Универзитет у Београду
Математички факултет

Развој нове методе за одређивање елемената за контролу генске експресије код бактерија

- Мастер рад -

Ментор:
др Ненад Митић

Кандидат:
Марко Стојићевић

Београд,
октобар 2011.

Ментор:

др Ненад Митић
Математички факултет,
Универзитет у Београду

Чланови комисије:

др Гордана Павловић-Лажетић
Математички факултет,
Универзитет у Београду

др Милош Бељански
Институт за општу и физичку хемију, Београд

Датум одбране: _____

Садржај

Списак слика

Списак табела

Предговор

Генска експресија има важну улогу у организму јер одређује шта се из генотипа издваја као фенотип. Део молекула РНК који регулише експресију гена зове се *riboswitch*. Постоји више фамилија *riboswitch*-ева, а свака се разликује по одређеном низу нуклеотида (*консензусу*) који је специфичан за ту фамилију. Циљ овог рада је да се развије метода за одређивање елемената за контролу генске експресије код бактерија, тј. метода за проналажење *riboswitch*-ева. Алгоритам методе се састоји из неколико корака, где се прво издваја део ДНК који испуњава почетне услове да се на том месту налази *riboswitch*. Потом се врши претрага да ли се у неком од издвојених делова налази косензус који би рекао о којој фамилији *riboswitch*-ева је реч. Уколико се консензус налази у издвојеном делу, потребно је да се у истом налази и одређена палиндромска секвенца како би постојала велика вероватноћа да је то заиста *riboswitch*. Проверу пронађене секвенце и утврђивање да ли се заиста ради о *riboswitch*-у, вршиће Лабораторија за молекуларну генетику индустријских микроорганизама, Института за молекуларну генетику и генетичко инжењерство, Универзитета у Београду. За развој методе, коришћена је секвенца генома соја *Lactobacillus paracasei subsp. paracasei BGSJ2-8*, која је власништво наведене Лабораторије.

Како тематика овог рада спада у новије научне дисциплине, не постоји велики број радова на ову тему. Стога је у оквиру рада, поред поменуте методе, приказан и кратак опис постојећих решења за проналажење *riboswitch*-ева.

Добијени резултати су упоређени са резултатима који су добијени неком од постојећих софтверских метода и такође су представљени у оквиру рада.

Захвалница

Велику захвалност за помоћ при изради мастер рада дугујем пре свега свом ментору, др Ненаду Митићу, професору Математичког факултета Универзитета у Београду, који ме је својим критикама и саветима водио кроз рад. Поред ментора, захвалност дугујем и др Милошу Бељанском са Института за општу и физичку хемију из Београда, који ми је помогао да разумем тематику и важност представљеног рада. Посебну захвалност дугујем сарадницима Института за молекуларну генетику и генетичко инжењерство, пре свега др Наташи Голић, др Јелени Беговић, као и др Бранку Јовчићу. Они су ми омогућили приступ подацима које сам користио при изради овог рада, али су ми и својим саветима и несебичном подршком помогли, уз ментора и др Бељанског, да рад изгледа овако.

Глава 1

Увод

Биоинформатика је научна дисциплина која решава проблеме из области биологије уз помоћ рачунара. Већи рачунарски ресурси, који су постали доступни последњих неколико година, условили су бржи развој ове области. Значај коришћења рачунара у биоинформатици се огледа пре свега у бржем израчунавању одређених задатака, или чак резултата извршавања експеримената. Наиме, раније су се експерименти изводили *in vivo* (на живим организмима), или *in vitro* (у вештачком окружењу). Данас је захваљујући рачунарима могуће да се симулира извршавање различитих експеримената за веома кратак временски период, на пример да се увиде одређени обрасци понашања унутар генома неког организма. Управо је то разлог што се биоинформатиком последњих година бави све више научника.

1.1 Основни појмови

Генетика је установљена 1866. године од стране Грегора Мендела, када је изнео своје резултате експеримената на баштенском грашку. Неколико година касније, 1869. године, Фридрих Мишер је изоловао ДНК (дезоксирибонуклеинска киселина) која заједно са РНК (рибонуклеинска киселина) спада у *нуклеинске киселине*. Основна функција нуклеинских киселина у организму је биосинтеза протеина, као и пренос генетичког материјала.

Нуклеинске киселине. ДНК је молекул у облику ланца који је испреплетан у *двоструки хеликс*, при чему се *нуклеотид* једног ланца спаја са својим комплементарним нуклеотидом из другог ланца (слика 1.1). Један ланац нуклеотида почиње $5'$ крајем, а завршава се $3'$ крајем (*кодирајући* ланац - *„sense"*) док се насупрот њега налази комплементаран ланац (*некодирајући*, или *темплатни* ланац - *„antisense"*). Некодирајући ланац је супротно оријентисан (антипаралелно) и његов $3'$ крај је постављен наспрам $5'$ краја кодирајућег ланца. Основне јединице ДНК ланца су нуклеотиди, који су сачињени од шећера ($2'$-дезоксирибоза), фосфорне киселине и азотних база. Азотне базе могу да буду *пуринске* (аденин и гуанин) или *пиримидинске* (тимин и цитозин) и међусобно су повезане водоничним везама, чиме спајају два ланца нуклеотида. Спарива-

ње се врши тако што је аденин једног ланца увек у пару са тимином из другог, наспрамног ланца. Такође, гуанин једног ланца је увек у пару са цитозином наспрамног ланца. Табеларни приказ нуклеотида је дат у табели 1.1, а детаљан приказ свих ознака за запис нуклеотида ДНК (РНК) ланца је приложен у додатку.

Поред ДНК, у нуклеинске киселине спада и РНК. Разлика између нуклеинских киселина (ДНК и РНК) је у следеће три особине:

- РНК је једноланчана,
- У састав РНК улази шећер рибоза,
- Уместо базе *тимина*, који се јавља у ДНК, код РНК се јавља *урацил*.

Ознака	**Назив**	**Где се јавља**
A	Аденин	ДНК и РНК
G	Гуанин	ДНК и РНК
T	Тимин	ДНК
C	Цитозин	ДНК и РНК
U	Урацил	РНК

Табела 1.1: Приказ могућих текстуалних записа нуклеотида у оквиру произвољног ДНК или РНК ланца

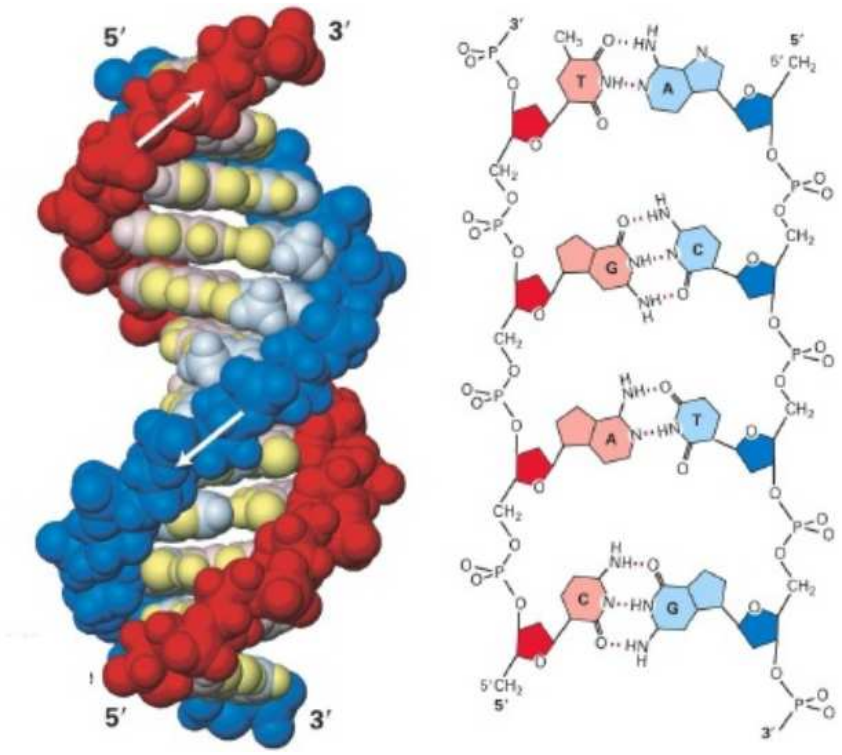
Постоје три типа рибонуклеинских киселина (РНК):

- информациона РНК - иРНК (енг. *Messenger RNA, mRNA*),
- транспортна РНК - тРНК (енг. *Transfer RNA, tRNA*),
- рибозомална РНК - рРНК (енг. *Ribosomal RNA, rRNA*).

Функција информационе РНК је пренос информације са гена до биосинтезе протеина (експресија гена). иРНК се синтетише на основу информације ДНК у процесу транскрипције (детаљније о транскрипцији на страни 3) и представља ,,отисак" на основу којег се добија протеин. Структура иРНК представљена је на слици 1.2.

Транспортна РНК је молекул РНК који је обично дужине од 73 до 93 нуклеотида. Њена функција је пренос амино киселине до одређеног места у ћелији где се врши биосинтеза протеина (до рибозома).

Рибозомална РНК је РНК део рибозома, те има градивну функцију (да улази у састав рибозома). рРНК омогућава механизам за превођење иРНК у аминокиселине и заједно са тРНК учествује у процесу транслације (детаљније о транслацији на страни 5) тако што омогућава активност пептидил трансферазе.



Слика 1.1: Секундарна структура ДНК. На слици је приказана секундарна структура молекула ДНК, као и парови азотних база ДНК и њихов начин везивања.

Геном. Геном је комплетан скуп наследних информација једног организма [1]. Код бактерија, најмањи познат геном је дужине 159662 (*Carsonella ruddii*), док најдужи познат садржи 9970000 нуклеотида (*Solibacter usitatus*). Просечне дужине генома код бактерија су између 3 и 4 милиона нуклеотида, и обично су кружног (циркуларног) облика, ређе су линеарни.

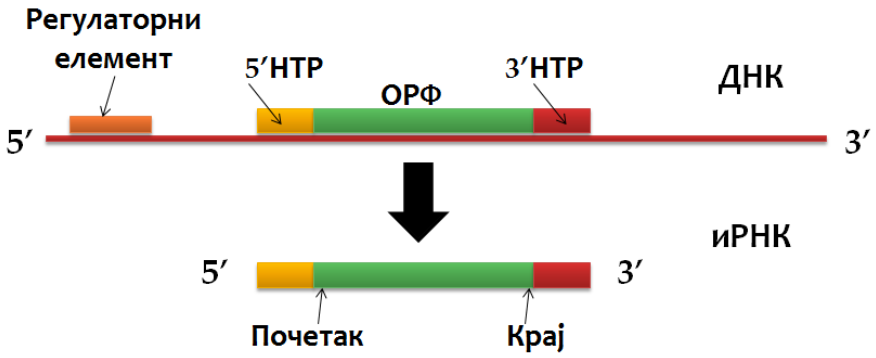
Транскрипција. *Транскрипција* је процес у коме се информација са ДНК преписује на иРНК посредством РНК полимеразе. РНК настаје на основу некодирајућег ДНК ланца, док кодирајући ланац садржи генетичке информације. У току транскрипције се у ланац иРНК додаје нуклеотид по нуклеотид, при чему је сваки нуклеотид комплементаран нуклеотидима некодирајућег ДНК ланца. Резултат овог процеса је РНК ланац који је идентичан кодирајућем ланцу ДНК са наведеном разликом да је тимин (T) замењен урацилом (U). Транскрипција се завршава посебним низом нуклеотида који формирају терминаторску петљу (*терминатор*).

Терминатор. Терминатор је нуклеотидна секвенца која означава крај гена. Код прокариота се разликују две врсте терминатора:

- Унутрашњи терминатор (или ρ-независни терминатор),
- ρ-зависни терминатор.

Показано је [6] да је унутрашњи терминатор механизам који код прокариота зауставља транскрипцију. Терминатор представља структуру дужине од седам до двадесет базних парова која има облик петље (омче). Пример ове структуре је приказан на слици 1.3.

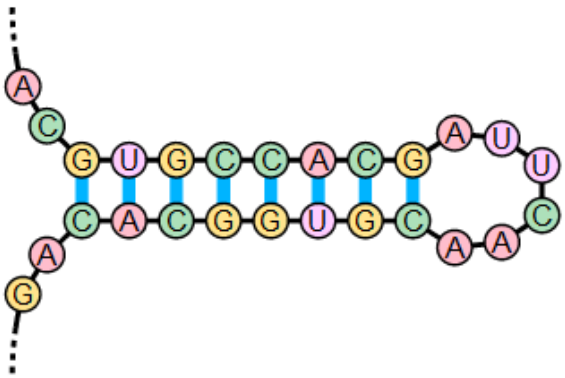
ρ-независни терминатор се често појављује у оквиру РНК, у облику елемената који се зову *рибо-прекидачи* (енг. *„riboswitch"-еви*).



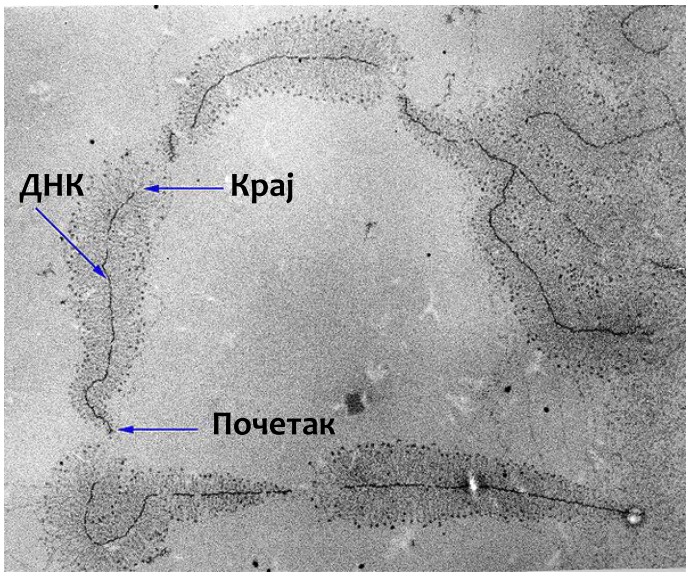
Слика 1.2: Структура информационе РНК. *Легенда:* НТР - нетранслирајућа РНК (енг. *Un-translated RNA, UTR*). ОРФ - отворени оквир читања (енг. *Open reading frame*)

Кодон. Кодон представља триплет база на информационој РНК који кодира једну амино киселину. За једну амино киселину постоји неколико кодона. Три врсте кодона су битне за потребе овог рада:

- *старт кодони* - који означавају почетак биосинтезе протеина (најчешћи је *AUG*),
- *стоп кодони* - који означавају завршетак биосинтезе протеина (најчешћи су *UAA, UAG, UGA*),
- *антикодони* - који се налазе на тРНК и комплементарни су кодону.



Слика 1.3: Пример терминатора у РНК

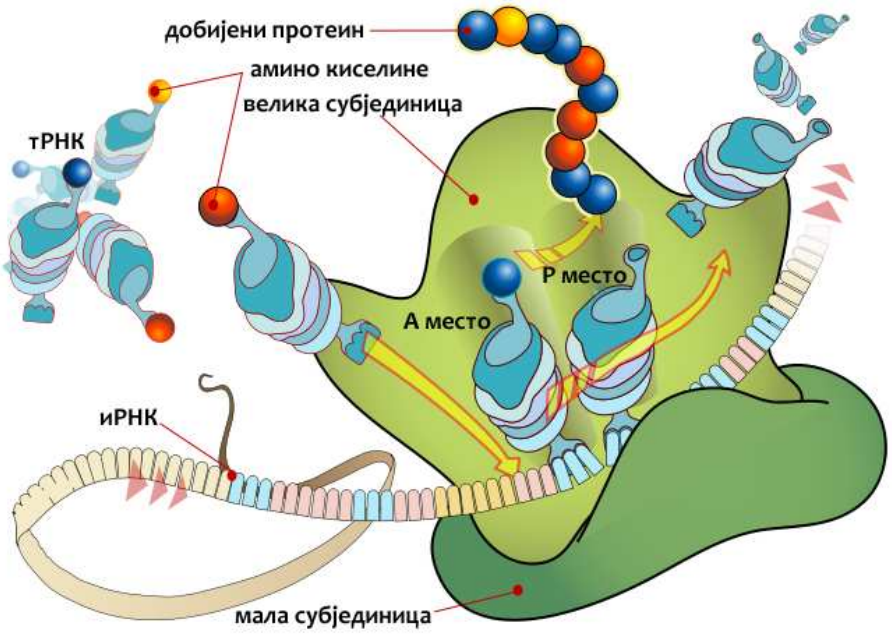


Слика 1.4: Електронска микрографија процеса транскрипције и траслације. Формиране иРНК нити су приказане као гране које се одвајају од ДНК ланца.

Амино киселина. Амино киселине су органске киселине које имају амино групу у свом саставу, на другом C атому. Мањим делом се налазе слободне у организму, а од важности су пре свега јер улазе у састав протеина.

Протеин. Протеини су биомакромолекули састављени од великог броја амино киселина које су повезане пептидном везом. Саме амино киселине повезане у ланац чине полипептид, чијим даљим увијањем настаје функционалан протеин.

Транслација. Транслација је процес у коме се информациона РНК, која је добијена процесом транскрипције, преводи у протеин (слика 1.5). Код прокариота се транскрипција и транслација одвијају скоро истовремено (слика 1.4) и то у цитоплазми. Прва фаза транслације обухвата процесе којима се иРНК везује са малом и великом субјединицом рибозома. У рибозому има два места, која се називају P и A место и за које могу да се вежу две тРНК. Прва тРНК која улази у рибозом садржи антикодон који је комплементаран старт кодону. Та тРНК носи аминокиселину *метионин* и везује се за P место. На A место у рибозому долази друга тРНК чији је антикодон комплементаран наредном кодону иРНК (први после старт кодона). Две аминокиселине су сада близу једна другој и долази до кидања везе између метионина и тРНК. Енергија која се тада ослободи се користи за стварање пептидне везе између прве две аминокиселине. Након овог корака, тРНК са P места напушта рибозом, а друга по реду тРНК прелази са A на P место. Тиме се рибозом помера (у $5'$ - $3'$ правцу) за један кодон дуж иРНК. Сада се на P месту налази тРНК за коју су везане две аминокиселине, а место A је слободно. Зато на место A улази следећа по реду тРНК која носи антикодон комплементаран трећем кодону иРНК и процес се понавља све док се у иРНК не стигне до стоп кодона.



Слика 1.5: Процес транслације

ОРФ. У генетици се под скраћеницом ОРФ (отворени оквир читања, енг.*Open Reading Frame*) подразумева низ нуклетиода који почиње старт кодоном, а завршава се непосредно испред стоп кодона. Код прокариота термин ОРФ је еквивалентан термину кодирајућа секвенца (енг. *Coding Sequence, CDS*) и представља део ДНК ланца који се преводи у протеин. Такође, код прокариота, термин ОРФ се поистовећује са термином *ген*. Илустрација позиције ОРФ-а је приказана на слици 1.2

1.2 Експресија гена

Експресија гена је процес у коме се информације које носи ген користе за синтезу функционалног производа гена (то су најчешће протеини, а код непротеинских кодирајућих гена под производом се подразумева функционална РНК). У генетици, генска експресија има важну улогу јер она одређује шта ће да се из генотипа испољи као фенотип.

1.2.1 *Riboswitch*

Riboswitch (*switch*- прекидач) је део некодирајуће РНК који регулише експресију гена. Регулација експресије гена се постиже мењањем конформације у односу на то да ли је лиганд (најчешће је то мали молекул, или јон) везан, или не. Једноставније речено, *riboswitch* садржи ,,прекидач", а експресија гена зависи од тога да ли је тај прекидач ,,укључен", или ,,искључен".

До сада је откривено неколико фамилија *riboswitch*-ева и оне су приказане у табели 1.2. Свака фамилија *riboswitch*-ева се разликује по домену аптамера, који обезбеђује место за везивање лиганда и могућност за експресију. Домен

аптамера (у даљем тексту консензус) је високо конзервисан у *riboswitch*-евима који припадају истој фамилији.

Редни бр.	**Фамилија**	**Редни бр.**	**Фамилија**
1.	*FMN*	13.	*preQ1-II*
2.	*TPP*	14.	*MOCO_RNA_motif*
3.	*SAM*	15.	*Mg_sensor*
4.	*Purine*	16.	*SAH_riboswitch*
5.	*Lysine*	17.	*rli52*
6.	*Cobalamin*	18.	*rli53*
7.	*glmS*	19.	*rli54*
8.	*Glycine*	20.	*rli55*
9.	*SAM_alpha*	21.	*rli56*
10.	*PreQ1*	22.	*rli61*
11.	*SAM-IV*	23.	*rli62*
12.	*T-box*		

Табела 1.2: Списак фамилија *riboswitch*-ева, преузет са *RFAM*-а

Лиганд. У општем смислу, лиганд представља мали молекул или јон, који везивањем за неки биомакромолекул омогућава испољавање њихове функције.

Лиганд, или како се још назива *афекторни* или *регулаторни* молекул, се код *riboswitch*-ева везује за РНК при одређеним физиолошким условима. Како постоји више фамилија *riboswitch*-ева, за сваку од њих је дефинисан посебан лиганд. Због тога се припадност некој од фамилија *riboswitch*-ева одређује на основу лиганда, односно поставље се питање да ли постоји место на ком би се афекторни молекул везао. Сам *riboswitch* се налази испред гена и везивањем лиганда долази до блокирања одређеног нуклеотида чиме се омогућава његова функција.

Riboswitch се налази у оквиру РНК, узводно од старт кодона, тј. нуклеотидна секвенца која дефинише *riboswitch* се у оквиру ДНК налази између два ОРФ-а. Тај простор се назива *интергенски регион* (скраћено, *ИГР*). Код прокариота један *riboswitch* може да контролише експресију већег броја гена, уколико су они позиционирани један иза другог. Овакви ОРФ-ови, који су контролисани једним регулаторним елементом, називају се *оперони*. Оперони најчешће кодирају протеине који су укључени у исти метаболички пут, или имају сличну функцију у ћелији и неопходно је да буду истовремено синтетисани у једнаком броју примерака. Тиме се такође постиже и енергетска уштеда која је од велике важности за прокариоте.

Riboswitch-еви, дакле, закључавају гене које регулишу чиме спречавају њихову експресију када она није пожељна. Обзиром да се ради о регулацији на нивоу транслације, могућа је бржа реакција бактеријске ћелије. Оваква реакција

на спољашње услове условљава и брже откључавање датих гена, те се и протеини синтетишу брже него код гена који су регулисани на нивоу транскрипције. Из тог разлога бактерије које користе *riboswitch*-еве као своје регулаторне елементе имају селективну предност у средини у којој живе у односу на бактерије које немају овај систем регулације.

За потребе овог рада нису од важности све постојеће фамилије *riboswitch*-ева, већ само оне које су наведене у табели 1.3. Кратак преглед ових фамилија приказан је у наставку поглавља, док је више информација (укључујући и детаљне консензусе) доступно на странама *Rfam* базе података (детаљније о *Rfam* бази је изложено на страни 12).

Назив фамилије	**Приступни код (*Rfam*)**	**Линк**
FMN	*RF*00050	`http://rfam.janelia.org/cgi-bin/getdesc?acc=RF00050`
TPP	*RF*00059	`http://rfam.janelia.org/cgi-bin/getdesc?acc=RF00059`
SAM	*RF*00162	`http://rfam.janelia.org/cgi-bin/getdesc?acc=RF00162`
Purine	*RF*00167	`http://rfam.janelia.org/cgi-bin/getdesc?acc=RF00167`
Lysine	*RF*00168	`http://rfam.janelia.org/cgi-bin/getdesc?acc=RF00168`
T-box	*RF*00230	`http://rfam.janelia.org/cgi-bin/getdesc?acc=RF00230`

Табела 1.3: Списак фамилија *riboswitch*-ева чији су консензуси обрађени у оквиру рада

Фамилија *FMN*

RFN елемент је високо конзервисана секвенца која се налази у оквиру 5′-нетранслирајућих[1] региона иРНК и задужен је за биосинтезу *флавин мононуклеотида* (*FMN*) и транспорт протеина. Овај елемент је метаболички зависан *riboswitch* који директно везује *FMN* у недостатку протеина. Ови *riboswitch*-еви највероватније контролишу генску експресију тако што изазивају рани завршетак транскрипције. Секундарна структура ове фамилије је приказана на слици 1.6а.

[1]Нетранслирајућа иРНК (енг. Un-translated mRNA) - редослед који не кодира протеин

Фамилија *TPP*

Витамин $B(1)$ у својој активној форми - *тиамин пирофосфат* (TPP) - је неопходни коензим који се систентише у бактерији. Раније су се тиамин регулаторни елементи означавали као *thi box*, а данас је знање о њима проширено. Регулаторни елементи који регулишу тиамин, имају високо конзервисану РНК секундарну структуру и зову се *THI* елементи. THI елемент је *riboswitch* који директно везује TPP за потребе регулисања генске експресије уз помоћ различитих механизама у бактеријама и еукариотама. Секундарна структура ове фамилије је приказана на слици 1.6б.

Фамилија *SAM*

Нуклеотидне секвенце које дефинишу *SAM riboswitch* (слика 1.6ц) се често налазе узводно од различитих гена који синтетишу протеине укључене у биосинтезу метионина. Ова фамилија *riboswitch*-ева делује на нивоу контролисања завршетка транскрипције. Структура ових елемената, која је делом садржана и у оквиру консензуса, састоји се од сложених региона петљи (омчи), где се сваки завршава терминаторским регионом.

Фамилија *Purine*

У оквиру *Bacillus subtilis* ова фамилија је пронађена на минимум пет раздвојених делова који се транскриптују и који кодирају 17 гена. Ти гени су углавном укључени у транспорт пурина и биосинтезу пуринских нуклеотида. Неки пронађени елементи ове фамилије су специфични, односно искључиви за аденин. Секундарна структура ове фамилије је приказана на слици 1.6д.

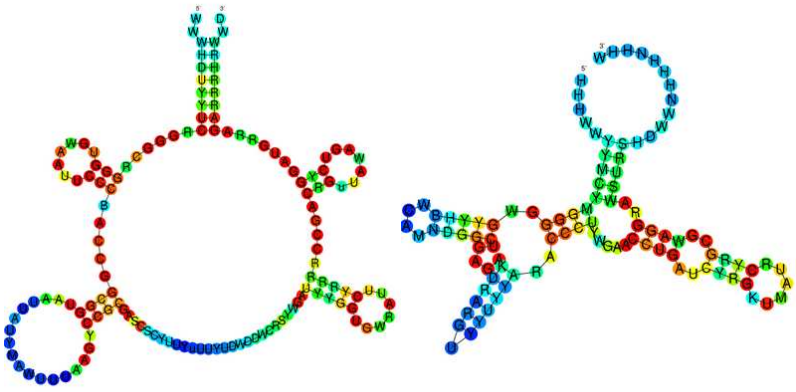
Фамилија *Lysine*

Ова фамилија укључује *riboswitch*-еве који препознају лизин у различитим генима који су укључени у метаболизам лизина. Њена секундарна структура је приказана на слици 1.6е.

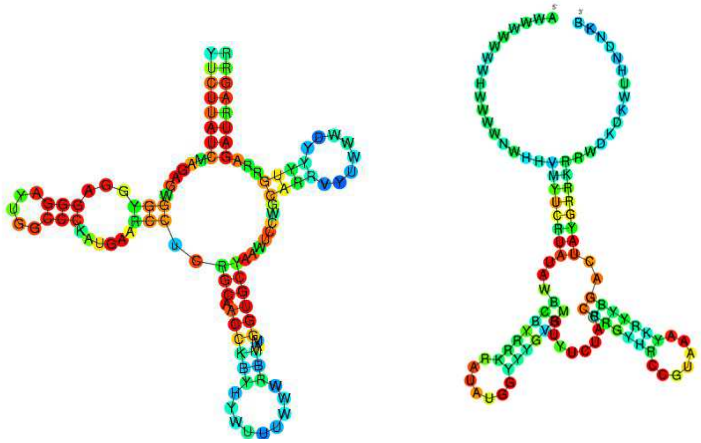
Фамилија *T-box*

T-box се налазе узводно од многих гена за аминоацил-тРНК синтетазу код Грам-позитивних бактерија[2]. Антикодон у тРНК се упарује са секвенцом унутар *T-box* мотива.

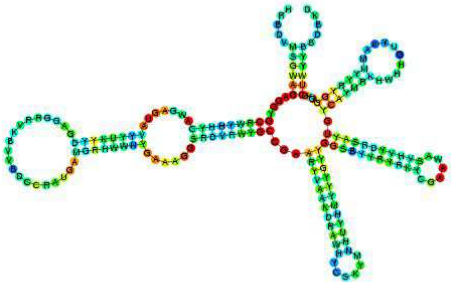
[2]Грам-негативне бактерије имају један слој липополисахарида који покрива њихов ћелијски зид, док Грам-позитивне бактерије немају тај слој. Услед тога, оне се на Грам тесту боје црвено, односно плаво, респективно.



(а) *FMN* (б) *TPP*



(ц) *SAM* (д) *Purine*



(е) *Lysine*

Слика 1.6: Секундарна структура неких фамилија *riboswitch*-ева. *Легенда:* Боје на слици означавају конзервисаност секвенце. Љубичаста боја има вредност 0, а црвена вредност 1, при чему 1 означава највећу конзервисаност

Целокупни живи свет је класификован у три *супер краљевства*: Архе бактерије, Бактерије и Еукарије. У природи постоје два типа ћелија: ћелије чије је једро (карион, нуклеус) одвојено двоструком мембраном од остатка цитоплазме, односно еукариотске (еу - прави, истински) и ћелије које не поседују једро одвојено двоструком мембраном од остатка цитоплазме, односно прокариотске (про- пре, пређашњи). Ћелије Архе бактерија и Бактерија спадају у групу прокариота јер оне немају развијено једро. Прокариоте су углавном једноћелијски организми који могу да живе заједно, у колонијама.

Елементи за контролу генске експресије, чије проналажење представља тему овог рада, се јављају и код прокариота и код еукариота, уз неке разлике. Како је тема овог рада проналажење нове методе за идентификацију регулаторних елемената код бактерија, у даљем току рада су обрађене искључиво методе за прокариоте. Стога је у наредном одељку објашњено како се сви набројани појмови представљају на рачунару, док је у следећој глави приказан кратак преглед досадашњих метода, а потом и детаљно објашење нове методе.

1.3 Рачунарска обрада података

ДНК и РНК, као што је познато, чини двоструки, односно једноструки ланац, састављен од нуклеотида. У рачунару се ДНК (РНК) складишти као низ карактера, где сваки карактер представља неку од азотних база. Који карактер представља коју базу приказано је у табели 1.1. Додатно, уколико се база налази са вероватноћом један на одређеној позицији у низу нуклеотида, она се обележава великим словом, а у супротном малим. У зависности од договора, могу да се уведу додатни карактери за опис ситуације када, на пример, не може са сигурношћу да се одреди која се од две базе налази на одређеној позицији. Уколико уопште не може да се одреди која би база могла да се налази на некој позицији, то је обично обележено словом N, или на пример, тачком, цртицом, итд.

Уколико се геном посматра као низ карактера, онда је и обрада података једноставнија. Обрада може да се изврши директно на рачунару (у локалу), или путем клијент-сервер архитектуре. Одабир места где ће подаци да се складиште (обрађују) зависи пре свега од тога шта је за ту одређену намену повољније. Метода која је развијена ради искључиво у локалу и у скоријој будућности неће бити развијана за друге намене. За потребе овог рада, због велике количине података, подаци су ради лакшег складиштења и обраде чувани у релационој бази података. База која је коришћена у овом раду је *IBM DB2* база података, верзија 9.7. Обрада складиштених података се делом вршила директним упитима, а делом уз помоћ додатних програма који су посебно писани за потребе овог рада. За писање додатних програма коришћени су програмски језици *C* и *Java*. Детаљније објашњење коришћених програма је приказано у глави 3.

Rfam. *Rfam* [9] (*RNA families database*) је *online* база података која садржи поравнавајуће секвенце и коваријансне моделе многих честих некодирајућих РНК фамилија. Подаци из ове базе могу да се користе за претрагу генома, или других ДНК секвенци, за структуре које су сличне познатим РНК фамилијама. *Rfam* база је од великог значаја за овај рад јер су подаци који су коришћени за препознавање фамилија *riboswitch*-ева, преузети управо одавде. Садржај ове базе се ажурира у одређеном временском интервалу, а тренутно је актуелна верзија 10.0 из јануара 2010. године са 1466 фамилија.

Податке је могуће преузети у више различитих записа, али најчешћи је Стокхолмски формат (енг. *Stockholm format*). То је формат за поравнање више секвенци (контига) и он је најчешће коришћен формат за *Rfam* базу података. Поред *Rfam* базе, овај формат се користи и у програмима у којима се користи вероватноћа за претрагу база података. За потребе овог рада коришћени су подаци *Rfam* базе у Стокхолмском формату записа.

Поред наведених података, за развој нове методе коришћен је и геном соја *Lactobacillus paracasei subsp. paracasei BGSJ2-8*. Све обраде овог соја су се заснивале на обради његових **ДНК секвенци**. Детаљан опис алгоритма методе, као и припреме и обраде података, изложени су у глави 3, док су у глави 2 приказана постојећа решења разматраног проблема.

Глава 2

Формулација проблема

Проналажење елемената за контролу генске експресије представља проналажење одређене нуклеотидне секвенце у оквиру ДНК. Регулаторни елементи су се раније проналазили искључиво уз помоћ експерименталних метода и то је био веома дуг и обиман лабораторијски посао. Рачунарска обрада података је значајна за решење овог проблема, јер обезбеђује бржу и сигурнију обраду података, уз скраћење времена потребног за вршење експеримената у лабораторијама. Међутим, проналажење регулаторних елемената путем рачунарских метода је такође тежак проблем јер:

- Нису математички прецизно дефинисани, већ постоје велика одступања између њих,
- Чини их релативно мали број базних парова, па се често једна секвенца (која означава регулаторни елемент) понавља више пута у оквиру генома (који има и до неколико милиона нуклеотида), а да не означава сваки пут регулаторни елемент.

Истраживање контроле генске експресије има велики значај, међутим и поред тога, не постоји велики број објављених софтверских решења. Како се овај рад бави проналажењем *riboswitch*-ева, као једних од елемената за контролу генске експресије, у овом поглављу су приказана нека од доступних софтверских решења. Такође, укратко је објашњен начин рада за неке од постојећих програма, као и њихове предности и недостаци.

2.1 Постојеће методе

Методе за претрагу *riboswitch* елемената у оквиру генома се развијају углавном на два начина. Један начин је да се на основу већ пронађених *riboswitch*-ева направи модел за сваку од фамилија, по коме ће моћи да се претраже и геноми других бактерија. Други начин је да се користе методе које препознају елементе за контролу генске експресије на основу њихових биолошких дефиниција.

Методе у којима се прави модел за препознавање произвољне фамилије се углавном заснивају на законима вероватноће. Раније су се ови модели дефинисали углавном уз помоћ *коваријансног модела* који даје добре резултате, али уз

одређене недостатке. Коваријансни модел је веома осетљив и с тога при обради обухвата велику количину података, што му смањује брзину извршавања [5]. У последње време, овај проблем се превазилази тако што се уместо коваријансног модела све чешће користе *скривени Марковљеви модели* (енг. *Hidden Markov Models, HMM*). Главна предност скривених Марковљевих модела у односу на коваријансни модел је пре свега у брзини извршавања. Међутим, показано је [5] да постоје фамилије *riboswitch*-ева за које скривени Марковљеви модели ипак дају лошије резултате у односу на коваријансни модел. При великим количинама података се ипак препоручује употреба метода које се заснивају на *HMM*, или профил скривеним Марковљевим моделима (енг. *profile Hidden Markov Models, pHMM*), пре свега због брзине извршавања, иако не дају увек најбоље резултате.

2.1.1 *RibEx*

RibEx (riboswitch explorer, url: `http://www.ibt.unam.mx/biocomputo/ribex.html` *)* је програм који је јавно доступан и коме може да се приступи путем клијент-сервер архитектуре. Овај програм претражује унете секвенце и у њима препознаје *riboswitch* елементе. Поред *riboswitch*-ева, *RibEx* претражује и друге, високо конзервисане, регулаторне елементе код бактерија. Препознавање консензуса у оквиру овог програма се врши захваљујући подацима из јавно доступне *Rfam* базе података.

Програм је писан у програмском језику *Perl* и подељен је у четири засебне датотеке, где свака има посебну функцију. Алгоритам прво претражује секвенце од 500 базних парова како би се нашли елементи који личе на *riboswitch*-еве (енг. *riboswitch-like elements, RLE*). Уколико су елементи пронађени у оквиру задате грешке претраге, резултати се чувају. Други корак алгоритма је претрага ОРФ-ова по унапред одређеним параметрима (које корисник може да измени). За претрагу ОРФ-ова, користе се предефинисани старт и стоп кодони, као и минимални број амино киселина унутар ОРФ-а. Када се пронађу позиције ОРФ-ова (и њихове оријентације), претражује се секвенца узводно (лево) од почетка ОРФ-а. Четврти део програма контролише претходна три и служи за графички приказ резултата. У тренутку писања овог рада, *RibEx* омогућава препознавање петнаест фамилија *riboswitch*-ева, а овај број је подложан промени у зависности од откривања нових консензуса.

Поред тога што је унос секвенце ограничен на дужину од 40.000 нуклеотида, што доста ограничава коришћење програма, постоје два додатна ограничења ове методе. Прво је то што регулаторни елементи морају да буду у близини барем једног кластера сличних протеина (енг. *cluster of orthologous groups of proteins, COG*). Поред тога, регулаторни елементи морају да се налазе и у минимум других пет, нередундантних генома. Из угла корисника, мана ове методе је у недовољном објашњењу могућих параметара за измену, па услед лошег постављања параметара, често се добијају нетачни резултати.

2.1.2 *Riboswitch finder*

Под називом *Riboswitch finder*, у овом поглављу се подразумева *online* програм (доступан на: `http://riboswitch.bioapps.biozentrum.uni-wuerzburg.de/`) за претрагу *riboswitch*-ева. Такође, овај програм може да се преузме и извршава на локалном рачунару.

Програм се састоји из неколико целина, где се обрада података врши на следећи начин: прво се врши претрага унете секвенце и проверава се да ли она садржи консензус за неку од фамилија *riboswitch*-ева. Уколико је консензус пронађен, издваја се секвенца и над њом се извршава даља провера (рачуна се минимум слободне енергије, као и да ли су израчунате вредности изнад одређене границе). На крају, издвојена секвенца се оцењује (добро, осредње, лоше) и алгоритам креће од почетка док се не обради цео унос.

Иако се овај програм брзо извршава, мана му је то што проналази само *riboswitch*-еве који регулишу биосинтезу пурина. Из тог разлога, резултати овог програма немају велики значај за овај рад. Додатно, мана програма је и та што не постоје параметри за додатно подешавање израчунавања (сем бирања *строг, средњи, лош* режим рада); а и постојећи нису детаљније објашњени [7].

2.1.3 *Infernal*

Infernal (*Inference of RNA alignment*) је софтверски пакет који прво омогућава дефинисање профила (модела) консензуса, а потом и њихову претрагу. Наиме, уз помоћ овог софтверског пакета прво се дефинише профил за препознавање консензуса ДНК (РНК) елемената, а потом се, на основу профила, врши и претрага *riboswitch*-ева. Такође, уколико је потребно, може да се изврши и поравнање секвенци и то на сличан начин на који се то ради и у *Rfam* бази података.

Због јаке математичке позадине, овај програм даје добре резултате, међутим, недостаци су му пре свега у компликованом коришћењу. *Infernal* ради под оперативним системом *Linux* и захтева одређену рачунарску писменост, тако да није једноставан за коришћење. Такође, споро врши обраду података и препоручује се коришћење што већег броја процесорских јединица.

2.2 Недостаци постојећих метода

Иако не постоји велики број програма за претрагу *riboswitch*-ева, они који су набројани у делу 2.1 спадају у коришћеније и дају релативно добре резултате. Међутим, поред релативно добрих алгоритамских решења, постоје пре свега биолошка ограничења која условљавају грешке у овим програмима. Нека од тих ограничења су:

- И поред високе конзервисаности фамилија *riboswitch*-ева, постоје разлике у оквиру фамилије *riboswitch*-ева за различите родове бактерија,
- Неки програми су развијани превасходно за одређене родове бактерија, па са другим родовима не дају добре резултате,

Назив програма	Место извршавања	Мане
RibEx *(Riboswitch Explorer)*	На серверу	- Регулаторни елементи морају да буду у близини барем једног кластера сличних протеина, - Регулаторни елементи морају да се налазе и у минимум других пет, нередундантних генома, - Максимална дужина унете секвенце је 40.000 нуклеотида.
Riboswitch finder	На серверу и у локалу	- Недовољно објашњени параметри, - Проналази само *riboswitch*-еве који регулишу биосинтезу пурина.
Infernal *(Inference of RNA aligment)*	У локалу (*Linux*)	- Захтева већу рачунарску писменост, - Спора обрада података, препоручује се већи број процесорских јединица.

Табела 2.1: Упоредни приказ програма за проналажење елемената који контролишу генску експресију (пре свега, *riboswitch*-ева)

- Временом се откривају нове фамилије *riboswitch*-ева, које не могу бити одмах обухваћене у старијим верзијама програма; потребно је одређено време да се обухвате нове фамилије, итд.

Фамилије *riboswitch*-ева које су распрострањене у великом броју бактерија, лакше се проналазе, јер је лакше уочити образац по коме се оне проналазе. Проблем настаје када се неке фамилије *riboswitch*-ева не налазе у неким бактеријама, па је онда и скуп на коме се може учити о консензусу мањи.

2.3 Циљ рада

Наведене мане постојећих метода условиле су потребу за развијањем нове методе која ће на ефикасан и поуздан начин да издвоји *riboswitch* секвенце, као једне од могућих елемената за контролу генске експресије. Нова метода је осмишљена пре свега за потребе Лабораторије за молекуларну генетику индустријских микроорганизама, Института за молекуларну генетику и генетичко инжењерство, Универзитета у Београду (у даљем тексту ИМГГИ). Наиме, доступна софтверска решења нису била одговарајућа, јер већина захтева познавање скривених Марковљевих модела, као и других статистичких модела, што их чини тежим за коришћење. Програми који се покрећу искључиво на *Linux* опе-

ративном систему, неопходно познавање различитих начина рачунарске обраде података, на пример релационе базе података, условиле су потребу сарадника из ИМГГИ за развијањем нове методе. Из наведених разлога, циљ овог рада је да се развије метода чији ће кораци бити детаљно разумљиви корисницима исте и која неће да захтева велико математичко, или рачунарско знање за каснију примену.

Како је од велике важности за ИМГГИ да се одреде *riboswitch* секвенце у оквиру генома соја *Lactobacillus paracasei subsp. paracasei BGSJ2-8*, развијена метода је дефисана над наведеним сојем. Из тог разлога су и параметри методе, као и резултати, представљени у односу на дати геном.

Начин на који је метода за проналажење елемената за контролу генске експресије развијена, представљен је у следећој глави. Метода обухвата оне фамилије *riboswitch*-ева који су се могли наћи у задатом соју бактерије. Један од разлога за овакав начин конструкције методе је био и тај што, у тренутку писања рада, још увек није у потпуности асемблиран комплетан геном. Детаљније информације о коришћеним фамилијама *riboswtich*-ева су изложене на страни 8.

Глава 3

Развој нове методе

Riboswitch-еви су један од начина регулације генске експресије код прокариота. Они закључавају гене које регулишу и тако спречавају њихову експресију када она није пожељна. Ови елементи се у оквиру ДНК налазе између два отворена оквира читања (ОРФ-а, тј. гена). Простор између два ОРФ-а где се налазе ови елементи, зове се *интергенски регион* (скраћено, ИГР). У овој глави је детаљно изложена метода за препознавање ових регулаторних елемената унутар генома бактерије.

3.1 Услови за постојање *riboswitch* секвенци

Потребни услови да би секвенца унутар интергенског региона била потенцијални *riboswitch* су:

- Оријентација ОРФ-ова мора да испуњава задате услове,
- Мора да постоји консензус (домен аптамера који је конзервисан за сваку од фамилија *riboswitch*-ева) унутар ИГР-а,
- Мора да постоји барем један комплементаран палиндромски низ нуклеотида унутар ИГР-а.

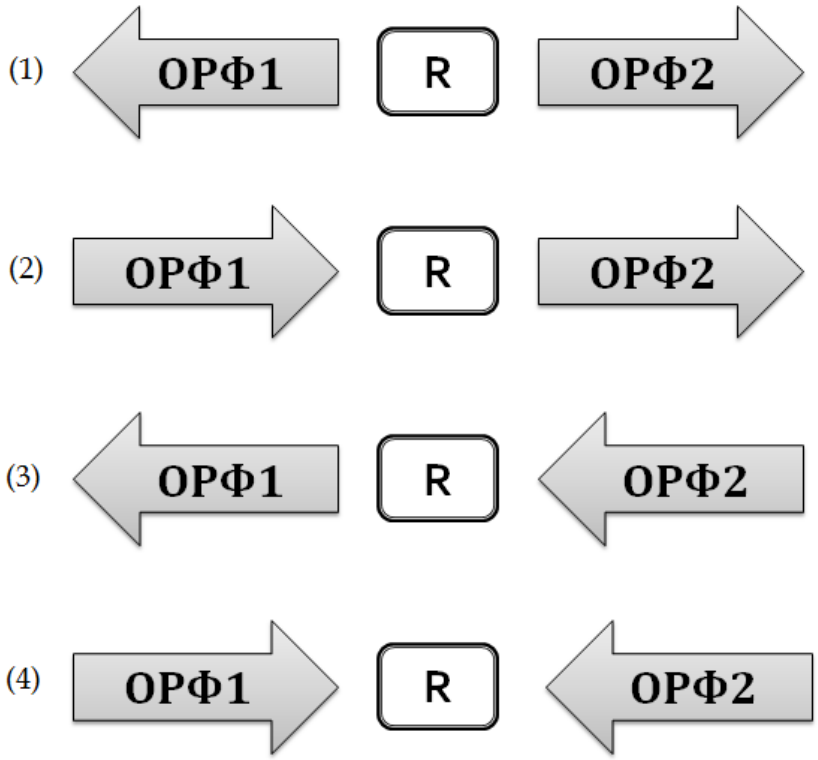
3.1.1 Оријентација ОРФ-ова

ОРФ је нуклеотидна секвенца која почиње старт кодоном, а завршава се непосредно испред стоп кодона. За потребе овог поглавља, ОРФ може да се посматра као оријентисани вектор чији почетак означава старт, а крај стоп кодон. На основу оваквог математичког модела, могу лакше да се дефинишу међусобне оријентације два ОРФ-а, између којих може да се налази *riboswitch* секвенца.

Почетни услов да се *riboswitch* налази између два ОРФ-а је испуњен уколико су они међусобно ,,добро" оријентисани. Под термином ,,добра" оријентација се подразумевају све међусобне оријентације изузев ,,крај - крај". Другим речима, *riboswitch* може да се налази између два ОРФ-а само уколико су они међусобно оријентисани ,,старт - старт", или ,,старт - стоп" (дозвољена је и оријентација

„стоп - старт”, у зависности да ли се ОРФ налази на кодирајућем, или некодирајућем ланцу). *Riboswitch* се у овако дефинисаном односу ОРФ-ова налази увек узводно од старт позиције ОРФ-а, односно налази се лево од старт кодона (уколико се ради о кодирајућем ланцу).

Графички приказ могућих оријентација ОРФ-ова је представљен на слици 3.1.



Слика 3.1: Могуће оријентације ОРФ-ова. Оријентације (1) – (3) приказују „добре” оријентације, док између ОРФ-ова приказаних на оријентацији (4) не може да се налази *riboswitch* елемент. *Легенда: R - riboswitch.*

3.1.2 Консензус фамилија

Консензус, који чине нуклеотиди у оквиру *riboswitch* секвенце у близини којих се везује лиганд, одређује фамилију *riboswitch*-ева. На основу консензуса је релативно лако да се препозна о којој фамилији регулаторних елемената је реч, јер та нуклеотидна секвенца има мала одступања у оквиру произвољне фамилије *riboswitch* елемената. То је разлог зашто је препознавање консензуса један од услова за постојање ових регулаторних елемената.

Дефинисање консензуса за целу фамилију *riboswitch*-ева није једноставан задатак. Потребно је да се препознају *riboswitch* елементи у оквиру неког соја бактерије и да се на основу појединачних консензуса дефинише консензус за целу фамилију. Квалитет консензуса зависи пре свега од два чиниоца:

- Количине обрађених података - што је више појединачних консензуса за различите сојеве бактерија, модел ће бити веродостојнији,
- Начина на који се дефинише консензус за фамилију.

Први услов, количина обрађених података, зависи од доступности фамилија *riboswitch*-ева за различите сојеве бактерија. За неке фамилије број пронађених појављивања у различитим сојевима се мери хиљадама, док су неке, углавном новије, препознате свега неколико пута. Уколико су консензуси пронађени мали број пута и ако их чини релативно мали број нуклеотида (до шест, седам), тешко да би било какав модел могао са великом сигурношћу да даје резултате. Други услов, дефинисање консензуса за фамилију, односно прављење модела на основу којег се врши препознавање, је сложен посао. За прављење модела се користе различити математички модели, пре свега засновани на вероватноћи и статистици.

На основу наведеног, постоје две врсте програма, једни који сами дефинишу консензус на основу унетих података од стране корисника, и други који користе већ готове податке. У већини случајева, исплативије је користити већ изведене консензусе који су јавно доступни у *Rfam* бази података, јер су они дефинисани на основу великог броја инстанци. У оквиру овог рада, коришћени консензуси су преузети из наведене базе података. Списак постојећих фамилија *riboswitch*-ева је приказан у табели 1.2.

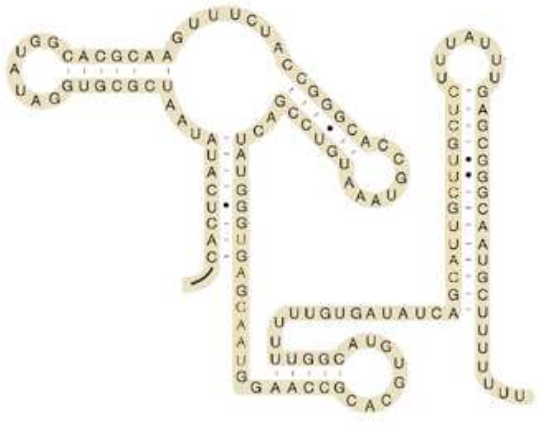
Због услова задатих од стране сарадника ИМГГИ, у овом раду је обрађено шест фамилија *riboswitch*-ева. Њихови називи, као и приступни линкови за *Rfam* базу су приказани у табели 1.3. Број обухваћених фамилија је условљен родом бактерије за коју је развијана метода, јер неке фамилије *riboswitch*-ева не могу да се очекују у роду *Lactobacillus*. Са друге стране, консензуси неких фамилија су били превише кратки тако да ни они нису узети у обзир. Описи коришћених фамилија су представљени на страни 8.

3.1.3 Комплементарне палиндромске ниске

Да би се у оквиру интергенског региона нашла потенцијална *riboswitch* секвенца, у датом сегменту мора да се налази барем један комплементаран палиндромски низ. Постојање оваквих низова омогућава настајање терминаторске и антитерминаторске петље, које могу да се преклапају у једном делу. На том потенцијалном месту преклапања настаје „прекидач" на основу којег долази до испољавања функције гена који се контролише. Пример једне могуће палиндромске секвенце је дат на слици 3.2

У оквиру *riboswitch* секвенце не мора да постоји искључиво један палиндромски низ, већ супротно. У природи се често дешава да се у оквиру једног *riboswitch*-а налазе два, три, па и више палиндрома. Појављивање већег броја палиндрома узрокује прављење великих петљи (омчи), које физички изгледају као терминатор. Палиндроми праве „стуб" док се на његовом врху налази омча коју дефинишу нуклеотиди између палиндрома. Уколико постоји више угњеждених палиндрома, секундарна структура је доста сложенија. Изглед *riboswitch*-а са више палиндромских секвенци приказан је на слици 3.3.

Слика 3.2: Пример палиндромске секвенце

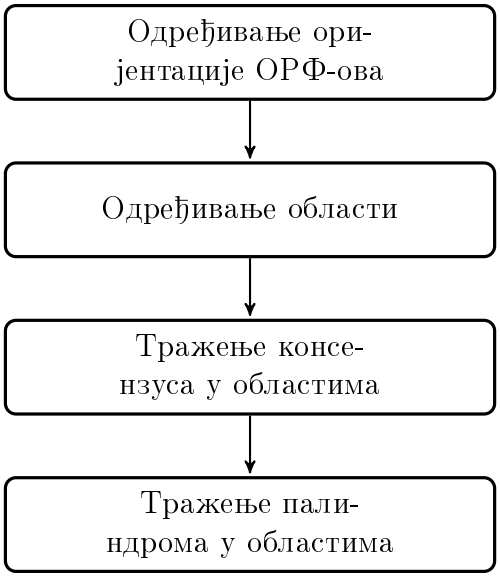


Слика 3.3: Пример *riboswitch* елемента који садржи неколико палиндромских секвенци

3.2 Алгоритам методе

На основу наведених услова за постојање *riboswitch* секвенце у оквиру интергенског региона, развијен је алгоритам који је представљен у оквиру овог поглавља. Алгоритам се састоји из четири корака чији је шематски приказ представљен на слици 3.4, док су сами кораци детаљније објашњени у оквиру поглавља.

Алгоритам је развијен на основу података добијених од стране сарадника ИМГГИ, који су ми обезбедили приступ геному соја *Lactobacillus paracasei subsp. paracasei BGSJ2-8*, који припада дивизији *Firmicutes*, породици *Lactobacillaceae*, роду *Lactobacillus* и врсти *Lactobacillus paracasei*. Како геном наведене бактерије још увек није склопљен, метода је коришћена на свакој од 207 *контига*, где је контига непрекидни скуп преклапајућих ДНК секвенци. У наведеном скупу контига, најкраћа контига је дужине 105, а најдужа 230919 нуклеотида. Укупна дужина контига је око три милиона нуклеотида. Проблем склапања генома из контига није тривијалан и он захтева посебно разматрање, што није тема овог рада.



Слика 3.4: Алгоритам проналажења потенцијалних *riboswitch* секвенци

3.2.1 *Корак 1:* Одређивање оријентације ОРФ-ова

У оквиру првог корака алгоритма врши се одређивање позиција ОРФ-ова, као и њихове оријентације. Оријентације ОРФ-ова су битне јер су оне један од услова за постојање *riboswitch* секвенци.

ИМГГИ је поред 207 контига генома доставио и датотеку са ОРФ-овима, тачније, са нуклеотидним низом који чини ОРФ и информацију у којој се контиги налази. На целом геному укупно је пронађен 2301 ОРФ, за које је требало да се одреде позиције у оквиру контига. Прва идеја је била да се поравнање ОРФ-ова и контига врши *online* уз помоћ *BLAST*[1] програма. Међутим, како је овакав процес поравнања спор, јер посао не може да се аутоматизује, одабран је други начин за поравнање.

Свих 207 контига је смештено у табелу у релационој бази података. Потом су ОРФ-ови пропуштени кроз програм који је написан за ову намену, а који од задате секвенце прави идентичну, али на супротном ланцу. Другим речима, програм је правио комплементарне палиндроме за унете секвенце. Овако обрађени подаци су упоређени са контигама како би се добиле тачне позиције ОРФ-ова у оквиру контига. Дакле, за сваки ОРФ су постојале две ,,идентичне" секвенце и једна од те две је пронађена у оквиру задате контиге. У односу на то који је запис ОРФ-а пронађен одређена је и његова оријентација. Сви добијени подаци су сачувани у посебној табели у оквиру релационе базе података. Структуре коришћених табела су приложене у додатку на посебном компакт диску.

У случају да се ради о произвољном геному за који нису унапред дефинисани ОРФ-ови, могу да се искористе додатни програми, тзв. *ORF finder*-и (срп. ОРФ претраживачи). Ови програми као унос примају геном, или део генома,

[1] *BLAST (Basic Local Alignment Search Tool)* је алгоритам за поређење биолошких секвенци.

а враћају позиције пронађених ОРФ-ова.

3.2.2 *Корак 2:* Одређивање области

У другом кораку алгоритма се на основу оријентација ОРФ-ова одређује које су ,,добре" оријентације и у ком интергенском региону могу да се налазе потенцијалне *riboswitch* секвенце. Могуће оријентације ОРФ-ова, на основу којих се одређују области, приказане су на слици 3.1.

Под термином ,,област" се у овом раду подразумева нуклеотидна секвенца дужине приближно хиљаду нуклеотида у којој постоји могућност да се налази *riboswitch*. Област се формира тако што се од старт кодона ОРФ-а који је ,,добро" оријентисан, узима максимално 500 нуклеотида узводно и максимално 500 нуклеотида низводно. У случају да се старт кодон налази на позицији мањој од 500 од почетка произвољне контиге, област ће да садржи све нуклеотиде са леве и 500 нуклеотида са десне стране. На пример, нека се старт кодон налази на позицији 400 у оквиру произвољне контиге. Формирана област ће бити дужине 900 нуклеотида, јер ће се узети 400 нуклеотида узводно и свих 500 нуклеотида низводно од старт кодона.

Коришћењем описане обраде ОРФ-ова, добијено је 2245 области.

3.2.3 *Корак 3:* Тражење консензуса у областима

Трећи корак алгоритма се заснива на тражењу консензуса фамилија *riboswitch*-ева у добијеним областима. Консензуси који су коришћени у оквиру ове методе су условљени фамилијама *riboswitch*-ева које се претражују у оквиру методе, а које су приказане у табели 1.3.

Произвољан консензус, записан у *Rfam* бази, може да изгледа овако

$$Cgg.aGCCGA.CgG..UAuAGU$$

и потребно је да се он запише на одговарајући начин у табелу базе података како би се лакше извршила обрада података. Велико слово у запису означава највећу вероватноћу да се ту налази баш тај нуклеотид који је записан. Мало слово означава највероватније нуклеотид који је записан, али могуће је да уместо њега стоји и неки други нуклеотид. Тачка у запису консензуса означава да на основу пронађених *riboswitch* елемената у бактеријама није могло да се закључи који се нуклеотид ту налази, већ да сви могу да се појаве са сличном вероватноћом. Такође, потребно је да се напомене да се подаци у *Rfam* бази односе на РНК, тако да је неопходно да се урацил свуда замени са тимином.

Проблем записа консензуса у табелу базе података је решаван у више итерација. У првој итерацији су узети сви консензуси наведених фамилија. Такође, вођено је рачуна о вероватноћама појављивања нуклеотида, односно, да на позицијама где су нуклеотиди представљени малим карактерима може да се налазе и други нуклеотиди. Резултат оваквог записа је био некористан, пре свега због огромне количине података. Из тог разлога су у другој итерацији коришћени консензуси дужи од шест нуклеотида, а сви нуклеотиди записани

малим словом су у бази записани са великим. Другим речима, због смањења скупа могућих решења уведен је строжији услов за претрагу, па није вођено рачуна о вероватноћама приликом записа нуклеотида малим словима. На основу изнетог, произвољан консензус који је наведен у оквиру овог поглавља, при претрази у оквиру ове методе је записан на следећи начин:

$$\%CGG\_AGCCGA\_CGG\_\_TATAGT\%.$$

Записивањем података на наведен начин, у областима је тражен укупно 41 консензус. Тиме је пронађено 929 појављивања консензуса (не у јединственим областима) у 59 различитих контига. Списак тражених консензуса се налази у табели 6.2, која је приложена у додатку.

Уколико је потребно да се изврши претрага генома неке друге бактерије, нове фамилије лако могу да се уврсте у алгоритам. Потребно је само да се у оквиру табеле консензуса додају консензуси жељених фамилија.

3.2.4 *Корак 4:* Тражење палиндрома у областима

Последњи корак алгоритма се састоји од претраге области у којима су пронађени консензуси. У датим областима се траже комплементарне палиндромске секвенце, односно делови *riboswitch* елемената који у одређеним физиолошким условима дефинишу „прекидач". Овај корак алгоритма је такође извршен у неколико итерација.

Поставља се питање за коју минималну дужину палиндрома је потребно извршити претрагу. Иако тачна биолошка граница не постоји, из искуства сарадника ИМГГИ, у природи се петље краће од осам нуклеотида тешко формирају. Такође, не постоји јасна граница ни за дужину размака између два палиндрома. Иако постоје случајеви где је разлика између два палиндрома велика, они се најчешће преклапају чиме формирање једне петље директно онемогућује формирање друге петље. На основу наведених података, програмом за претрагу палиндромских секвенци су претражене све области у којима су пронађени консензуси.

Прво је извршена претрага за палиндромске секвенце минималне дужине осам. На тај начин је пронађено укупно 7648 палиндрома и то у 812 различитих области у којима су већ пронађени консензуси. Како се за вредност минималне дужине палиндрома осам добија превише велики скуп резултата, било је потребно да се поставе оштрији услови претраге. У другој итерацији је извршена претрага палиндрома минималне дужине десет, међутим и то је дало велики скуп потенцијалних решења. На крају, области су претражене за палиндромима минималне дужине дванаест где је добијено 150 палиндрома у 122 области. Упоредни приказ резултата за различите минималне дужине палиндрома дат је у табели 3.1.

Најдужа палиндромска секвенца која је пронађена у оквиру 207 контига је дужине 25, са девет нуклеотида између њих. Такође, пронађен је и велики број области у којима су палиндроми угњеждени са такође релативно малим бројем разлика између њих. Резултат овог корака је 122 области у којима се

Број итерације	**Минимална дужина палиндрома**	**Број палиндрома**	**Број области**
1.	*Осам*	7648	812
2.	*Десет*	711	424
3.	*Дванаест*	150	122

Табела 3.1: Упоредни приказ скупа потенцијалних *riboswitch* елемената у зависности од дужине палиндромских секвенци

налазе палиндроми (за минималну дужину палиндрома од дванаест) и тај број представља и број потенцијалних *riboswitch* секвенци пронађених овом методом. Детаљнија анализа резултата је изложена у следећој глави.

Глава 4

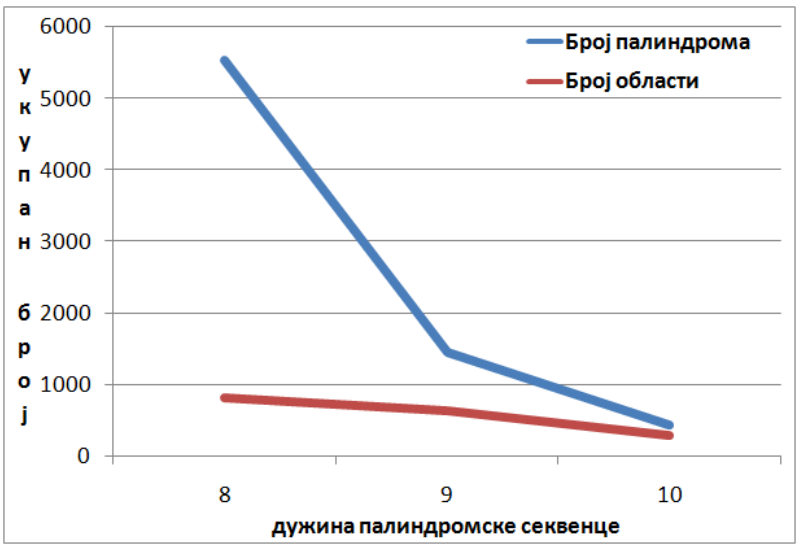
Анализа резултата

Резултати који су добијени применом описане методе су детаљно изложени у оквиру ове главе. На основу 207 контига и 2301 ОРФ-а, одређено је 2245 области које задовољавају задате услове. Добијене области су претражене за 41 консензус и пронађено је 929 појављивања. Последњи корак методе је одредио палиндромске секвенце унутар простора скупа могућих решења и добијено је 122 области, што може да се представи и као 122 потенцијалне *riboswitch* секвенце. Резултати који су добијени у међукорацима су графички приказани на слици 4.1.

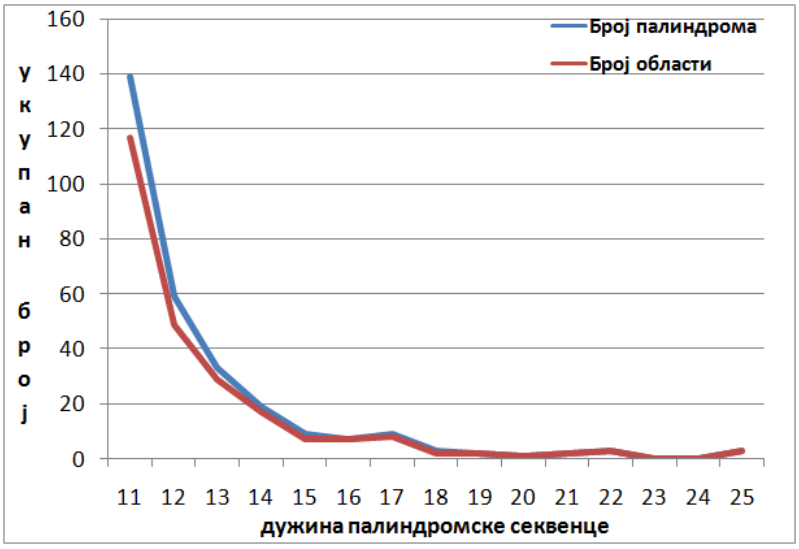
Слика 4.1: Сужавање скупа могућих решења

Крајњи резултат методе чине 122 потенцијалне *riboswitch* секвенце, добијене на основу услова да је минимална дужина тражених палиндрома *дванаест*. Уколико се смањи минимална вредност за дужине палиндрома, потенцијални скуп решења ће да се повећа. Због велике разлике између броја палиндрома и области у односу на дужину палиндрома, графички приказ не даје репрезентативне резултате. Из тог разлога је графикон упоредног приказа броја и дужине

палиндрома подељен на два графика у оквиру ове главе. График над целокупним скупом података је приказан у оквиру додатка, уз табеларни приказ. У оквиру ове главе се на слици 4.2 налазе графикони за палиндроме дужине од осам до десет нуклеотида (слика 4.2а) и за палиндроме дужине од 11 до 25 (слика 4.2б).



(а) Граф 1: за дужине од 8 до 10



(б) Граф 2: за дужине од 11 до 25

Слика 4.2: Упоредни приказ броја палиндрома и броја области у зависности од дужине палиндромске секвенце. *Легенда:* Број палиндрома представља укупан број палиндрома у областима у којима су пронађени консензуси. Број области се односи на оне области у којима је пронађен барем један консензус и барем једна палиндромска секвенца

Најдужа палиндромска секвенца која је пронађена у задатом соју је дужине 25 нуклеотида. Та секвенца се налази у области која је у контиги 96, орф 280 и обухвата нуклеотиде од 119671. до 120671. позиције у контиги. У оквиру

наведене области је пронађен и консензус

$$\_\_AC\_C\_A\_G\_C\_\_\_$$

који припада фамилији *T-box*. Наведена палиндромска секвенца има следећи изглед (подвучени нуклеотиди означавају палиндромску секвенцу): *GAGCATAAGGGCCTTGAATCTAAATGGCTGGGCTCTGGCCATTTAGATTCAAGGCCCTTATGTGT*.

Као један од начина провере веродостојности методе, добијени резултати су упоређени са резултатима који се добијају коришћењем доступних софтверских решења. Претпоставимо да је скуп потенцијалних *riboswitch* секвенци добијених изложеном методом величине 122. Са друге стране, коришћењем постојећих програма добијено[1] је 25 потенцијалних *riboswitch* секвенци. У прилог новој методи говори то да све пронађене позиције добијене коришћењем постојећих програма се налазе у областима које су представљене као резултати ове методе. Овај податак даје велику вероватноћу да се у преосталим областима налазе *riboswitch* секвенце које постојећи програми нису могли да препознају.

4.1 Пример резултата

Резултат који је представљен садржи палиндромску секвенцу дужине 16 нуклеотида. Потенцијални *riboswitch* се налази у оквиру контиге 126, прецизније, око старт кодона ОРФ-а 46. Детаљи ове области су приказани у табели 4.1. У овој области је пронађен консензус за фамилију *Lysine* који је релативно кратак и чини га седам нуклеотида: *CTGATGA*.

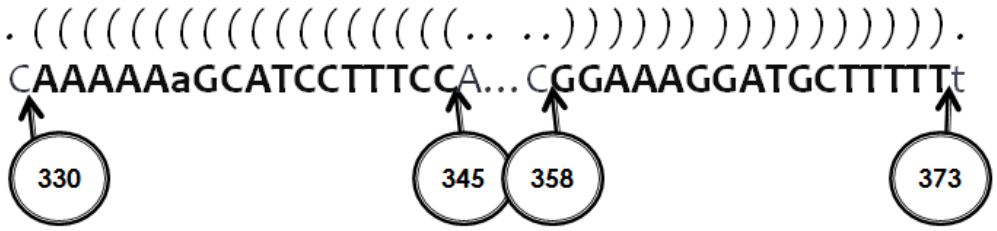
Контига	**ОРФ**	**Позиција старт кодона у контиги**	**Област у оквиру контиге**
126	46	22849	од 22349. до 23349.

Табела 4.1: Параметри за одређивање области - пример

Дужина између ове две палиндромске секвенце је свега 14 нуклеотида, што даје већу вероватноћу за стварно постојање *riboswitch* елемената. Наиме, иако не постоје тачне биолошке дефиниције о размаку између две палиндромске секвенце, на основу јавних база издвојених *riboswitch*-ева, може да се закључи да је тај број обично мали, односно да разлику чини свега неколико нуклеотида. Изглед палиндромске секвенце дужине 16 је дат на слици 4.3.

Овај резултат је добијен на основу сужавања скупа решења, односно постављањем минималне дужине палиндрома на дужину од дванаест нуклеотида. Међутим, након сужавања скупа решења, у свакој од 122 области су потражени сви палиндроми, без обзира на дужину (мора бити барем осам). На овај начин

[1]На основу резултата изложених у мастер раду Стефановић Александра, Математички факултет, Универзитет у Београду, октобар 2011.



Слика 4.3: Изглед палиндромске секвенце дужине 16 у оквиру потенцијалног *riboswitch* елемента - пример

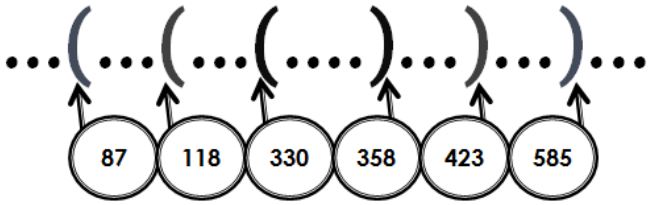
је пронађено чак 11 палиндромских секвенци у оквиру приказане области. Детаљан приказ пронађених палиндрома је представљен у табели 4.2. Позиције свих палиндрома су важне за даље посматрање резултата јер је битно да се уочи где би могли да се формирају ,,угњеждени" палиндроми. За формирање оваквих палиндрома постоји неколико комбинација, а не може да се зна која је права без експерименталне провере. Из тог разлога је за потребе илустрације на слици 4.4 приказана једна могућа комбинација која обухвата три различите палиндромске секвенце. Дужине спољашњих палиндрома су 9 нуклеотида, док је дужина унутрашње палиндромске секвенце 16 нуклеотида.

Секвенца	**Дужина**	**Стартна позиција (у контиги)**	**Стартна позиција палиндрома**
TGTCGATT	8	18	771
CCTTGATG	8	123	163
CGCTGAAA	8	434	687
TGATGACC	8	858	918
TTGCCGGAA	9	87	585
CGGCGCCTT	9	118	423
TCAACATGC	9	172	809
CGTTGTCGAT	10	151	772
TCGTTGTCGA	10	601	773
TGATGATGACCC	12	235	917
AAAAAGCATCCTTTCC	16	330	358

Табела 4.2: Палиндромске секвенце у оквиру потенцијалног *riboswitch* елемента - пример

4.2 Наредни кораци и потврда резултата

Добијени резултати представљају само потенцијалне области у којима се налазе *riboswitch* секвенце. Кораци који би даље требало да се примене на ове резултате, а који не спадају у кораке развијене методе су:



Слика 4.4: Изглед три палиндромске секвенце у оквиру потенцијалног *riboswitch* елемента - пример. Спољашње палиндромске секвенце су дужине девет нуклеотида, док је унутрашња палиндромска секвенца дужине шеснаест нуклеотида.

- Пропуштање резултата кроз неки од РНК *fold* програма,
- Сужавање добијених области на краће нуклеотидне секвенце (тачније лоцирање *riboswitch* елемената),
- Експериментално потврђивање резултата.

Како је у појединим областима пронађено и по два консензуса, потребно је да се одреди да ли је уопште један од њих исправан. Такође, у случају пронађеног већег броја палиндромских секвенци, потребно је да се одреди које од њих улазе у ,,грађу" *riboswitch*-а. Ове информације могу да се добију као резултат РНК *fold* програма. Дакле, као резултат ових програма добијају се структуре петљи и њихове слободне енергије које се изражавају минусом јер се ради о отпуштању, а не коришћењу енергије. Што је добијена слободна енергија већа, већа је и вероватноћа да се таква петља формира у природи. Зато уколико се за неки потенцијални *riboswitch* добије могућност формирања две петље (два палиндромска низа) са великим слободним енергијама, постоји шанса да је у питању прави *riboswitch*.

На добијене податке који су приказани у оквиру ове главе могу да се примене различите методе. Међутим, примена било које рачунарске методе не може да са вероватноћом један потврди постојање *riboswitch* елемената. Зато су сви додатно наведени кораци опциони, јер мора да се изврши функционална провера резултата од стране колега из ИМГГИ. Тек након те провере ће се знати коначни резултати развијене методе.

Глава 5

Закључак

У овом раду је приказан метод за одређивање елемената за контролу генске експресије, тачније *riboswitch*-ева, код бактерија. Иако је метода развијана за геном соја *Lactobacillus paracasei subsp. paracasei BGSJ2-8*, она уз мале измене може да се примени и на геном неког другог соја бактерије. У раду је изнет опис проблема, постојећи начини за решење проблема, детаљан приказ нове методе, као и добијени резултати.

5.1 Значај рада

Овај рад је значајан јер омогућава одређивање позиције потенцијалне *riboswitch* секвенце у новом генетичком материјалу што знатно смањује трошкове и скраћује потребно време у односу на одређивање позиција класичним методама, са једне, и, са друге стране, даје већи скуп потенцијалних кандидата у односу на друге рачунарске методе. Специфичан значај *riboswitch*-ева се посматра у зависности од гена које регулишу. На пример, један од најпознатијих примера *riboswitch*-ева су они који регулишу гене за патогеност код генома *Listeria* и других патогених бактерија. *Riboswitch*-еви у оквиру ових бактерија регулишу гене који обезбеђују могућност за њихову адаптацију код људи и животиња. Патогене бактерије (код којих су *riboswitch*-еви први пут пронађени) укључују вирулентне гене када нема довољно слободних нутритиената, па онда бактерија узима потребне нутриенте од домаћина у коме живи. Додатно, најновији радови говоре и о значају *riboswitch* елемената у фармацеутској индустрији, тј. о њиховој могућој употреби као мета за антибиотску терапију. Предлаже се да се праве нови лекови који би ,,гађали" одговарајуће *riboswitch*-еве тако што би их држали закључане и спречавали вирулентност бактерија.

На основу наведеног значаја *riboswitch* елемената, види се и значај већег разумевања генске експресије и начина њеног контролисања. Добрим резултатима које ова метода постиже, сарадницима из ИМГГИ је омогућено да наведене регулаторне елементе лакше лоцирају, како би могли даље да их изучавају. Дефинисање могућих *riboswitch*-ева чини основу за даљи рад у оквиру њихове Лабораторије. Сама идентификација гена који су регулисани *riboswitch*-евима би могла потенцијално да им укаже на то који су гени овој бактерији важни за

преживљавање у природној средини.

Поред значаја за наведену Лабораторију, овај рад пружа могућност откривања *riboswitch* елемената и у оквиру других генома, што повећава број потенцијалних корисника ове методе.

5.2 Будући рад

Иако је ова метода дала добра решења, постоји места за њено унапређење. У будућности би метода могла да се унапреди на следеће начине:

- Кораци методе би могли да се аутоматизују,
- Да се у оквиру методе дода и програм за идентификацију ОРФ-ова,
- Да се повећа број фамилија које метода претражује,
- Да се не користе готови консензуси, већ да се на основу пронађених *riboswitch* елемената дефинише нови модел (да се не преузима са *Rfam*-а).

Резултати ове методе се налазе на провери у ИМГГИ. Највећи успех и значај овог рада је управо тај што су добијени резултати од користи за ИМГГИ и што ће њихова доступност да смањи потребно време за ручно проналажење *riboswitch* елемената.

Литература

[1] Andreas D. Baxevanis, B. F. Francis Ouellette, *BIOINFORMATICS - A practical guide to the analysis of genes and proteins*, Wiley- Interscience, 2001, ISBN 0-471-38391-0, стр. 67-79

[2] Jean-Michel Claverie, Cedric Notredame, *Bioinformatics for dummies*, Wiley Publishing, 2007, ISBN 0-470-08985-7, стр. 29-104, 146-149

[3] Jin Xiong, *Essential Bioinformatics*, Cambridge University Press, 2006, ISBN-10 0-521-84098-8, стр. 97-123

[4] Michiel Wels, Tom Groot Kormelink, Michiel Kleerebezem, Roland J. Siezen, Christof Francke, An in silico analysis of T-box regulated genes and T-box evolution in prokaryotes, with emphasis on prediction of substrate specificity of transporters: *BMC Genomics*, BioMed Central, 2008, DOI 10.1186/1471-2164-9-330

[5] Payal Singh, Pradipta Bandyopadhyay, Sudha Bhattacharya, A Krishnamachari, Riboswitch Detection Using Profile Hidden Markov Models: *BMC Bioinformatics*, BioMed Central, 2009, DOI 10.1186/1471-2105-10-325

[6] Peggy J. Farnham, Terry Plat. Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription in vitro: *Nucleic acids research 9 (3)*, 1981, PMID 7012794

[7] Peter Bengert, Thomas Dandekar, Riboswitch finder- a tool for identification of riboswitch RNAs: *Nucleic Acids Research*, Vol. 32, 2004, DOI: 10.1093/nar/gkh352

[8] Cei Abreu-Goodger, Enrique Merino, RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements: *Nucleic Acids Research*, Vol. 33, 2005, Web Server issue doi:10.1093/nar/gki445

[9] RFAM baza podataka, `http://rfam.janelia.org`, *Howard Hughes Medical Institute*, poslednji datum izmene: januar 2010, datum pristupa: jul 2011.

Глава 6

Додатак

6.1 Кодне ознаке нуклеинских киселина

Карактерска ознака	**Назив азотне базе**
A	аденин
C	цитозин
G	гуанин
T	тимин
R	$G\ A$ (пурин)
Y	$T\ C$ (пиримидин)
K	$G\ T$ (кето)
M	$A\ C$ (амино)
S	$G\ C$ (јаке везе)
W	$A\ T$ (слабе везе)
B	$G\ T\ C$ (све сем A)
D	$G\ A\ T$ (све сем C)
H	$A\ C\ T$ (све сем G)
V	$G\ C\ A$ (све сем T)
N	$A\ G\ C\ T$ (било која)

Табела 6.1: Најчешћи карактери коришћени за запис азотних база ДНК (*IUPAC- International Union of Pure and Applied Chemistry*)

6.2 Запис коришћених консензуса

Фамилија *riboswitch*-ева	**Оригинални консензус *RFAM***	**Записан консензус**
FMN	*aauuauccU.C.a..G.G.G.Ca.G.GG.U.GA.AAUU*	*%AATTATCCT_C_A__G_G_G_CA_G_GG_T_GA_AATT%*
FMN	*CCC.aAC.CGGCGGUAAaaaaauaaa*	*%CCC__AAC_CGGCGGTAAAAAAATAAA%*
FMN	*aaa.....aAGc*	*%AAA_____AAGC%*
FMN	*CCGCGAGCgauuaa*	*%CCGCGAGCGATTAA%*
FMN	*aaaaaaagucaGcaGA.u.c.cG.G.UGAaAuU*	*%AAAAAAAGTCAGCAGA_T_C_CG_G_TGAAATT%*
FMN	*Cgg.aGCCGA.CgG..UAuAGU*	*%CGG_AGCCGA_CGG__TATAGT%*
FMN	*.CcGG.AUG.g.gA...G..A..g..g*	*%_CCGG_ATG_G_GA___G__A__G__G%*
Lysine	*aauaaaggu.AGAGGu.GCgac.u.uuc.AugAGUA.au..uu.u.ucgG*	*%AATAAAGGT_AGAGGT_GCGAC_T_TTC_ATGAGTA_AT__TT_T_TCGG%*
Lysine	*AGG.....gagugaaaucCgAUGA*	*%AGG_____GAGTGAAATCCGATGA%*
Lysine	*.cga.aa.a.au.GAAAGG...gaa*	*%_CGA_AA_A_AT_GAAAGG___GAA%*
Lysine	*a.gucGCCGAAacaa.a.u.u.gaaauc...cuca.a*	*%A_GTCGCCGAAACAA_A_T_T_GAAATC___CTCA_A%*
Lysine	*uuucaau.u.u..g.uUGGgccu.g.uauuc.GAAuA*	*%TTTCAAT_T_T__G_TTGGGCCT_G_TATTC_GAATA%*
Lysine	*aaua.caggaCUGuCAcaaua..uuuauu.....*	*%AATA_CAGGACTGTCACAATA__TTTATT_____%*
Lysine	*auugUGg.AGuGCUac*	*%ATTGTGG_AGTGCTAC%*
Lysine	*cugauga*	*%CTGATGA%*
Purine	*aaaaaaaaaaaaaaaaucac.u.CgUAUAAu.cccggg.AAUAUGG*	*%AAAAAAAAAAAAAAAATCAC_T_CGTATAAT_CCCGGG_AATATGG%*
Purine	*cccggga..GUUUCUACCaggcaaCC..GUAAA ∼∼∼∼∼*	*%CCCGGGA__GTTTCTACCAGGCAACC__GTAAA%TTGCC%*

Табела 6.2: Списак консензуса коришћених у оквиру методе (1)

Фамилија *riboswitch*-ева	**Оригинални консензус *RFAM***	**Записан консензус**
Purine	*u...G.ACUAcG.agugaaauuau uaaaaau*	*%T__G_ACTACG_AGTGA AATTATTAAAAAT%*
SAM	*uuC.uuAU.C.aAGAG.aGG.c. .GG..AG.GGA..cuGG*	*%TTC_TTAT_C_AAGAG_A GG_C__GG__AG_GGA__CTG G%*
SAM	*.CC..C..uAUGAA...gC*	*%_CC__C__TATGAA___GC%*
SAM	*C..CgGC.AACC..gucauauaa. .....aaa*	*%C__CGGC_AACC__GTCAT ATAA_____AAA%*
SAM	*gacAa..GGUGC.cA.A.uUC.. Cag..Caggaaa*	*%GACAA__GGTGC_CA_A_T TC__CAG__CAGGAAA%*
SAM	*ccuG.Aa.A.GAUaaGaa*	*%CCTG_AA_A_GATAAGA A%*
T − box	*AuAAAaaC.ga.U...GA..Aa.a gG....AAa..A.G..U.Aau.u.u.*	*%ATAAAAAC_GA_T___GA_ _AA_AGG____AAA__A_G__T_ AAT_T_T_%*
T − box	*uuauu.....aaaauu..u..ua..... .....*	*%TTATT_____AAAATT__T_ _TA_____%*
T − box	*cAGAGA...g..c.u.gg...u..GGU u.....GgUGagA*	*%CAGAGA___G__C_T_GG___ T__GGTT_____GGTGAGA%*
T − box	*..ac.c.a.g.c...*	*%__AC_C_A_G_C___%*
T − box	*..a..aaaaa....aaaa...a.a...a.u.G AAa....uucA....Cc.*	*%__A__AAAAA____AAAA___A _A___A_T_GAAA____TTCA____ CC_%*
T − box	*u.u....gGA...Guu.....uc. u.a.u.u.cuGAA..*	*%T_T___GGA__GTT_____ _____TC_T_A_T_T_CTGAA__%*
T − box	*a.....aua aaAGUA*	*%A_____ _____ATAAAAGTA%*
T − box	*ag.aa.u..a.ga.C.G.gu*	*%AG_AA_T__A_GA_C_G_GT%*
T − box	*.....uaauauaa aaAGa...Gg*	*%_____T AATATAAAAAGA___GG%*
T − box	*auaaua..... .....aauu*	*%ATAATA_____ _____AATT%*
T − box	*uauuauc....AAa*	*%TATTATC____AAA%*
T − box	*...UAgGG.UGG.UA...CC.g.C Gaa.....AuaA*	*%__TAGGG_TGG_TA__CC _G_CGAA_____ATAA%*
T − box	*.....uuCG.uCCcUuuu*	*%_____TTCG_TCCCTTT T%*

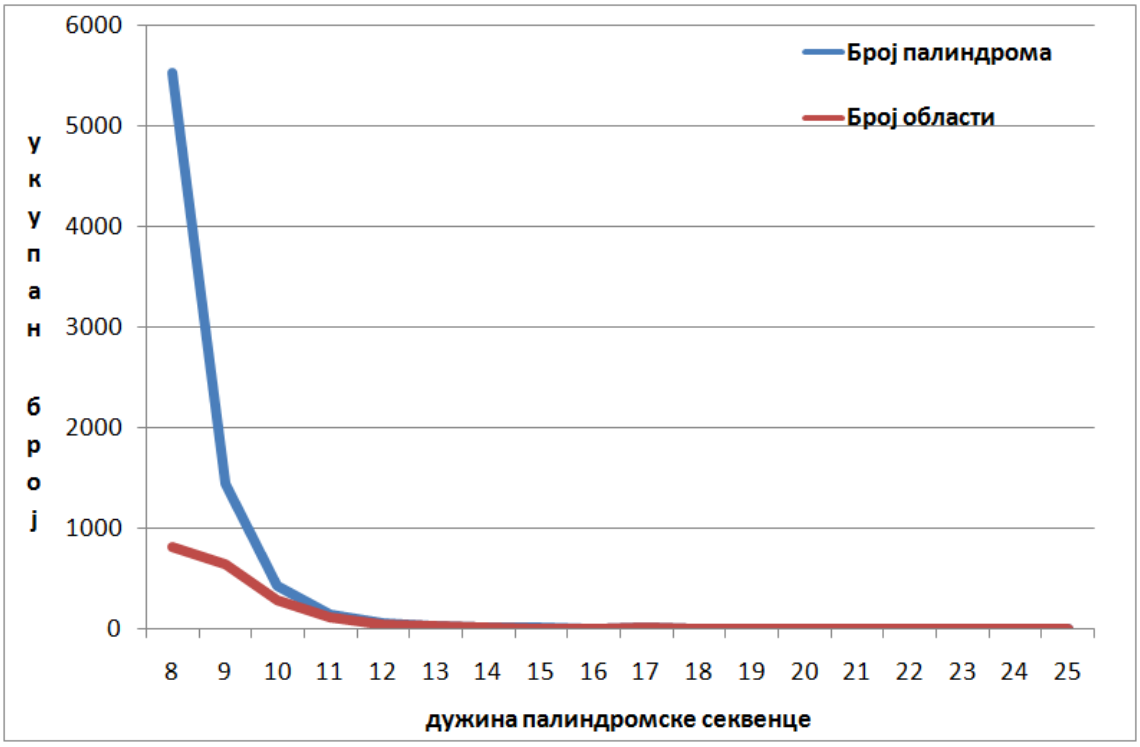
Табела 6.3: Списак консензуса коришћених у оквиру методе (2)

Фамилија *riboswitch*-ева	**Оригинални консензус *RFAM***	**Записан консензус**
TPP	*aaaaacca.c..u.a..g...G.G.GuG Ccccaaa.....*	*%AAAAACCA_C__T_A__G___ G_G_GTGCCCCAAA_____ _____%*
TPP	*ggg.GCUG.AGA.ug.gaagu*	*%GGG_GCTG_AGA_TG_G AAGT%*
TPP	*ccaaA.C.Cc.u....U....u..G.A..... .....ACCUg.A.*	*%CCAAA_C_CC_T____T____T __G_A_____ACCTG_ A_%*
TPP	*..UCc.G.G..u.U.A.AUA*	*%__TCC_G_G__T_T_A_ATA%*
TPP	*CCgG.CG..uAGGGA....ag.ug g.aaaaaaaaaaau*	*%CCGG_CG__TAGGGA____ AG_TGG_AAAAAAAAAAA T%*

Табела 6.4: Списак консензуса коришћених у оквиру методе (3)

6.3 Однос броја палиндрома и броја области у зависности од дужине палиндромских секвенци

6.3.1 Графички приказ броја палиндрома и броја области у зависности од дужине палиндромских секвенци



Слика 6.1: Упоредни приказ броја палиндрома и броја области у зависности од дужине палиндромске секвенце. *Легенда:* Број палиндрома представља укупан број палиндрома у областима у којима су пронађени консензуси. Број области се односи на оне области у којима је пронађен барем један консензус и барем једна палиндромска секвенца

6.3.2 Табеларни приказ броја палиндрома и броја области у зависности од дужине палиндромских секвенци

Дужина палиндрома	Број пронађених палиндрома	Број области у којима су пронађени палиндроми
8	5528	812
9	1445	636
10	422	282
11	139	117
12	59	49
13	33	29
14	19	17
15	9	7
16	7	7
17	9	8
18	3	3
19	2	2
20	1	1
21	2	2
22	3	3
23	0	0
24	0	0
25	3	3

Табела 6.5: Упоредни приказ броја палиндрома и броја области у зависности од дужине палиндромске секвенце. *Легенда:* Број палиндрома представља укупан број палиндрома у областима у којима су пронађени консензуси. Број области се односи на оне области у којима је пронађен барем један консензус и барем једна палиндромска секвенца