukleotidna lanca obavijena oko centralne ose uz stvaranje dvostrukog heliksa. Dva polinukleotidna lanca su postavljena antiparalelno. Purinske i pirimidinske baze nukleotidnih jedinki se nalaze unutar dvostrukog heliksa, tako što se uvek naspram *adenina* nalazi *timin*, a naspram *guanina* *citozin* (slika 1.2).



Slika 1.2: Sekundarna struktura molekula DNK

1.3 Gen i genom

Genom predstavlja kompletan skup naslednjih informacija jednog organizma[2]. On se sastoji od duge sekvence¹ nukleinskih kiselina koja obezbeđuje informaciju neophodnu za organizam da se izgradi i funkcioniše. Preko složenog niza interakcija, redosled nukleotida u molekulima DNK se koristi za stvaranje RNK i proteina na odgovarajućem mestu i u odgovarajuće vreme. Nastali proteini učestvuju u metaboličkim reakcijama neophodnim za život.

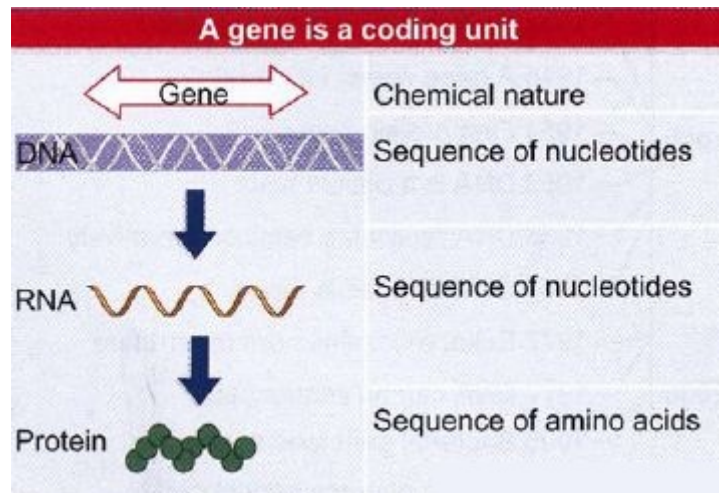
Genom se, fizički posmatrano, sastoji od jednog ili više molekula DNK, dok se funkcionalno posmatrano sastoji iz gena koji nose zapis za različite proteine i molekule RNK. Broj gena u genomima različitih organizama se veoma razlikuje. Genom može da se definiše i kao određen redosled nukleotida u svim molekulima DNK iz kojih se sastoji.

Imajući u vidu prethodno navedeno, može da se napravi sledeće poređenje:

- genom = biblioteka
- molekuli DNK = knjige
- geni = poglavlja knjige
- redosled nukleotida u molekulima DNK = tekst knjige
- kodoni = reči
- nukleotidi = slova.

Gen je fizička i funkcionalna jedinica nasleđivanja koja prenosi naslednu poruku iz generacije u generaciju[1]. Predstavljen je delom molekula DNK koji se prepisuje u RNK i zasebno je regulisan. Drugim rečima, gen je redosled nukleotida u molekulu DNK koji određuje hemijsku strukturu specifičnog polipeptida ili molekula RNK. Pokazano je da redosled nukleotida DNK određuje redosled aminokiselina u proteinu, jer se redosled nukleotida molekula DNK prepisuje u redosled nukleotida RNK procesom *transkripcije*, a redosled nukleotida molekula RNK prevodi u redosled aminokiselina u proteinu procesom *translacije* (slika 1.3).

¹ Sekvenca nukleinskih kiselina predstavlja redosled nukleotida.

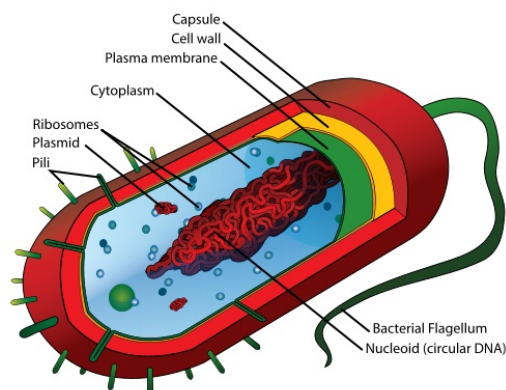


Slika 1.3: Redosled nukleotida u genu *kodira* protein

1.3.1 Organizacija genoma

Prokarioti su jednoćelijski organizmi čije ćelije su male, jednostavne građe, obavijene ćelijskim zidom i membranom, a nemaju jedro niti ćelijske organele, osim ribozoma. Prokariotski organizmi su arheo bakterije i bakterije, dok su ostali organizmi, bilo jednoćelijski, bilo višećelijski, eukarioti. 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.

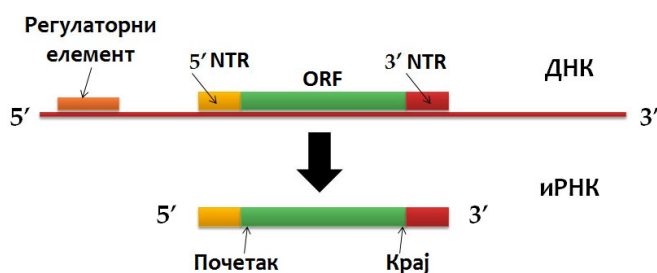


Slika 1.4: Struktura prokariotske ćelije

1.4 Regulacija ekspresije gena kod prokariota

Pod ekspresijom gena podrazumevamo sintezu funkcionalnog proteinskog proizvoda za čiju je sintezu genetička informacija zapisana u nekom genu. Ekspresija gena započinje transkripcijom kako bi se prepisala genetička informacija sadržana u genu u molekulu informacione RNK. Molekul informacione RNK se vezuje sa ribozomima i tokom procesa translacije se odvija biosinteza proteina.

Informaciona RNK (iRNK) nastaje prepisivanjem (transkripcijom) strukturnih gena koji sadrže upustvo za sintezu proteina. Uloga iRNK je da tu informaciju za sintezu polipeptida prenese do ribozoma (mesto sinteze proteina), tj. da posluži kao neposredna matrica za njegovu sintezu. Tako se prenosi informacija za sintezu proteina. Sinteza iRNK počinje kada je ćeliji potreban neki protein, a kada se obezbedi dovoljna količina proteina ona biva razgrađena. Informaciona RNK uspostavlja različitost sastava proteina u ćeliji. Zato je iRNK po strukturi i veličini najraznovrsniji tip RNK u ćeliji, a po količini iznosi svega 5-10%.

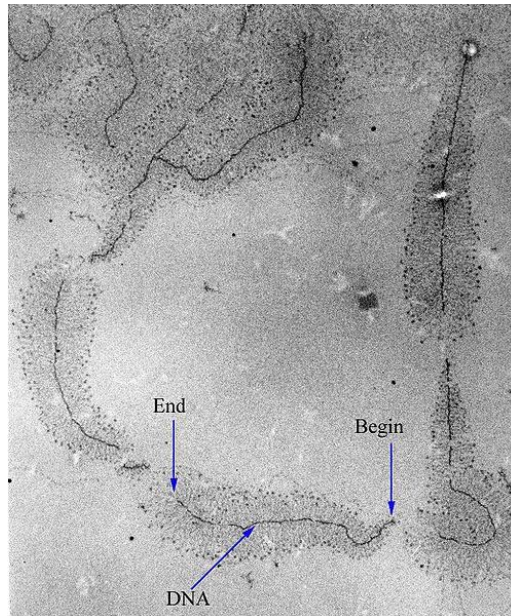


Slika 1.5: Struktura informacione RNK

[Legenda: ORF - otvoreni okvir čitanja, NTR - netranslirajuća RNK]

U biologiji, okvir čitanja je način čitanja nukleotidne DNK (RNK) sekvence po tri slova (po kodonu) koji predstavljaju aminokiseline. Otvoreni okvir čitanja (ORF) je okvir čitanja koji sadrži start kodon i sekvencu koja se može posmatrati kao više grupa od po 3 nukleotida, ali ne sadrži stop kodon u datom okviru čitanja[11]. Start kodon označava mesto gde ribozom počinje translaciju (prevođenje) RNK u aminokiselinu. Stop kodon je triplet nukleotida u okviru iRNK koji označava završetak translacije. Najčešći start kodon je AUG, dok su najčešći stop kodoni UAG, UGA, UAA.

Bakterije moraju da budu vrlo prilagodljive i fleksibilne, pa čitav svoj metabolizam prilagođavaju promenama sredine u kojoj žive. Razvijen je čitav niz adaptacija koje uključuju to da se proces ekspresije gena odvija na vrlo efikasan način. Jedna od tih adaptacija jeste da se transkripcija i translacija odvijaju istovremeno. Nema procesa obrade informacione RNK, pa je ekspresija gena znatno kraća nego kod eukariotskih ćelija (slika 1.6).



Slika 1.6: Elektronska mikrografija procesa transkripcije i translacije

U molekularnoj biologiji *riboswitch* sekvenca predstavlja deo molekula informacione RNK koja kontroliše ekspresiju gena, odnosno proces u kome se nasledne informacije koje nosi gen koriste za sintezu proteina ili funkcionalne RNK. Cilj ovog rada je određivanje pozicija *riboswitch* sekvenci u skupu kontiga genoma bakterije *Lactobacillus paracasei subsp. paracasei* BGSJ2-8 sekvenciranog za potrebe IMGGI Univerziteta u Beogradu. Pomenuta bakterija pripada diviziji *Firmicutes*, porodici *Lactobacillaceae*, rodu *Lactobacillus* i vrsti *Lactobacillus paracasei*.

Glava 2

Riboswitch sekvence

U molekularnoj biologiji, *riboswitch* (*switch* = "prekidač") je deo molekula informacione RNK koji može direktno vezati mali ciljani molekul, a čije vezivanje utiče na aktivnost gena[10]. *Riboswitch* sekvenca je palindromska sekvenca¹ nukleinske kiseline (DNK ili RNK) koja je identična bez obzira da li se čita na kodirajućem ili nekodirajućem lancu sa kojim gradi dupli heliks.

Glavna uloga *riboswitch* sekvenci je u regulaciji genske ekspresije, jer zaključavaju gene koje regulišu i tako sprečavaju njihovu ekspresiju u momentu kada ona nije poželjna. S obzirom da je u pitanju regulacija na nivou translacije (prepisivanje sa informacione RNK na proteine), moguća je brža reakcija bakterijske ćelije na spoljašnje uslove, pa je otključavanje odgovarajućih gena brže i proteini se sintetišu pre nego kod gena koji su regulisani na nivou transkripcije (prepisivanje sa DNK na informacionu RNK). Razlog je taj što u ćeliji već postoji spremna dovoljna količina informacione RNK, što bakterijama koje koriste regulaciju *riboswitch* sekvencama daje selektivnu prednost u sredini u kojoj žive u odnosu na bakterije koje nemaju ovaj sistem regulacije.

Najpoznatiji primer *riboswitch* sekvenci su geni za patogenost kod bakterije *Listerije* i nekih drugih bakterijskih patogena, gde su ovim sekvencama regulisani geni koji obezbeđuju prilagodljivost bakterija za život u gastrointestinalnom traktu ljudi i životinja. Na primer, patogene bakterije kod kojih su *riboswitch* sekvence prvi put pronađene uključuju virulentene gene u situacijama kada nema dovoljno slobodnih nutrienata. Tada bakterija uzima potrebne nutrijente od domaćina u kome živi. Takođe, najnoviji radovi govore i o mogućoj upotrebi *riboswitch* sekvenci kao meta za antibiotsku terapiju, za proizvodnju lekova male molekulske mase koji bi selektivno napadali odgovarajuće *riboswitch* sekvence, tako što bi ih držali zaključane i sprečavali virulentnost bakterija.

¹Invertovane sekvence predstavljaju dva niza nukleotida koji su komplementarni jedan drugom kada se čitaju u suprotnom smeru. Na primer, 5'-GACTGCNNNNNNGCAGTC-3'. Invertovane sekvence se mogu međusobno spariti formirajući stukturu ukosnice u jednonančanoj nukleinskoj kiselini, ili krstastu stukturu u dvolančanoj nukleinskoj kiselini. Takav niz se zove palindromska sekvenca i ona je identična kada se u dva lanca čita u istom smeru.

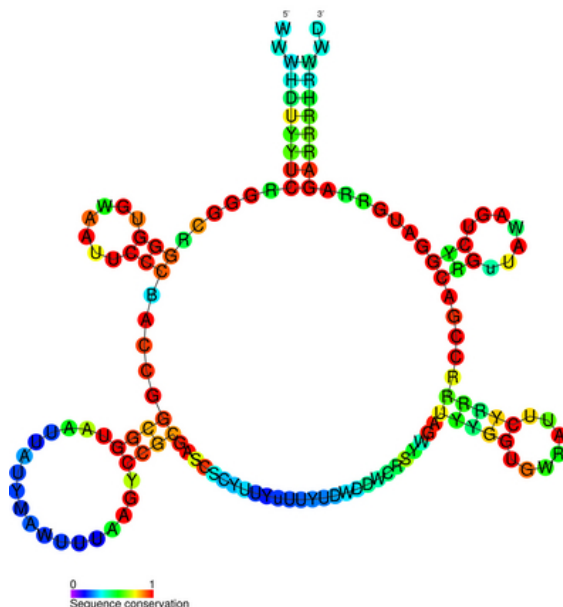
2.1 Familije *riboswitch* sekvenci

Do sada su anotirane 22 familije *riboswitch* sekvenci koje su dostupne na internet adresi: <http://rfam.janelia.org/>. Za potrebe ovog rada je razmatrano 6 familija koje su opisane u daljem tekstu.

Boje na slikama koje predstavljaju sekundarne strukture familija *riboswitch* sekvenci prikazanih u daljem tekstu, označavaju njihove konzervisanosti. Ljubičasta boja ima vrednost 0 a crvena vrednost 1, pri čemu 0 označava najmanju, a 1 najveću konzervisanost.

2.1.1 *FMN riboswitch (RFN element)*

RFN element je *riboswitch* koji direktno vezuje flavin mononukleotid (*FMN*) u nedostatku proteina (slika 2.1).

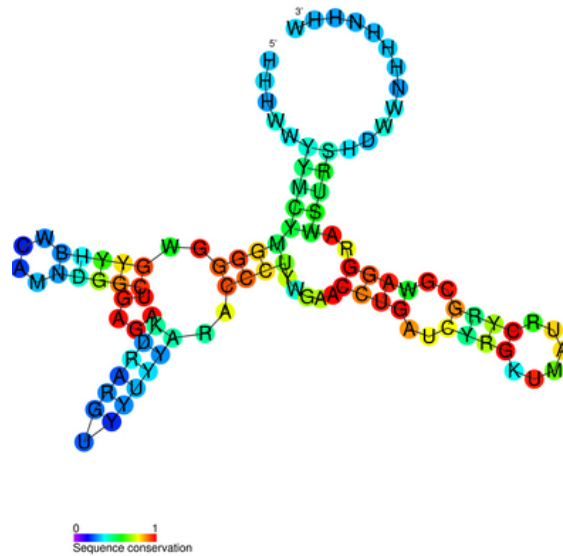


Slika 2.1: Sekundarna struktura *FMN riboswitch* sekvence

Više informacija o familiji može da se nađe na internet adresi: <http://rfam.janelia.org/cgi-bin/getdesc?name=FMN>.

2.1.2 *TPP riboswitch (THI element)*

THI element je *riboswitch* koji direktno vezuje *TPP* (tiamin pirofosfat), čime uz pomoć različitih mehanizama reguliše ekspresiju gena kod bakterija (slika 2.2).



Slika 2.2: Sekundarna struktura *TPP riboswitch* sekvence

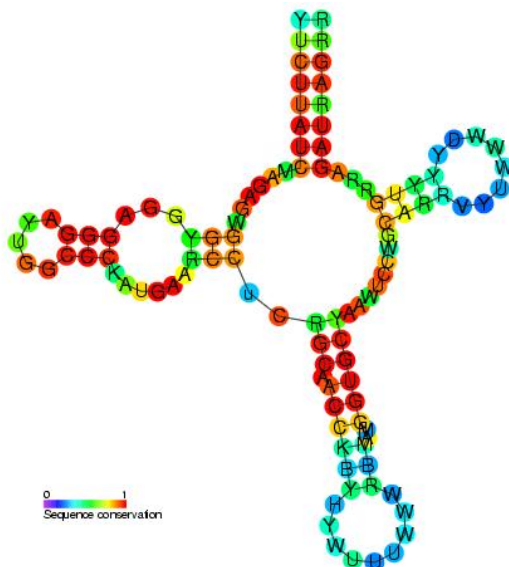
Više informacija o familiji može da se nađe na internet adresi: <http://rfam.janelia.org/cgi-bin/getdesc?name=TPP>.

2.1.3 *T-box leader riboswitch*

T-box riboswitch sekvence se nalaze uzvodno od gena za sintezu amino kiselina na transportnoj RNK, kao i gena za biosintezu aminokiselina u Gram-pozitivnim bakterijama. Više informacija o familiji može da se nađe na internet adresi: <http://rfam.janelia.org/cgi-bin/getdesc?name=T-box>.

2.1.4 *SAM riboswitch (S box leader)*

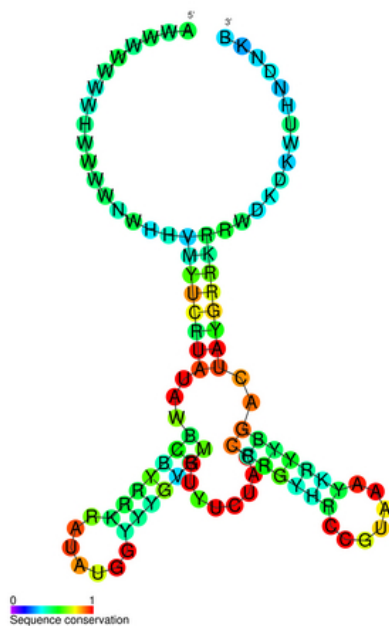
SAM riboswitch je pronađen uzvodno od brojnih gena koji kodiraju proteine uključene u biosintezu metionina kod Gram-pozitivnih bakterija. *SAM riboswitch* deluje na nivou završetka transkripcije (slika 2.3).

Slika 2.3: Sekundarna struktura *SAM riboswitch* sekvence

Više informacija o familiji može da se nađe na internet adresi: <http://rfam.janelia.org/cgi-bin/getdesc?name=SAM>.

2.1.5 *Purine riboswitch*

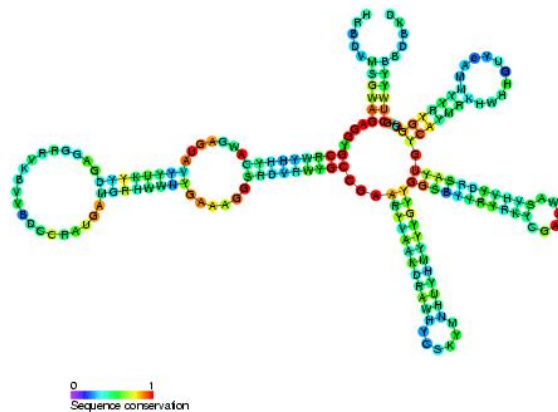
Purin riboswitch je familija *riboswitch*-eva koji selektivno prepoznaju guanin i postaju zasićeni na koncentracijama manjim od 5 nM (slika 2.4).

Slika 2.4: Sekundarna struktura *Purine riboswitch* sekvence

Više informacija o familiji može da se nađe na internet adresi: <http://rfam.janelia.org/cgi-bin/getdesc?name=Purine>.

2.1.6 *Lysine riboswitch*

Ova familija uključuje *riboswitch*-eve koji su osetljivi na lizin, i to u brojnim genima koji su uključeni u metabolizam lizina, uključujući *lysC* (slika 2.5).



Slika 2.5: Sekundarna struktura *Lysine riboswitch* sekvence

Više informacija o familiji može da se nađe na internet adresi: <http://rfam.janelia.org/cgi-bin/getdesc?name=Lysine>.

2.2 Pronalaženje *riboswitch* sekvenci

U nastavku rada je prikazano određivanje *riboswitch* sekvenci genoma *Lactobacillus paracasei subsp. paracasei* BGSJ2-8 koji se nalazi u vlasništvu Instituta za molekularnu genetiku i genetičko inženjerstvo Univerziteta u Beogradu. Određivanje bi bilo znatno jednostavnije da postoji asemliran kompletan genom. Kako za sada to nije slučaj, postupak pronalaženja *riboswitch* sekvenci je primenjen na skup od 207 kontiga. Celokupan skup kontiga je propušten kroz postojeće softverske alate:

- *Rfam*
- *RibEX*
- *Riboswitch finder* i
- *RiboSW*.

Kako bi se utvrdila njihova korektnost, potencijalne *riboswitch* sekvence su dalje analizirane u smislu zadovoljenosti uslova koji se odnose na biološke definicije *riboswitch* sekvenci.

Glava 3

Pronalaženje *riboswitch* sekvenci u genomu bakterije *Lactobacillus paracasei* subsp. *paracasei* BGSJ2-8

Jedan od najvažnijih resursa kada se radi sa familijama RNK je **Rfam** baza podataka. Sa verzijom 10.1, **Rfam** sadrži ukupno 1973 familije od kojih je 26 anotirano kao *riboswitch* familije. Na **Rfam** serveru postoji mogućnost lociranja pozicija *riboswitch* sekvenci, ali pošto implementirani algoritam zahteva mnogo izračunavanja, dužina ulazne sekvence je ograničena na maksimalno 10000 nukleotida.

RibEx je takođe veb server koji omogućava korisniku jednostavno pretraživanje mogućih *riboswitch* sekvenci u proizvoljnoj ulaznoj sekvenci[14]. Kako je većina *riboswitch*-eva povezana sa *attenuator*-ima, uključena je opcija pretraživanja transkripcionih i translatorskih *attenuator*-a koji pomažu u izboru najverovatnijih kandidata za *riboswitch* sekvence. Izlaz je predstavljen u obliku vizuelnog prikaza otvorenih okvira čitanja i njihovih odgovarajućih regulatornih elemenata. Svaki mogući *riboswitch* element je povezan sa listom gena za koju se predviđa da je predmet njegove regulacije.

Pored ova dva programa za lociranje *riboswitch* sekvenci najčešće se koriste i **RiboSW**, alat za pretragu *putative* familija *riboswitch*-eva u ulaznoj sekvenci, kao i **Riboswitch finder** koji u trenutnoj verziji podržava samo pretragu *purine-sensing* *riboswitch*-eva koji regulišu biosintezu purina i reaguju na prisustvo tj. odsustvo purina u ćeliji[12, 13].

3.1 Rfam baza podataka

Rfam baza podataka sadrži informacije o RNK familijama i anotacijama miliona RNK gena. Svaki slog u **Rfam** bazi podataka sadrži višestruka poravnanja sekvence, sekundarnu strukturu i modele verovatnoća poznatije kao kovarijanski modeli. **Rfam** projekat je pokrenut sa ciljem da na jednom mestu objedini RNK

molekule i učini ih dostupnim široj naučnoj zajednici. Za svaku familiju RNK, veb interfejs omogućava korisnicima pregled višestrukih poravnanja sekvence, čitanje anotacije i ispitivanje distribucije vrsta članova posmatrane familije. Korisnici putem veb sajta mogu da pristupe literaturi, kao i drugim RNK bazama podataka.

Molekuli RNK u bazi podataka mogu da se podele u tri kategorije:

- nekodirajući RNK geni,
- strukturni *cis*-regulatorni elementi i
- samospojivi molekuli RNK.

Svaka od ovih klasa RNK je u bazi podataka podeljena u familije koje dele zajedničkog pretka. **Rfam** omogućava korisnicima pretragu baze podataka po ključnim rečima, nazivu familije i genomu. Informacije čuvane u bazi su besplatne i dostupne za preuzimanje i snimanje na lokalnoj mašini.

Poslednja objavljena verzija **Rfam** baze podataka je 10.1 iz januara 2010. godine. Ona sadrži 1973 anotirane familije RNK. Baza podataka je dostupna na internet adresi: <http://rfam.janelia.org/>. U okviru sajta se nalazi server koji omogućava pretragu ulazne sekvence za poznatim RNK familijama. Grafički korisnički interfejs programa za pretragu RNK familija je prikazan na sledećoj slici. Veličina ulazne sekvence je ograničena na 10000 baznih parova.

Sequence search

Find Rfam families within your sequence of interest. Paste your nucleotide sequence into the box below, to have it searched for matching Rfam families. [More...](#)

Sequence

Submit Reset Example

You can also see the Rfam hits for a given sequence. Enter an EMBL sequence accession.

Look up sequence

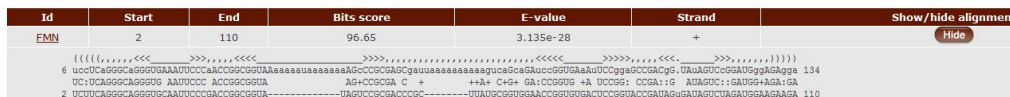
Submit Reset Example

Slika 3.1: Grafički korisnički interfejs programa za pretragu RNK familija u ulaznoj sekvenci na Rfam veb serveru

3.1.1 Primer

Na sledećoj slici je prikazan izlaz iz programa za pretragu RNK familija u ulaznoj sekvenci na **Rfam** veb serveru. Izlaz je predstavljen u tabelarnom obliku gde su izdvojeni detalji o svakom poklapanju sa postojećim RNK familijama. Pored informacija o početnoj i krajnjoj poziciji pronađene RNK familije u ulaznoj

sekvenci, dostupne su i procenat poklapanja, lanac na kome je pronađeno poklapanje, kao i komplementarne palindromske sekvence unutar pronađene familije.



Id	Start	End	Bits score	E-value	Strand	Show/hide alignment
FMN	2	110	96.65	3.135e-28	+	Hide

Slika 3.2: Izlaz iz programa za pretragu RNK familija u ulaznoj sekvenci na Rfam veb serveru

3.1.2 Rezultati

Uz pomoć **Rfam** baze podataka i implementiranog programa za pretragu RNK familija, u 12 kontiga je pronađena 21 moguća *riboswitch* sekvenca.

Familije <i>riboswitch</i> sekvenci	Broj pronađenih sekvenci
<i>FMN riboswitch (RFN element)</i>	1
<i>TPP riboswitch (THI element)</i>	2
<i>T-box (T-box leader)</i>	15
<i>SAM riboswitch (S box leader)</i>	0
<i>Purine riboswitch</i>	0
<i>Lysine riboswitch</i>	3

Tabela 3.1: Broj pronađenih *riboswitch* sekvenci po familijama uz pomoć Rfam baze podataka

3.2 RibEx: veb server za određivanje pozicija *riboswitch* sekvenci i drugih konzerviranih bakterijskih regulatornih elemenata

U ovom odeljku će detaljnije biti predstavljen **RibEx** (*riboswitch explorer*), veb server koji može da pretražuje poznate *riboswitch* sekvence u proizvoljnoj sekvenci, kao i druge visoko konzervirane bakterijske regulatorne elemente. On omogućava vizuelni pregled identifikovanih motiva u odnosu na *attenuator*-e i otvorene okvire čitanja (*ORF*). Svaki *ORF* ili regulatorni element može da se dobije klikom i prosledi **BLAST** programu na *NCBI* serveru. **RibEx** je dostupan na internet adresi <http://132.248.32.45/cgi-bin/ribex.cgi>.

3.2.1 Veb server

Server je podeljen na nekoliko modula napisanih u **PERL** programskom jeziku. U ovom odeljku biće predstavljen kratak opis svakog od modula.

- *Mogući riboswitch elementi.* Program uzima ulaznu sekvencu i deli je u preklapajuće prozore dužina 500 nukleotida. Dalje se svaka od novoformiranih manjih sekvenci pretražuje **MAST**¹ programom u cilju pronalaženja mogućih *riboswitch* elemenata. Svaki od mogućih *riboswitch* elemenata se definiše kao skup nekoliko nepreklapajućih motiva. Ako mogući *riboswitch* elementi prođu zadati *E-value* uslov, onda se za svaki od njih vizuelno prikazuju pozicija i veličina svakog od motiva koji ga grade.
- *Otvoreni okviri čitanja.* Ovaj modul predviđa otvorene okvire čitanja. Podrazumevane vrednosti su za rezultujući protein dužine najmanje 80 amino kiselina, koji počinje *start* kodonom (ATG, GTG ili TTG), a završava se *stop* kodonom (TAA, TAG ili TGA). Ne prikazuju se otvoreni okviri čitanja koji se u potpunosti preklapaju.
- *Attenuator-i.* Algoritmi za pronalaženje *attenuator*-a neće biti predstavljeni u ovom radu. Oni predviđaju sekundarnu strukturu svakog *attenuator*-a koja može biti prikazana u novom prozoru.
- *Veb izlaz.* Izlaz iz programa je prikazan veb stranom generisanom **PERL** skriptom koja kontroliše preostale module. Interaktivnost na web strani omogućena je *JavaScript*-om.

Grafički korisnički interfejs **RibEX** veb servera je prikazan na sledećoj slici.

Slika 3.3: Grafički korisnički interfejs RibEX veb servera

¹**MAST** server pretražuje proteinsku bazu podataka u cilju pronalaženja najboljih pogodaka za skup motiva ili blokova bez raskoraka.

3.2.2 Primer

Na slici 3.4 je prikazan tipičan **RibEx** izlaz. Ulazna sekvenca dužine 28406 nukleotida predstavlja kontigu 192 genoma bakterije *Lactobacillus paracasei subsp. paracasei* BGSJ2-8. Neposredno iznad jednog od otvorenih okvira čitanja (predstavljenih plavom strelicom) mogu da se vide tri motiva koji čine **T-box riboswitch**, kao i transkripcioni *attenuator*-i (predstavljeni crnom figurom).



Slika 3.4: RibEx pronalazi T-box *riboswitch*

[Legenda: plava strelica - predviđeni ORF na jednom lancu, narandžasta strelica - predviđeni ORF na naspramnom lancu, lizalice - transkripcioni ili translatorski *attenuator*-i, crveni kvadrati - motivi T-box *riboswitch* sekvence]

3.2.3 Rezultati

Uz pomoć **RibEX** programa za pretragu RNK familija, u 15 kontiga su pronađene 24 moguće *riboswitch* sekvence.

Familije <i>riboswitch</i> sekvenci	Broj pronađenih sekvenci
<i>FMN riboswitch (RFN element)</i>	1
<i>TPP riboswitch (THI element)</i>	1
<i>T-box (T-box leader)</i>	22
<i>SAM riboswitch (S box leader)</i>	0
<i>Purine riboswitch</i>	0
<i>Lysine riboswitch</i>	0

Tabela 3.2: Broj pronađenih *riboswitch* sekvenci po familijama uz pomoć RibEX programa

3.3 RiboSW: alat za pretraživanje *riboswitch* sekvenci

RiboSW je alat za pretraživanje *putative riboswitch*-eva u sekvenci. Program je razvijen kako bi omogućio pretragu 12 familija *riboswitch*-eva.

Implementiran je i dostupan u obliku veb servera na internet adresi <http://ribosw.mbc.nctu.edu.tw/index.html>. Preciznost algoritma pretrage je uporediva sa algoritmima drugih metoda koji su opisani u ovoj glavi. Na slici 3.5 je prikazan veb interfejs **RiboSW** veb servera.

RiboSW - an effective tool for searching riboswitches [Run webserver](#) [About](#)

Citing RiboSW
T.H. Chang, L.C Wu, C.T. Yeh, B.J. Liu, H.D. Huang and J.T. Horng. (2009) "Computational Identification of Riboswitches Based on RNA Conserved Functional Sequences and Conformations ", RNA, 15: 1426-1430.

► **Input sequences** (fasta format; up to 10k bps):

or upload from

► **Selecte riboswitch models to search:** (Note: the web version only provides strict search)

<input checked="" type="checkbox"/> Purine	<input type="checkbox"/> TPP	<input type="checkbox"/> Cobalamin	<input type="checkbox"/> SAM alpha
<input type="checkbox"/> Lysine	<input type="checkbox"/> Glycine	<input type="checkbox"/> yybP-ykoY	<input type="checkbox"/> PreQ1
<input type="checkbox"/> FMN	<input type="checkbox"/> SAM	<input type="checkbox"/> ykkC-yxkD	<input type="checkbox"/> glmS

Department of Computer Science and Information Engineering, National Central University, Tao-yuan, Taiwan. Contact with site administrator: [Tzu-Hao Chang](mailto:Tzu-Hao.Chang@nctu.edu.tw)

Slika 3.5: RiboSW veb interfejs

3.3.1 Primer

Na slici 3.6 je prikazan izlaz iz **RiboSW** programa za pretragu RNK familija u ulaznoj sekvenci. Izlaz sadrži informacije o početnoj poziciji pronađene RNK familije u ulaznoj sekvenci, kao i komplementarnim palindromskim sekvencama unutar pronađene familije.

Search Results:

► Purine:

No detection of any putative Purine riboswitch

► SAM:

No detection of any putative SAM riboswitch

► Glycine:

No detection of any putative Glycine riboswitch

► Lysine:

No detection of any putative Lysine riboswitch

► FMN:

```
>contig00051|FMN|Start:1|Free energy: 1.31|HMM E-value:1.1e-14
gucuuacagggcagggugcaauucccgaccggcgguauaguccgaccgcuuauagcggggaaccggugugacuccgguaccgauagugauagucuagaggaagaagau
((((((.....(((.....))).....(((.....))).....(((.....))).....))))))
```

► TPP:

No detection of any putative TPP riboswitch

Slika 3.6: Izlaz iz RiboSW programa

3.3.2 Rezultati

Uz pomoć **RiboSW** programa za pretragu RNK familija, u 3 kontige su pronađene 3 moguće *riboswitch* sekvence.

Familije <i>riboswitch</i> sekvenci	Broj pronađenih sekvenci
<i>FMN riboswitch (RFN element)</i>	1
<i>TPP riboswitch (THI element)</i>	0
<i>T-box (T-box leader)</i>	0
<i>SAM riboswitch (S box leader)</i>	0
<i>Purine riboswitch</i>	0
<i>Lysine riboswitch</i>	2

Tabela 3.3: Broj pronađenih *riboswitch* sekvenci po familijama uz pomoć RiboSW programa

3.4 Uporedna analiza rezultata

Propuštanjem svih 207 kontiga kroz **RibEx** pronađeno je 24 moguće *riboswitch* sekvence, dok je **Rfam** pokazao lošije rezultate. **RiboSW** je pronašao najmanji broj mogućih *riboswitch* sekvenci što se objašnjava nemogućnosti pretrage za *T-box riboswitch* sekvencama kojih, na osnovu prethodna dva programa, ima najviše u skupu kontiga posmatrane bakterije. Uz pomoć **Riboswitch finder** programa nije pronađen ni jedan *riboswitch* zbog specifičnosti *riboswitch* sekvenci

koje ova metoda može da odredi.

Familije <i>riboswitch</i> sekvenci	RibEx	Rfam	RiboSW	Ukupno
<i>FMN riboswitch (RFN element)</i>	1	1	1	1
<i>TPP riboswitch (THI element)</i>	1	2	0	2
<i>T-box (T-box leader)</i>	22	15	0	22
<i>SAM riboswitch (S box leader)</i>	0	0	0	0
<i>Purine riboswitch</i>	0	0	0	0
<i>Lysine riboswitch</i>	0	3	2	3
Ukupan broj različitih mogućih <i>riboswitch</i> sekvenci				28



Tabela 3.4: Broj pronađenih *riboswitch* sekvenci po familijama koristeći RibEx, Rfam i RiboSW

Veći broj mogućih kandidatskih sekvenci koji je dobijen **RibEx** programom može da se objasni time što **RibEx** omogućava pretragu na većim sekvencama, a pruža i bolji pregled regulatornog potencijala same sekvence, prikazujući *ORF*-ove i predviđene *attenuator*-e. Dobijene sekvence predstavljaju samo potencijalne *riboswitch*-eve. Da bi se videlo da li su i zaista u pitanju *riboswitch*-evi potrebno je da se ispita i ispunjenost dodatnih uslova koje svaka *riboswitch* sekvenca zadovoljava.

Glava 4

Analiza rezultata i ispitivanje korektnosti potencijalnih *riboswitch* sekvenci

Propuštanjem svih 207 kontiga genoma *Lactobacillus paracasei subsp. paracasei* BGSJ2-8 dobijeno je 28 potencijalnih *riboswitch* sekvenci iz ukupno 4 familije. One su sačuvane u tabeli relacione baze podataka čija struktura izgleda ovako:

Key	Name	Data type	Length	Nullable
	FAMILIJA	VARCHAR	20	No
	ID_CONS	SMALLINT	2	No
	KOD	VARCHAR	200	Yes

Slika 4.1: Struktura tabele RIBOSVICEVI u lokalnoj bazi podataka



Kako bi se utvrdilo da pronađene sekvence zaključavaju gene koje regulišu i tako sprečavaju njihovu ekspresiju u momentu kada ona nije poželjna, sve potencijalne *riboswitch* sekvence analizirane su u smislu ispitivanja zadovoljenosti sledeća četiri uslova:

- *Riboswitch* sekvenca se nalazi između dva ORF-a ili ulazi u prethodni uzvodni ORF;
- *Riboswitch* sekvenca se nalazi između početka jednog ORF-a i kraja drugog ili početka jednog ORF-a i početka drugog;
- *Riboswitch* sekvenca reguliše gene nizvodno od nje u sličnim organizmima iz roda *Lactobacillus*;
- *Riboswitch* sekvenca je palindromska sekvenca.

4.1 Otvoreni okviri čitanja

Podaci o pozicijama otvorenih okvira čitanja u pojedinačnim kontigama su dobijeni od strane saradnika sa IMGGI. Oni su iskorišćeni za punjenje tabele u

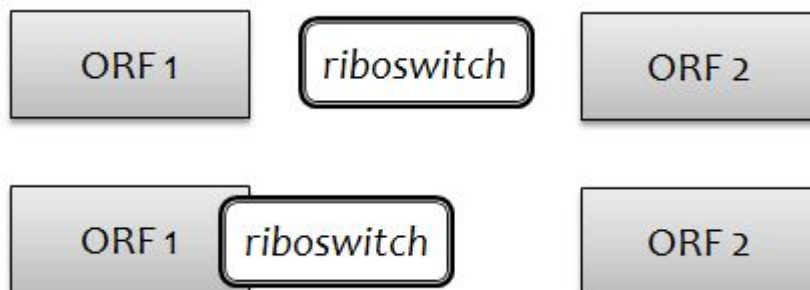
relacionoj bazi podataka čija struktura izgleda ovako:

Key	Name	Data type	Length	Nullable
	KONTIGA	CHARACTER	11	No
	ORF	VARCHAR	20	No
	POCETAK	SMALLINT	2	Yes
	DUZINA	SMALLINT	2	Yes
	KOD	VARCHAR	8200	Yes
	KODR	VARCHAR	8200	Yes
	ORIJENTACIJA	CHARACTER	3	Yes
	POCETAK_U_KONTIGI	INTEGER	4	Yes
	KRAJ_U_KONTIGI	INTEGER	4	Yes

Slika 4.2: Struktura tabele ORFOVI u lokalnoj bazi podataka

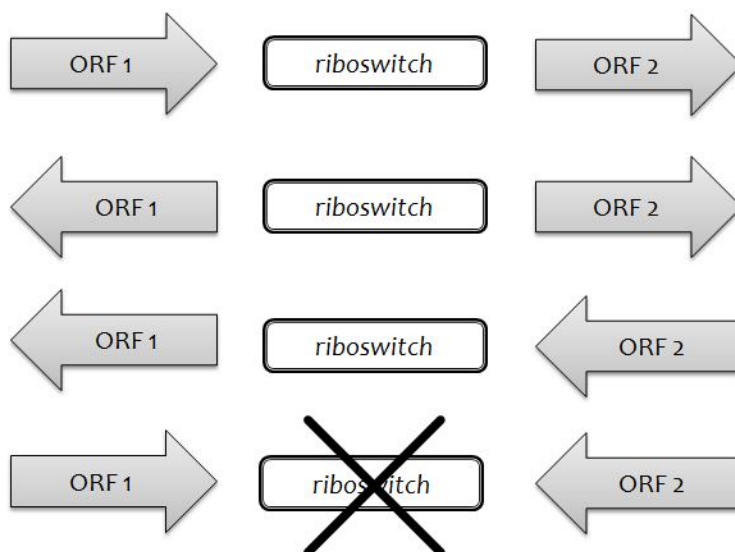
Na osnovu izdvojenih informacija o početnoj i krajnjoj poziciji potencijalnih *riboswitch* sekvenci u kontigama kojima pripadaju ,kao i pozicija otvorenih okvira čitanja, dalje su analizirani položaji i orijentacije otvorenih okvira čitanja koji okružuju potencijalne *riboswitch* sekvence.

Prvi uslov koji potencijalna *riboswitch* sekvenca treba da zadovolji jeste da se nalazi između dva ORF-a ili da ulazi u prethodni uzvodni ORF. Nijedna druga mogućnost nije dozvoljena.



Slika 4.3: Dozvoljeni položaj ORF-ova u odnosu na *riboswitch* sekvencu

Drugi uslov je da orijentacija ORF-ova koji okružuju potencijalnu *riboswitch* sekvencu zadovoljava kriterijume prikazane na sledećoj slici.



Slika 4.4: Dozvoljene orijentacije okolnih ORF-ova

Kod dve kandidatske *riboswitch* sekvence uslov položaja *riboswitch* sekvence u odnosu na okolne ORF-ove nije zadovoljen, zbog čega su isključene iz daljeg razmatranja. Za svaku preostalu *riboswitch* sekvencu je dalje utvrđivano postojanje komplementarnih palindroma.

4.2 Ispitivanje postojanja palindroma u potencijalnim *riboswitch* sekvencama

Riboswitch sekvenca treba da sadrži dva palindroma koja naizmenično formiraju petlje u zavisnosti od fiziološkog stanja ćelije. Ti palindromi mogu da se preklapaju i najčešće se i preklapaju, tako da kada se formira jedna petlja ne može da se formira druga. Takve petlje se nazivaju *in vivo*. Što se tiče dužine, mogu biti različite, ali što je sekvenca palindroma duža to je veća verovatnoća da se petlja formira tj. slobodna energija koja se oslobađa prilikom formiranja palindroma je veća. U skladu sa tim je uz pomoć postojećeg programa napisanog za ovu namenu ispitivano postojanje palindromskih sekvenci u pronađenim *riboswitch* sekvencama.



Slika 4.5: Primer postojanja komplementarnog palindroma u mogućoj *riboswitch* sekvenci

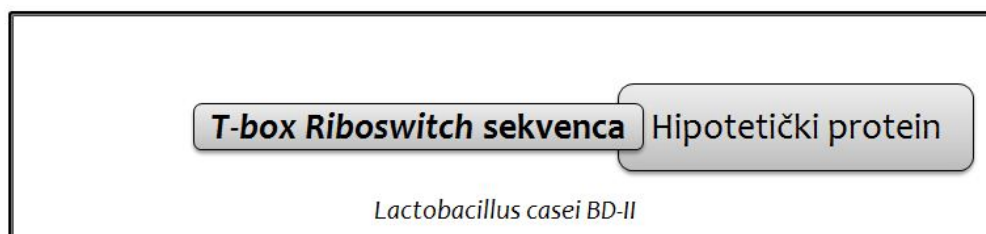
Sve moguće *riboswitch* sekvence zadovoljavaju uslov postojanja komplementarnih palindroma. U narednom koraku sekvence prolaze kroz proces blastovanja u cilju određivanja nizvodnog gena koji *riboswitch* reguliše u sličnim organizmima iz roda *Lactobacillus*.

4.3 Blastovanje mogućih *riboswitch* sekvenci

BLAST je program koji nepoznatu ulaznu sekvencu upoređuje, tj. poravnava sa svakom sekvencom koja se nalazi u bazi podataka DNK sekvenci ili proteina. Kao rezultat vraća listu sekvenci koje su najbližije ulaznoj. Dostupan je na internet adresi: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>, i to u dve verzije:

- nukleotidni blast (**blastn**) i
- proteinski blast (**blastp**).

Koristeći **BLAST**, pronađene su DNK sekvence koje su najbližije svakoj mogućoj *riboswitch* sekvenci. Na osnovu izdvojenih najbližijih organizama iz roda *Lactobacillus*, određeni su nizvodni geni koji su regulisani od strane *riboswitch* sekvence u pomenutim organizmima. Na ovaj način je zbog ulaska u gen koji bi trebalo da reguliše utvrđena nekorektnost još jedne *riboswitch* sekvence (slika 4.6), čime je skup mogućih *riboswitch* sekvenci sužen na 25.

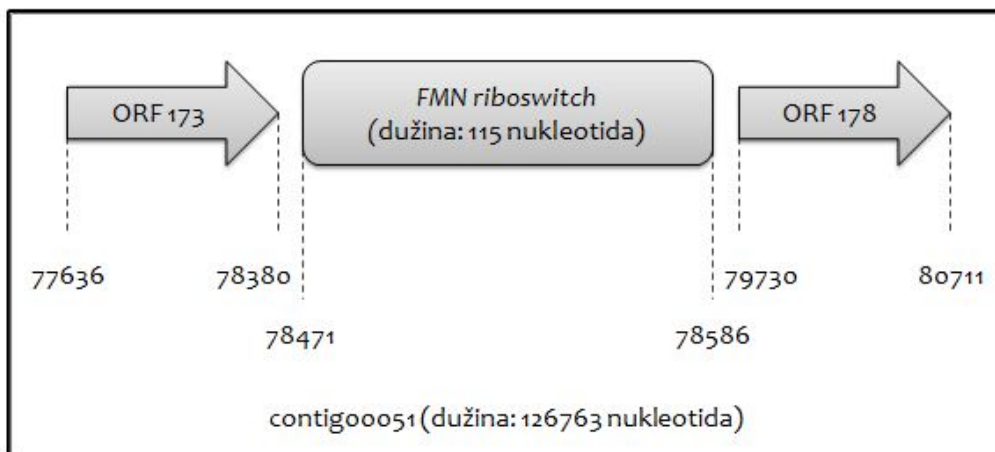


Slika 4.6: Primer nekorektne *riboswitch* sekvence

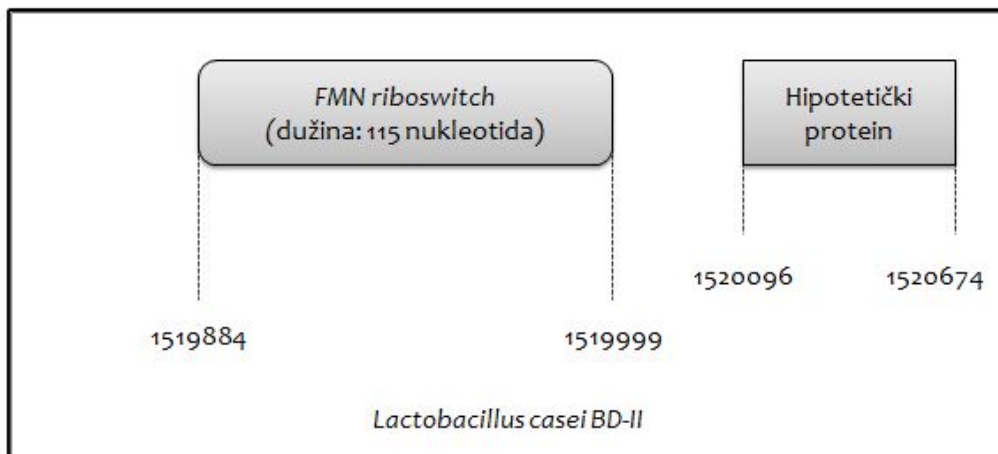
Sve *riboswitch* sekvence koje su prošle navedene uslove poslate su IMGGI Univerziteta u Beogradu na dalju eksperimentalnu proveru.

4.4 Primer korektne *riboswitch* sekvence

Jedna od *riboswitch* sekvenci koja je zadovoljiva sve definisane uslove pripada familiji *FMN*. Njena dužina je 115 nukleotida, a pronađena je u kontigi 51. Okružena je ORF-ovima 173 i 178, čiji su krajevi pravilno orijentisani (slika 4.7).



Slika 4.7: Primer dobro orijentisanih ORF-ova kod *riboswitch* sekvence



Slika 4.8: Primer korektne regulacije nizvodnog gena kod *riboswitch* sekvence u sličnom organizmu

Propuštanjem kroz **blast**, *riboswitch* sekvencija je pokazala 100% poklapanja sa sledećim organizmima:

- *Lactobacillus casei* BD-II, complete genome
- *Lactobacillus casei* LC2W, complete genome
- *Lactobacillus casei* str. Zhang, complete genome

- *Lactobacillus casei* BL23 complete genome, strain BL23 i
- *Lactobacillus casei* ATCC 334, complete genome.

U svakom od sličnih organizama iz roda *Lactobacillus* je utvrđeno da sekvenca reguliše sintezu hipotetičkog proteina koji se nalazi nizvodno od nje (slika 4.8).

U dodatku su navedne karakteristike svih *riboswitch* sekvenci koje su prošle sve navedene provjere.

Glava 5

Zaključak

U ovom radu je prikazan postupak pronalaženja *riboswitch* sekvenci u skupu od 207 kontiga genoma *Lactobacillus Paracasei subsp. Paracasei BGSJ2-8* koji se nalazi u vlasništvu IMGGI Univerziteta u Beogradu. Za potrebe rada je razmatrano 6 familija *riboswitch* sekvenci od ukupno 22 anotirane familije dostupne na <http://rfam.janelia.org/>.

Celokupan skup kontiga je propušten kroz postojeće softverske alate, uz pomoć kojih je pronađeno ukupno 28 različitih potencijalnih *riboswitch* sekvenci. Kako bi se utvrdila njihova korektnost, potencijalne *riboswitch* sekvence su dalje analizirane u smislu zadovoljenosti uslova koji se odnose na biološke definicije *riboswitch* sekvenci.

Za svaku potencijalnu *riboswitch* sekvencu je određena početna i krajnja pozicija u kontigi kojoj pripada, kako bi se odredila njena udaljenost od okolnih ORF-ova. Nakon toga je za svaku od njih izvršena provera zadovoljenosti uslova koji se odnosi na položaj ORF-ova u odnosu na *riboswitch* sekvencu.

Nakon provere svih uslova preostalo je 25 *riboswitch* sekvenci koje su prosleđene IMGGI na eksperimentalnu proveru.

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: *Computational 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] E. Nudler, A. S. Mironov: *The riboswitch control of bacterial metabolism*, Trends Biochem Sci, 29 (1): 117, 2004.
- [11] Richard Deonier, Simon Tavaré, Michael Waterman: *Computational Genome Analysis: an introduction*, Springer-Verlag, ISBN 0387987851, 2005.
- [12] T. H. Chang, L. C. Wu, C. T. Yeh, B. J. Liu, H. D. Huang and J. T. Horng: *Computational Identification of Riboswitches Based on RNA Conserved Functional Sequences and Conformations*, RNA, 15: 1426-1430, 2009.
- [13] Peter Bengert, Thomas Dandekar: *Riboswitch finder- a tool for identification of riboswitch RNAs*, Nucleic Acids Research, Vol. 32, 2004.

- [14] Cei Abreu-Goodger, Enrique Merino: *RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements*, Nucleic Acids Research, Vol. 33, 2005.

Glava 6

Dodatak

6.1 Pronađene *riboswitch* sekvence u genomu bakterije *Lactobacillus paracasei* subsp. *paracasei* BGSJ2-8

riboswitch sekvenca 1

Familija: FMN

Kod:

```
TTCAGGGCAGGGTGCAATTCCTCCGACCGGCGGTATAGTCCGCGACCCGCTTATGCGGTGGAACCGGTGTGA  
CTCCGGTACCGATAGTGATAGTCTAGATGGAAGAAGATGGGG
```

Kontiga: 51

Početna pozicija u kontigi: 78471

Krajnja pozicija u kontigi: 78586

Udaljenost od levog ORF-a: 91

Udaljenost od desnog ORF-a: 1144

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvenca 2

Familija: TPP

Kod:

```
AGAAACATTTGGGGTGCTGTAAAGCTGAGATGATACCCATGGAACCTGATGCAGTTAGGACTGCCGCAGG  
GAAATGT
```

Kontiga: 36

Početna pozicija u kontigi: 26062

Krajnja pozicija u kontigi: 26139

Udaljenost od levog ORF-a: 1080

Udaljenost od desnog ORF-a: 741

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvenca 3

Familija: T-box

Kod:

ATTCATCCATGAAACCGATGATGTGGAGATCAAGATCATTGTGCGCGCAAAGAGAGTTGCCGGTGCTGAG
AGTGCAATCGAACGCAGATGATTAATGGACCACGGAGGGTTCTCCGAAAAGACTTCAAGTACGGAGAAA
CGTAAAGGCATACGTCAGTTGTGCGGAGAGATGTTGGTCTCTCGCTAAGCGTGATTAATGGGGATTCTTA
GACAAATCAGATGCTATTATTTGTTGAAGTCTCGAATTAATCGAACTAAGGTGGCACCGCGGAAAATCCG
TCCTTAAC

Kontiga: 193

Početna pozicija u kontigi: 23574

Krajnja pozicija u kontigi: 23862

Udaljenost od levog ORF-a: 202

Udaljenost od desnog ORF-a: 188

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvenca 4

Familija: T-box

Kod:

CGTCCATGAACAAACCGATGATGTGGAGATCAAGATCATTGTGCGCGCAAAGAGAGTTGCCGGTGCTGAG
AGTGCAATCGAACGCAGATGATTAATGGACCACGGAGGGTTCTCCGAAAAACTTCAAGTACGGAGAAA
CGTAAAGGCATACGTCAGTTGTGCGGAGAGATGTTGGTCTCTCGCTAAGCGTGATTAATGGGGATTTTA
GACAAATCAGATGCTATTATTTGTTGAAGTCTCGAATTAATCGAACTAAGGTGGCACCGCGGAAAAATCC
GTCCTTAA

Kontiga: 193

Početna pozicija u kontigi: 20872

Krajnja pozicija u kontigi: 21160

Udaljenost od levog ORF-a: 105

Udaljenost od desnog ORF-a: 928

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvenca 5

Familija: T-box

Kod:

CCACTTAATTTCAAACAAAAAGCCGCCCTATCTACATTAGGACGGCTCAGCCCGGTACCACCTAAGTTC
TGCTTTACAGCGCTCAAAGGAGATAACGGTCCGGCCGGCTGTGGCTACTAATTTACCCCAGCAACTTC
AAAGGTCAGTTCGGTCTTTGCTTCGTCAGACTCGCACCACCGCTGCTCTCTGTCACCC

Kontiga: 192

Početna pozicija u kontigi: 24084

Krajnja pozicija u kontigi: 24283

Udaljenost od levog ORF-a: 11

Udaljenost od desnog ORF-a: 368

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvence 6

Familija: T-box

Kod:

```
GGTCTCCAGAGAGTCGGCAATGCTGTGACGTCGATGACCATGGTAGGCAAAGTTGGGTCGTTTGC GTTAT  
GTTCCGGTGGGTGCGCCTTAAGCCAAGAGAGGTCCGCAAGCGGACAAGTTAGGTGGTACCGCGCGCTTGC  
GTCCTATGTTGTTATGACATAGGGCGCTTTTTTTtACGATAACGCGCTC
```

Kontiga: 129

Početa pozicija u kontigi: 67270

Krajnja pozicija u kontigi: 67458

Udaljenost od levog ORF-a: 209

Udaljenost od desnog ORF-a: 160

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvence 7

Familija: T-box

Kod:

```
CTGACTCGCGCTCACAAAAAaGCCGTCCATAACAAACATCTTGTTAAGGACGGTTTCCCGCGGTGCCAC  
CTTAGTTTGATCAAAGAGACTGCGCCTGTGTGATGGTAGCATCTGGCAGTGTGATCAGCCTTAGCCCG  
GGGCAACACCCCGCGCAACTGACGTATGTCAACACGTTTCCCGTACTCGaAAAACTTTCGgGGGAA  
CCCTCAGTGGTCCATTTACTGACCTGCGTTTCGGCCGACTCGCAGCATCGCCGGCTCTCTGTGCGTGCAC  
AATCAGCTTGATCTCCACTTCAAAGGTTTGCATGAACGGT
```

Kontiga: 125

Početa pozicija u kontigi: 19881

Krajnja pozicija u kontigi: 20201

Udaljenost od levog ORF-a: 103

Udaljenost od desnog ORF-a: -1

Sekvenca nije zadovoljila uslov koji se odnosi na poziciju sekvence u odnosu na okolne ORF-ove.

riboswitch sekvence 8

Familija: T-box

Kod:

```
TCATATTGACTTATGCGTTGAAGAGAAGAGTAAGCGTGAATCCGTTTAAAGAGACTGCCCGGAGGTGTGA  
GGGTGGAACGGTGAAGCACTGAAGATGAGCTCTGAGCATCGGCAACCGTGAAAGCGGGCGCCGCTCAGG  
ACTCGATACCGTCCCAGACATCGTTTTGACGGTGTGTTGAGTTGGCGGGAGCAATCCCAGCCAAGAATT  
AGGGTGGTACCACGAAGGTCTTCGTCCCTGTATTAACATACAGGGGCGAAGATCTTTTTTtGATTAAC  
AGGCAAAG
```

Kontiga: 103

Početna pozicija u kontigi: 118249

Krajnja pozicija u kontigi: 118537

Udaljenost od levog ORF-a: 127

Udaljenost od desnog ORF-a: -18

Sekvenca nije zadovoljila uslov koji se odnosi na poziciju sekvence u odnosu na okolne ORF-ove.

riboswitch sekvenca 9

Familija: T-box

Kod:

```
AACCTTGCGCTGCCACAAAAACGCCcTTAACTACTTTTGTGTAGTCAAGGGGCGACAGTGTGCGGTT
ACCACCCTTATTCACGATGTTGATACATCGTCTCCAGTACCTGTTGATACCTGACAGCATAACGGCCGTA
ACCGACTCGTGAGCGTTAACCGCACGAATAACTATTTCGAAGCTCATCTTCACCTTCTTGCAGATGACCCT
TTTCCATCAACCGGGCTTCTACAGCGTCTGCGCCAAGGTTACTCTTCTTCTCAAAGTGTGTTTTtATA
AA
```

Kontiga: 103

Početna pozicija u kontigi: 90306

Krajnja pozicija u kontigi: 90588

Udaljenost od levog ORF-a: 56

Udaljenost od desnog ORF-a: 197

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvenca 10

Familija: T-box

Kod:

```
AAATCTCAGTGACACAAAAaGCCTTCCATCTCGTTGGGACGAAAAGGCCATTCCGCGGTGCCACCCAAGT
TTGACGCCTATGCGCCACCCTCAATAACTGATAACGGTCAGTGGCCGTGCAGTTCATCACTACCCACTAA
GGCGCTGGAATATGGATGGCTAACAGCTTTCACCAACCGCTGTCTCTCTTGATCAT
```

Kontiga: 97

Početna pozicija u kontigi: 1550

Krajnja pozicija u kontigi: 1746

Udaljenost od levog ORF-a: 34

Udaljenost od desnog ORF-a: 141

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvenca 11

Familija: T-box

Kod:

GCTGACTTACGAGAATAAAAAAaCcGCCTAaTTAAAAAATTaGGACGAGTCAGCTCGCGGTACCACCTAA
aTtGGGAAACAATCCCCACTTGGTTCAGGATAACGGCTGTCACCGGCCGCTCCCTACTCACTTCAGGAC
GGCAGCTCACAAGCTACAACCATAGCCTTTTTTCAAGGTTTGCACCAACCACCTTGTGCGCTGGAAAAAAaG
ACCATGACCCTCTTGATCATTGCATTAGCTGAATTAT

Kontiga: 96

Početna pozicija u kontigi: 44772

Krajnja pozicija u kontigi: 45019

Udaljenost od levog ORF-a: 49

Udaljenost od desnog ORF-a: 74

Sekvenca je zadovoljila sve definisane uslove.

***riboswitch* sekvenca 12**

Familija: T-box

Kod:

GCAAGAGGACGCTTTTTGCGTGGTACCACCTCAGTTTGCTGATTGTTTACACAATCAGCCTCACGTGGTTG
AAAAATCATGTCCAAGAAAGATGTCCGGTAAACAAAAAGCCGTCCTCCTACAGCAAATTGCTGTAAGAGG
ACGGCTTTTACCGCGGTTCCACCTCAGATTTGTTAACTGCTCACACAGCTAACCTCATGTAGTTATCTCA
ACCACGGCGATAACGAGCCCAACCGAGACGACCTACTTCGTTCAATCGTCCACTCGTGGATGTGTGATCA
ATGGATGTCTATCCGGTTCCTATCATCCGGACTCGCTTCAAGAC

Kontiga: 96

Početna pozicija u kontigi: 34408

Krajnja pozicija u kontigi: 34733

Udaljenost od levog ORF-a: 218

Udaljenost od desnog ORF-a: 122

Sekvenca je zadovoljila sve definisane uslove.

***riboswitch* sekvenca 13**

Familija: T-box

Kod:

CTTTATTAATTGAAATCAAAAAAGCCCGCCCTAATATTAGGACGAGCAAACGTGGTACCACCTAACTTCA
GGG

Kontiga: 96

Početna pozicija u kontigi: 15729

Krajnja pozicija u kontigi: 15802

Udaljenost od levog ORF-a: 35

Udaljenost od desnog ORF-a: 162

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvenca 14

Familija: T-box

Kod:

```
TATTACTAATACCCTCGGTGATTGCCTTAACTGTGGTGCCATCAGGAAACTGTTTGACGTTACCATCAGG
AAATGTAATATCTAATGCCATGAGTAAGACTCCTTTATTGTTTTtCGAAGAAACCAAAAAaTCCCTGAC
ATCTCTTAAGAGACATCAGGGACGATCAGTGATCGCGGTCCACCCAGTTTGCCAA
```

Kontiga: 53

Početna pozicija u kontigi: 149478

Krajnja pozicija u kontigi: 149674

Udaljenost od levog ORF-a: -142

Udaljenost od desnog ORF-a: 423

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvenca 15

Familija: T-box

Kod:

```
ATGCATTTGAAGGCATAAAAAaGTCCCTCTCGAAAAAGGGACGAAGCATCGCGGTACCACCCTTGTTC
TGG
```

Kontiga: 53

Početna pozicija u kontigi: 124387

Krajnja pozicija u kontigi: 124460

Udaljenost od levog ORF-a: -71

Udaljenost od desnog ORF-a: 1035

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvenca 16

Familija: T-box

Kod:

```
CGTCCTTTTTCGCTCTAAAAAAaTACACCCCTCGATTGCTCGAAGGGTTCGGACACGGTACCACCTTGTT
TTTTtCTTTGCCATAACGGTGGCGTTCCGGAACCTGTTACCTCAGAAGTGTCTTTTCGATCGTTCCCGG
CTTTCACCATCTCCGGGTCGCTTAAAGTG
```

Kontiga: 48

Početna pozicija u kontigi: 114090

Krajnja pozicija u kontigi: 114259

Udaljenost od levog ORF-a: 37

Udaljenost od desnog ORF-a: 235

Sekvenca je zadovoljila sve definisane uslove.

***riboswitch* sekvenca 17**

Familija: *T-box*

Kod:

ATTGGCTTCGGGACaCGAAAAAGTCCCTACGCCAATTAATGGCGTAGGGACGCTTATGCGCGGTTCCACC
CTACTTCAGATAGAAAAAAaCTACCTGCGGCTTCATCATTGGTTTCGAAAACGCCAATCACAACAACCAT
TATCAGGCTTACTTTCCCCGACTCGCTTCAATAG

Kontiga: 48

Početna pozicija u kontigi: 114090

Krajnja pozicija u kontigi: 100794

Udaljenost od levog ORF-a: 159

Udaljenost od desnog ORF-a: 1462

Sekvenca je zadovoljila sve definisane uslove.

***riboswitch* sekvenca 18**

Familija: *T-box*

Kod:

aTCGCCCTCGAAAaCGAAAAaCGTTTCCGAGGGCGATTTTtGACCGCGGTACCACCTCAATTTTGTTTA
GCTCTTGCGACCTAACCTCACcTGCCAAGTAACTGTCAGCAATTGCGAACAGCAGTTGGCAGCAGTGAT
AACGCAACGGATGCGTGACAGCTTACTTAGTTTCAGCCGTCAAGTTCATAGATGTGTGTTTCAGCTGGCTAA
TCGCGTCTTTTCAGCAATAGACGCTCTCTGTTGATCAGGATCAGCTTACTTTtCTAATCACAACCTTCAT
GTTTGTTAA

Kontiga: 48

Početna pozicija u kontigi: 18441

Krajnja pozicija u kontigi: 18730

Udaljenost od levog ORF-a: 99

Udaljenost od desnog ORF-a: 181

Sekvenca je zadovoljila sve definisane uslove.

***riboswitch* sekvenca 19**

Familija: *T-box*

Kod:

CTCGCGCTCACAAAaTAAAAaGCACCATTGCTGCCAAGGGCGCTTGACGCGGTACCACCTTGTTGCAA
T

Kontiga: 44

Početna pozicija u kontigi: 42660

Krajnja pozicija u kontigi: 42731

Udaljenost od levog ORF-a: 140

Udaljenost od desnog ORF-a: 232

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvenca 20

Familija: T-box

Kod:

AGGTCTAAGAAAAaCGGTGGTTGCTGCAAACCGTTCCGCCTTTGATCACGAATTACACCTCAATGAAAGC
CTGGTTAACGGGTCATGCCGATATCCATTAGAGTGATGCAAAGCATAATGTGGGTGGTACCGCGGTGAAA
ATCGTCCCTGATCGTTGAAGATCAGGGGCTTTTTTTtATGCCATTGTGAAT

Kontiga: 41

Početna pozicija u kontigi: 82654

Krajnja pozicija u kontigi: 82844

Udaljenost od levog ORF-a: 140

Udaljenost od desnog ORF-a: 526

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvenca 21

Familija: T-box

Kod:

ACAAACAATAAAAAACGATGAGTAGAAGAGTACCTGAGTTGTCGGTGTGACGAGTCTCGGTTGGTGGGA
AGAGACCTCTGGTAGCTTTGGGAAAGTAGTCTACGAGTTGCCGAGGTGACTTTGGAAGCTTCGGTTGGGA
TTGCGCCCAATTATAGCAATCGTGTATCGCCAGGCAACTGGCCGTACACGGTAAGGCCGCTTACGGTGACG
TAGCGGTAAATTAAGGTGGTACAGCGTGATTCGCGCCCTTAACAGTCTCGACTGTTAAGGGCTTTTCTTT
tGGTTAATTAGCCA

Kontiga: 37

Početna pozicija u kontigi: 9254

Krajnja pozicija u kontigi: 9548

Udaljenost od levog ORF-a: 378

Udaljenost od desnog ORF-a: 69

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvenca 22

Familija: T-box

Kod:

GGTCCGCAGCGAGTTGTCATTTGGTGAAAGACAATGATCGGGCACTGGCGAAGATGGTCTTTTGAAGCAT
TCAGTTGCGAACACATCTTGCAAGCAACTGACGGGCACTCCCGTTACTGAGTCAGGGTATAACGCAAACT
TGCGAAGTACTTAGAGAGAATCGATTTGATCGGTTAATCAAGGTGGTAACGCGGAAAATCCGTCCTTGT
C

Kontiga: 28

Početna pozicija u kontigi: 69946

Krajnja pozicija u kontigi: 70157

Udaljenost od levog ORF-a: 327

Udaljenost od desnog ORF-a: 167

Sekvenca je zadovoljila sve definisane uslove.

***riboswitch* sekvenca 23**

Familija: *T-box*

Kod:

```
AATCGGACAAAAGCTTTTTTtGATTAATTTCAATTAGTTATTGACAAGAAAACAATCCATGCTAAACTA  
AAAAATCAAATATAAATGCAACGCGAAAAGACAAGTACCCGTGTCTGTGAATACAGCGAGCTTCGATTGGTG  
TGAGGAAGCATTTCATGATATGGCGAATATCCCTTTCAAGCAGGACACCGAACTTCAGTAACGTGTCATGG  
GGCACCATTACAGTGTAGCATATCGTCATTGGACCGTATGCGATGAAGACCAAGGCGCAGCTGGCTT  
TGGTGAATTAAGGTGGTACCGCGCAAAGTCGCCCTTAGAA
```

Kontiga: 28

Početna pozicija u kontigi: 66417

Krajnja pozicija u kontigi: 66737

Udaljenost od levog ORF-a: 462

Udaljenost od desnog ORF-a: 308

Sekvenca je zadovoljila sve definisane uslove.

***riboswitch* sekvenca 24**

Familija: *T-box*

Kod:

```
AAACGGCAGCGAATATGGTCTTGGTGATGCAGGAACGCTGAGCGATTGCGAGGCGGTCATGGGGACGCC  
ATTATTGCGTTAGGTCTCACAGCTATTTGCTGTGCGACTGATTGAGGAAAGTCGTGTGAACGGCTTTCAA  
ACCAAGGTGGTACCACGGTAAGCCGTCCTTGGCGCAGACAATTTGTCTGCGCTGAGGGCGGCTTTTTtGA  
CGTTCACGATTTT
```

Kontiga: 28

Početna pozicija u kontigi: 47730

Krajnja pozicija u kontigi: 47953

Udaljenost od levog ORF-a: 782

Udaljenost od desnog ORF-a: 129

Sekvenca nije zadovoljila uslov koji se odnosi na regulaciju gena nizvodno od nje.

***riboswitch* sekvenca 25**

Familija: *TPP*

Kod:

```
TTCCGACCATTACAAATCCCTACGCTCGCATTAACGAACAGGTTCTGAAGGGTCCGGGTTTCCCCGTTCT  
CAGCCAATGGCTCCCCTTTGTGAAGGT
```

Kontiga: 53

Početna pozicija u kontigi: 204882

Krajnja pozicija u kontigi: 204979

Udaljenost od levog ORF-a: -10

Udaljenost od desnog ORF-a: 81

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvenca 26

Familija: Lysine

Kod:

CGATTTCGTTAGCGCCCCGCGATGTTTCGCGACAGTCCAGCGGATGTTGTCCAACCTGGCCCAATGTCGGATT
TCTGAACAAATCCGGCATTTCGGCATCCACCCCTTTTGCAACCGCGCATCTGGCGTTCAGGCCATTGCGG
TTAGGATTCATGACAGATGCAACCTCTAATCGTTAT

Kontiga: 97

Početna pozicija u kontigi: 7016

Krajnja pozicija u kontigi: 7192

Udaljenost od levog ORF-a: 301

Udaljenost od desnog ORF-a: 59

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvenca 27

Familija: Lysine

Kod:

CAAACAGATAGAGGTTGCAACCAACATGAGTAGCTTGATCGAGCGGGATCCTTCGTGATGACATCAAGCC
AAAGGGGCGGTTGCCGAAATGGGCGCGCCGGGATCGGCAAGCCTGTTGGGCCAGGATTGAACAAATCCTG
TACTGTCGCTTAGAAACCTAAGCGGGGCGCTATCATTCAA

Kontiga: 193

Početna pozicija u kontigi: 204

Krajnja pozicija u kontigi: 384

Udaljenost od levog ORF-a: 5

Udaljenost od desnog ORF-a: 98

Sekvenca je zadovoljila sve definisane uslove.

riboswitch sekvenca 28

Familija: Lysine

Kod:

CCAAGAAATAGAGGTTGCGGTGGGCAACGAAAACCTGGGGAGTGCCGTATCATGTTGAACCAGTTGTAAGT
GCCACCGCCGAAATGTTGCAGATATTCGGATCGTGCGCATTGGGCCATGGGTGCATAACCCATGGACTG
TCGAAGCACTCGCTTCGGGGCGCTATGATTCTG

Kontiga: 125

Početna pozicija u kontigi: 59367

Krajnja pozicija u kontigi: 59540

Udaljenost od levog ORF-a: 426

Udaljenost od desnog ORF-a: 103

Sekvenca je zadovoljila sve definisane uslove.