SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 903

ANALIZA KODIRAJUĆIH REGIJA U GENOMU

Mirna Bokšić

Zagreb, lipanj, 2009.

Hvala mentoru, prof.dr.sc. Damiru Seršiću na svoj pomoći pri pisanju završnog rada, na svim savjetima i na pozitivnoj energiji.

Hvala mojim biocurama na svim savjetima i druženju u beskrajnim satima provedenim u labosu.

Hvala dr.sc. Mili Šikiću na razumijevanju i odgovorima na bezbroj pitanja.

Sadržaj

1. Uvod

Otkrićem strukture DNA 1950-tih godina razjašnjen je način kodiranja genetičkih informacija u stanici. Poznavanje genoma, načina kodiranja i faktora koji na to utječu u početku je bilo zanimljivo samo biolozima i molekularnim biolozima. Razvojem informatike javlja se jedno novo područje koje se bavi tim problemom, bioinformatika. Bioinformatika povezuje znanja iz biologije i informatiku.

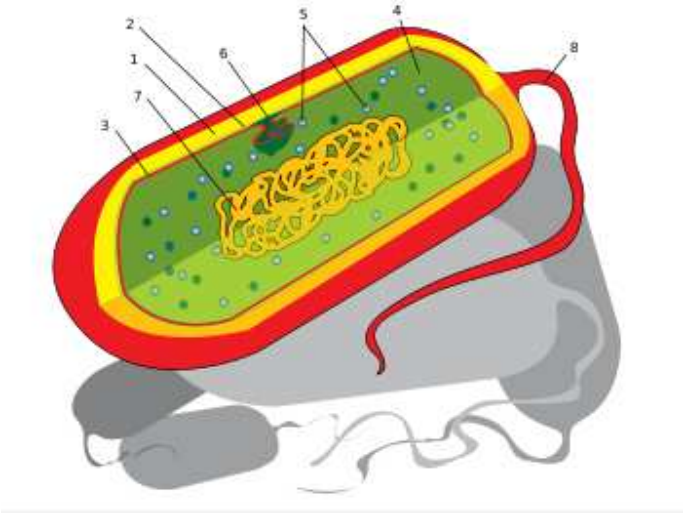
Istraživanje genoma postaje sve popularnije. Molekularni biolozi su otkrili potpuni genom nekih živih vrsta, poput C. Elegansa, ali ipak potpuni genomi većine živih organizama su i dalje nepoznanica. Dostignuća biologa informatičarima su od velike koristi. Informatičari genom ne promatraju kao trodimenzionalnu molekulu, već kao nizove koje čine četiri slova A, T, G i C. Nizovi predstavljaju ulazne podatke koji se obrađuju, pripremaju u oblik potreban statističkim paketima, a zatim statistički obrađuju. Rezultati statističe obrade se analiziraju i na kraju predstavljaju izlaznu informaciju.

Bioinformatika na taj način može pomoći biolozima u istraživanjima, jer smanjuje koštanje eksperimenata, predviđa rezultate istraživanja koja biolozi moraju potvrditi i ubrzava otkrivanje potpunih genoma organizama, a time i razvoj molekularne biologije. Bioinformatičari se bave odeđivanjem mjesta i načina izrezivanja nekodirajućih dijelova gena, pronaleženjem mjesta grananja u intronima, načinom na koji se aminokiseline povezuju u proteine i mnogim drugim stvarima.

2. Biološke osnove

2.1. Prokarioti i eukarioti

Prokarioti su stanična živa bića jednostave građe koja imaju staničnu stijenku i membranu, ali nemaju staničnu jezgru i organele (osim ribosoma). U citoplazmi prokariotskih stanica DNA se nalazi slobodno kao jezgrin ekvivalent, nukleoid. U tim stanicama najčešće se nalazi samo jedan kromosom koji ne sadrži histoproteine kao eukariotske stanice. Kromosom bakterija najčešće se sastoji od samo jedne DNA molekule. Samo neki prokarioti sadrže linearne kromosome. Veličina prokariota je između 0,2 i 700 µm. Prokariotski organizmi su bakterije i modrozelene alge.



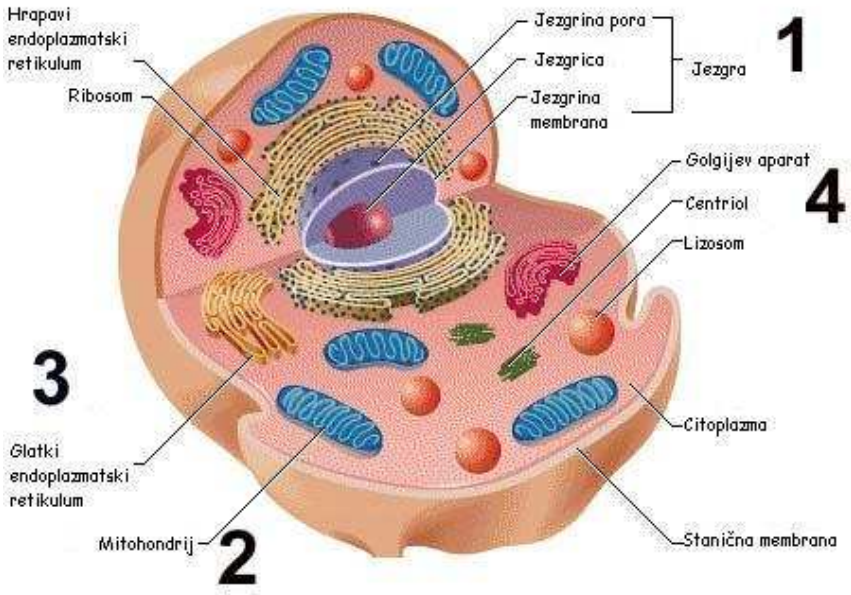
Slika 1. Prokariotski organizam

1- kapsula; 2-stanična stijenka; 3-stanična membrana; 4-citoplazma; 5-ribosomi; 6-mezosomi; 7-nukleoid; 8-flagelum

Kod eukariotskih organizama, odnosno stanica, nasljedni materijal je smješten u jezgri obavijenoj posebnom jezgrinom membranom. U eukariotskoj stanici razvile su se i brojne stanične organele kojih nema kod prokariotskih organizama, među kojima su: endoplazmatiski retikulum, Golgijev aparat, lizosomi i dr.

Eukarioti se dijele u pet carstva: čovjek, životinje, biljke, gljive i protisti. Izuzev protista, svi su prokarioti višestanični organizmi. Eukariotske stanice su uglavnom

puno veće od prokariotskih. Sastoje se od unutrašnjih membrana i struktura, zvanih organeli, i citoskeletona sastavljenog od mikrotubula koji igraju važnu ulogu u definiranju organizacije i oblika stanice. Eukariotska DNK je podijeljena u nekoliko kromosoma, koji su odvojeni mikrotubularnom vretenom tijekom razdiobe jezgre. Jezgra je okružena dvostrukom membranom koja omogućava protok tvari van i unutra, sadrži većinu staničnog genoma, ukupnu nasljednu poruku jednog organizma koja je pohranjenu u kromosomima. Smatra se da je jezgra nastala da bi se molekule DNA razdvojile od citoplazmatskih aktivnosti stanice.



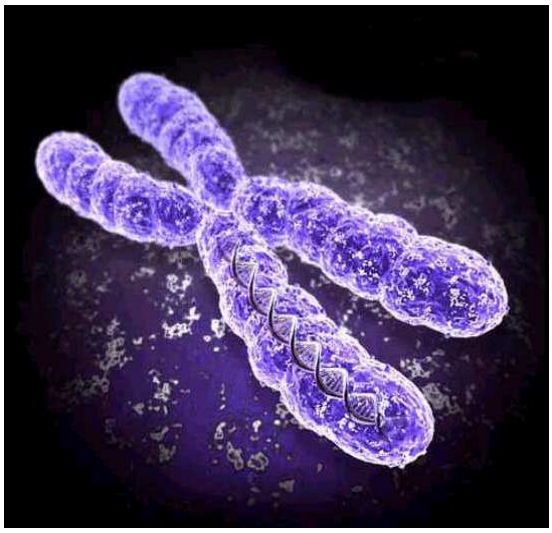
Slika 2. Građa eukariotske stanice

2.2. Kromosomi i kromatin

Genom prokariota sadržan je u jednom kromosomu, koji je obično kružna molekula DNA, dok je genom eukariota sastavljen od više kromosoma, od kojih svaki sadrži linearnu molekulu DNA. Molekula DNA je jako velik, negranajući, linearni polimer, koji može sadržavati mnogo milijuna nukleotida nepravilno, ali ne i nasumce, složenih u slijed. DNA eukariota često je vezana za male bazične protine koji u staničnoj jezgri pravilno pakiraju DNA. Kompleks između eukarotske DNA i proteina je kromatin, koji sadrži dvostruko više proteina nego DNA. Glavni proteini kromatina

su histoni. Osnovna stukturna jedinica kromatina je nukleosom (ponavljajuće jedinice od 200 parava baza). DNA se pakira uz pomoć histona čime nastaje kromatinsko vlakno, što mu skraćuje duljinu za 6 puta. Stupanj kondenzacije kromatina mijenja se tijekom životnog ciklusa stanice.

Nasljedna poruka stanice sadržana je u točno određenom linearnom redosljedu nukleotida u molekuli DNA. Svaka molekula DNA pakirana je u odvojeni kromosom .
Genetički kód, pisan "šifrom" od tri nukleotida (svaki kodon određuje jednu aminokiselinu), jednostavno rješava problem raspoređivanja velike količine nasljednih poruka u mali prostor.



Slika 3. Kromosom

Genom je ukupna nasljedna poruka jednog organizma pohranjena u kromosomima.

2.3. Složenost eukariotsih genoma

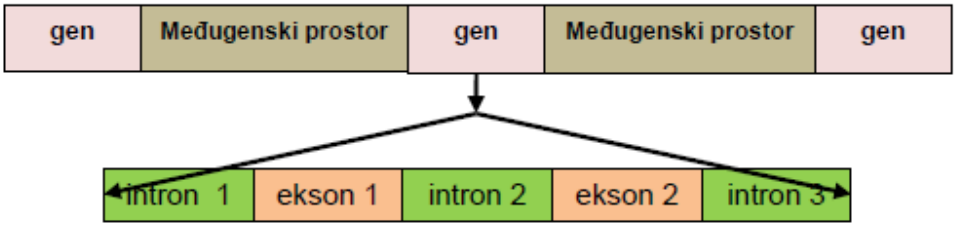
Genomi eukariota su uglavnom složeniji od genoma prokariota, ali ipak veličina genoma nije u srazmjeru s genetičkom složenošću. To je posljedica toga što genomi većine eukariotskih stanica ne sadrže samo funkcionalne gene već i velike količine DNA sljedova koji ne kodiraju proteine.

2.4. Egzoni i introni

Jedna DNA čini više različitih gena međusobno odvojenih nekodirajućim prostorom. Gen je dio DNA koji se pojavljuje u vidu funkcionalnog produkta u obliku RNA ili polipeptida. Dio nekodirajućih DNA su dugi sljedovi koji leže u prostoru između gena, ali i unutar većine eukariotiskih gena nalaze se velike količine

nekodirjućih DNA. Takvi geni imaju podijeljenu strukturu. Kodirajući sljedovi, egzoni, su ispresjecani nekodirajućim sljedovima, intronima. Količina DNA u intronskim sljedovima uglavnom je veća nego u egzonima.

Introni su prisutni u većini gena složenih eukariota, ali ponekada mogu i nedostajati, tako da se može zaključiti da oni nisu neophodni za funkcioniranje gena u eukariotskim stanica. Introni nisu prisutni u nekim jednostavnijim eukariotskim organizimima, ali se mogu pronaći i u nekim prokariotskim. Introni nisu bitni za određivanje sinteze nekog staničnog proizvoda iako postoje neki koji kodiraju funkcionalne RNA ili proteine. Oni igraju važnu ulogu u kontroli genetske ekspresije. Prisutnost intona omogućava da se egzoni mogu spajati u različim kombinacijama, što povećava broj različitih proteina koji se mogu sintetizirati iz pojedinig gena (alteranativno prekrajanje).



Slika 4. DNA i gen

Introni su odigrali važnu ulogu u evoluciji omogućavajući ekosnima različite načine međusobnog povezivanja između gena. To je rezultiralo nastankom novih gena koji omogućavaju nove kombinacije za kodiranje proteina.

2.5. Ponavljajući (repetativni) sljedovi DNA

Introni čine veliki dio genomske DNA, ali još veći dio eukariotskih genoma sastoji se od visoko ponavljajućih nekodirajućih sljedova DNA. Slijedovi koji se u genomu nalaze u više kopija reproducijraju se puno većom brzinom nego slijedovi koji nemaju svoje kopije u genomu. Više od 50% DNA sisavaca sastoji se od visoko repetativnih slijedova. Postoji više vrsta slijedova koji se ponavljaju. Jedna od tih vrsta je i ponavljanje jednstavnih slijedova, koji se ponavljaju nekoliko tisuća puta, a sastoje se od 1 do 500 nukleotida.Takvi se slijedovi često zovu satelitski slijedovi, zbog svoje

odvojesnosti od glavne pruge DNA, a čine oko 10% ukupne DNA. Statelitski slijedovi ne predstavljaju funkcionalnu genetičku informaciju. Ostali ponavljajući slijedovi su raštrkani po genomu, takvi slijedovi najviše doprinose veličini genoma, npr. čine oko polovine ljudsokog genoma. Ti segmenti postoje u genomu zbog vlastite sposobnosti umnožavanja, a ne zbog neke velike koristi.

2.6. Duplikacija gena i pseudogena

Veličini gena doprinosi i to što mnogi geni postoje u više kopija, od kojih su neke nefunkcionalne. Višestruke kopije koriste se u slučajevima proizvodnje RNA ili proteina. S druge strane neke srodne skupine gena mogu se prepisivati u različtim tkivima ili u različitim stadijima razvoja. Neke kopije gena zbog mutacija postaju nefunkcionalne i tako povećavaju veličinu genoma, a ne daju funkcionalni doprinos.
Duplikacija gena može nastati duplikacijom dijela DNA i njenim prijenosom na novu lokaciju ili obrnutim prepisivanjem mRNA, nakon kojeg se nastala cDNA ugrađuje na novo kromosomsko mjesto, pri tome nedostaju introni i nastaje inaktivni dorađeni pseudogen.

2.7. Sinteza i dorada RNA

DNA u genomu ne upravlja sama sintezom proteina, već to čini njen intermedijar RNA. Kada je stanici potreban određeni protein samo se određeni dio velike DNA, smještene u kromosomu, kopira u RNA. Nastale RNA su direktni kalupi za sintezu proteina. Tako da se može zaključiti da protok genetičkih informacija u stanici teče od DNA do RNA i na kraju do proteina – centralna dogma molekuralne biologije.

Usprkos dogmi postoje i određene varijacije u protoku genetičkog sadržaja.
Transkriptna RNA subjekt je mnogih procesnih koraka u jezgri, uključujući izrezivanje pre-mRNA, prije njenog izlaska iz jezgre i nastanka proteina. Ti koraci mogu promijeniti značenje RNA molekule i zato su ključni za razumijevanje kako eukariotske stanice čitaju genom. Za neke gene nastanak RNA je konačni produkt. Te RNA čine točno određene trodimenzonalne strukture koje imaju katalitičke i strukturne uloge u stanici.

Prvi korak u dekodiranju genoma je proces transkripcije uz pomoću kojeg RNA nastaje iz DNA. Zatim slijedi proces obrade RNA i na kraju nastanak proteina.

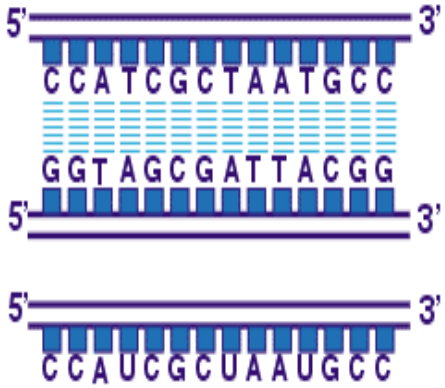
2.8. Od DNA do RNA

Transkripcija i translacija su načini na koje stanica izražava svoje genetičke instrukcije. Mnoge identične kopije RNA mogu nastati iz jednog gena i svaka RNA može upravljati sintezom mnogih istih proteina, pa stanica može brzo sintentizirati velik broj proteina ako su joj potrebni. Ali na svakom genu transkripcija i translacija mogu biti izvršene s različitom učinkovitošću, tako dopuštajući stanici da proizvede veliku količinu jednog proteina, a malu drugih, koji joj u tome trenutku nisu potrebni. Stanica može promijeniti ekspresiju svakog gena ovisno o svojim potrebama u danom trenutku.

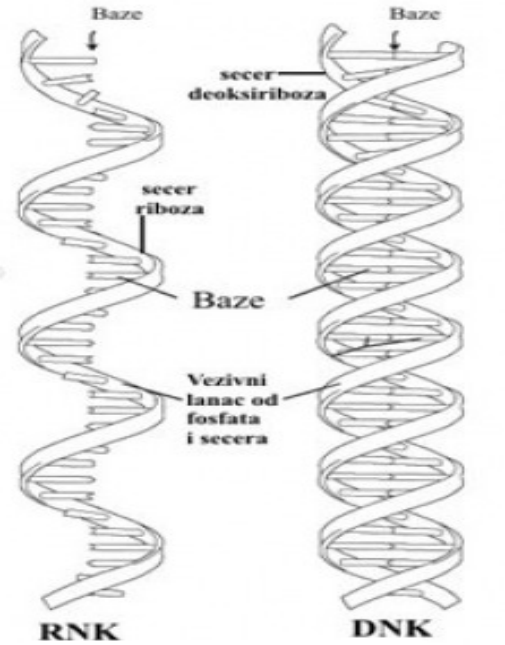
3. Proces transkripcije

Prvi korak koji stanica radi da bi dobila određeni dio genetičke upute je kopiranje dijela DNA, gena, u RNA nukleotidnu sekvencu. Informacija je i dalje napisana istim jezikom kao i kod DNA, zapisana je nukleotidima. Taj se proces zove transkripcija.

Kao i DNA i RNA je linearni polimer, koji čine četiri nukleotida međusobno povezana kovalentnim vezama. RNA se ipak razlikuje u dvije stvari od RNA: nukleotidi u RNA su ribonukleotidi (sadže šećer ribozu, ne deoksiribozu), RNA sadrži adenin (A), guanin (G), citozin i za razliku od DNA umjesto timina (T) RAN sadrži uracil (U). Uracil se povezuje, kao i timin, vodikovim vezama s adeninom, a guanin i citozin se međusobno vežu, kao i u DNA, vodikovim vezama.



Slika 5. Transkripcija



Slika 6. RNA i DNA

Unatoč samo malim kemijskim razlikama, DNA i RNA se stukturno jako razlikuju. DNA se uvijek pojavljuje u obliku dva spiralno povezana lanca, dok RNA sadrži samo jedan lanac. RNA lanac može imati različite oblike. Mogućnost RNA da tvori različite trodimenzionalne oblike omogućava nekim RNA molekulama da imaju strukturne i katalitičke funkcije.

3.1. Stanice proizvode nekoliko tipova RNA

- Glasnička RNA, mRNA- nastaje kopiranjem iz gena, konačni produkt je sama RNA, kod za nastanak proteina. Glasnička RNA služi kao kalup za sintezu proteina.
- snRNA - male RNA koje upravljaju izrezivanjem pre-mRNA , da bi nastala mRNA
- snoRNA - mala nuklearna RNA, koja tvori kemijski modificirane RNA
- rRNA - ribosomske RNA, tvore jezgru ribosoma i kataliziraju sintezu proteina, djelatne u procesu translacije
- tRNA - transfer RNA, ključna pri sintezi proteina kao adaptor između mRNA i aminokisleline
- Druge nekodirajuće RNA - u sintezi telomera, transportu proteina, inaktivaciji X kromosoma

Većinu RNA u stanici čine rRNA, mRNA čini samo 4-5% RNA u stanicama sisavaca.

3.2. Tanskripcijom iz dva lanca DNA nastaje jedan RNA

Sve RNA u stanici nastaju transkripcijom DNA. Transkripcija počinje otvaranjem i razmotavanjem malog dijela DNA, tako se razdvoje lanci DNA. Jedan lanac DNA tada djeluje kao kalup za sintezu RNA. Nukleotidna sekvenca RNA lanca nastaje komplementarnim povezivanjem baza između nukleotida i DNA kalupa. U RNA lanac se dodaje jedan po jedan ribonukleotid i tako nastaje RNA nukleotidna sekvenca komplementarna lancu DNA, koji se koristi kao kalup. Odmah iza dijela gdje se doda jedan nukleotid RNA lancu, RNA lanac se odvaja od DNA, koja se ponovo stvara. RNA je mogo kraća od DNA jer nastaje samo od jednog malog dijela cijele DNA.

Enzimi koji izvršavaju transkripciju su RNA polimeraze. RNA polimeraze kataliziraju formiranje kovalentnih veza koje povezuju nukleotide tako da tvore linearan lanac. RNA polimeraza se kreće korak po korak u DNA, odmotavajući jezgru DNA točno ispred aktivnog mjesta za polimerizaciju da bi otkrili novu regiju kalupa

lanca za komplementarno vezanje baza. Na taj način rastući RNA lanac se produžuje nukleotid po nukleotid u smijeru 5' do 3'.

RNA polimeraza je složen enzim koji se sastoji od više polipeptidnih lanaca. RNA polimeraza katalizira polimerizaciju ribonukleozoid-5'-trifosfata usmjerenu kalupom DNA i može započeti RNA lanac bez početne klice. Umjesto toga transkripcija započinje na specifičnim mjestima na početku gena. Transkripcija ne mora biti toliko točna kao replikacija. RNA ne pohranjuje trajno genetičke informacije u stanici. RNA polimeraza čini grešku na svaki 10^4 kopirani nukleotid, a posljedice pogreške su puno manje nego kod replikacije. Unatoč tome RNA polimeraze imaju mehanizam za popravljanje. Ako se krivi nukleotid doda na rastući lanac RNA polimeraza može izvršiti suprotnu reakciju od polimerizacije i time ispraviti pogrešku.

3.3. Signali kodirani u DNA

Iako se transkripcija u svim stanicama odvija po istom temeljnom mehanizmu, u eukariotksim je stanicama taj proces znatno složeniji nego u bakterijama. Dvije su izrazite razlike između prokariotskog i eukaritskog sustava. U bakterijama se svi geni prepisuju pomoću jedne RNA polimeraze, eukariotske stanice imaju nekoliko različitih RNA polimeraza. Eukariotkse RNA polimeraze ne vežu se izravno na promotorske slijedove nego trebaju interakciju s nizom dodatnih proteina da bi započele transkripciju. To povećava složenost eukariotske transkripcije, olakšava regulaciju genske ekspresije potrebne za usmjeravanje aktivnosti velikog broja različitih stanica višestaničnom organzmu.

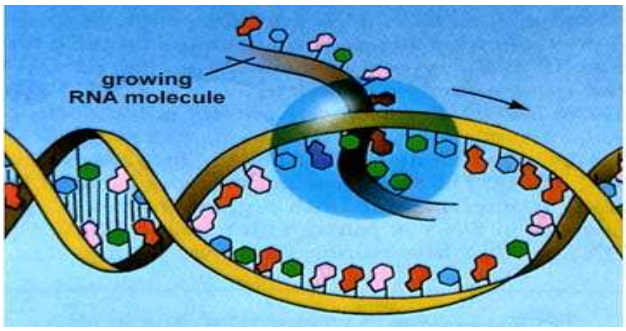
3.4. Transkripcija kod bakterija

Kod transkripcije prokariotksih stanica postoji samo jedna RNA polimeraza. Da bi se izvršila točna transkripcija gena, RNA polimeraza mora prepoznati gdje početi, a gdje završtiti proces transkripcije. RNA polimeraza veže se u regiju DNA koja se zove promotor, koji predstavlja posebnu sekvencu nukleotida koja označava početak sinteze RNA. RNA polimeraza prepoznaje tu regiju uz pomoć kontakta s dijelovima baza koje su izložene s vanjske strane lanca. Nakon što se RNA polimeraza veže za promotor, ona otvara lance DNA i otkriva kratki odsječak nukleotida na svakom lancu.

Odmotavanjem DNA, jedan od dva lanca se ponaša kao uzorak za komplementarno sparivanje s nadolazećim ribonukleotidima, od kojih se dva spajaju uz pomoć polinukleaze na početak RNA lanca. Nakon sinteze prvih desetak nukleotida RNA lanca sigma faktor se razdvaja od polimeraze , za vrijeme toga u polimerazi se događaju strukturne promjene koje joj omogućavaju da se nastavi brzo kretati po jednom lancu DNA. Prevođenje prestaje kad polimeraza u DNA lancu dođe do znaka stop, koji čine određene kombinacije nukleotida.

3.5. Signali početka i kraja transkripcije kod eukariota

Poroces početka i kraja transkripcije uključuje seriju kompliciranih strukturnih izmjena u proteinu, DNA i RNA molekulama.U jezgri eukatiotskih stanica postoje tri različite vrste polimeraze: RNA polimeraza I, RNA polimeraza II, RNA polimeraza III. Sve tri polimeraze su strukturno jednake. One dijele neke uobičajene podjedinice i mnoge strukturne osobine, ali one vrše transkripciju na različitim vrstama gena. RNA polimeraza I i III koriste se kod transkripcije ribosomne RNA i nekih malih RNA molekula. RNA polimeraza II prevodi većinu gena, uključujući i one koji kodiraju proteine.

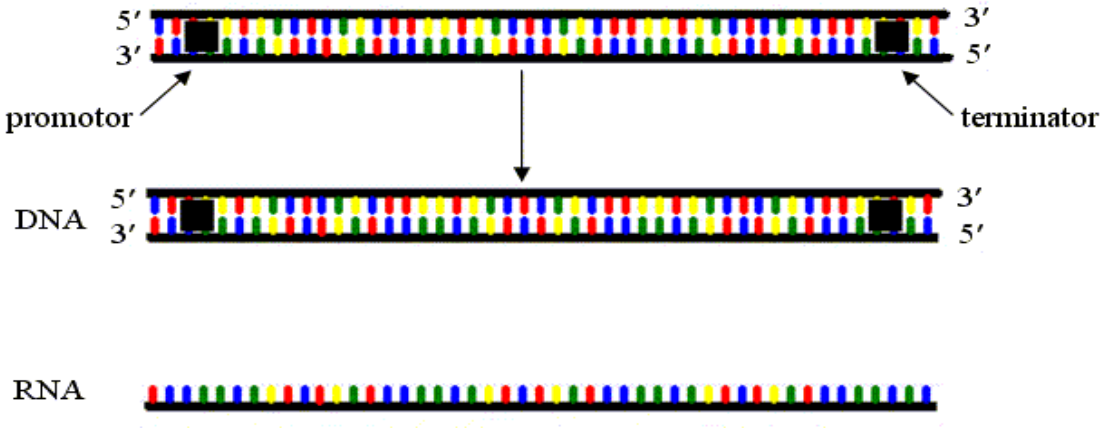


Slika 7. Vezanje RNA polimeraze za DNA i stvaranje lanca RNA

Proteini poznati kao transkripcijski aktivatori se vežu na posebne sekvence DNA i pomažu u privlačenju RNA polimeraze II na mjesto početka transkripcije. To je potrebno za pomoć RNA polimerazi i transkripcijskim faktorima za prevladavanje poteškoća vezanja na DNA koja se nalazi u kromatinu. Inicijacija transkripcije *in vitro* zahtjeva i prisustvo medijatora, koji omogućava aktivatorima da komuniciraju s

polimerazom II i transkripcijskim faktorima. Transkripcija u stanici često zahtjeva i ojačanje uz pomoć enzima koji modificiraju kromatin. Ti enzimi su korepresori.

RNA polimeraza ne može sama započeti transkripciju (kod eukariota), već ona zahtjeva pomoć mnogih proteina, transkripcijskih faktora (TFI), koji se zajedno s polimerazom spajaju na promotor prije početka transkripcije. Transkripcijski faktor pomaže RNA polimerazi da se veže točno za promotor, pomaže u razdvajanju lanaca DNA i oslobađa RNA polimerazu od promotora nakon početka transkripcije. TFII faktor ima istu funkciju kao i sigma faktor kod bakterija. Proces spajanja započinje povezivanjem TFIID na kratku sekvencu dvolančane DNA, koju primarno čine A i T nukleotidi. Ta sekvenca je poznata kako TATA slog (a prepoznaje ju podjedinica TFII, TBP). TATA slog se tipično nalazi 25 nukleotida uzvodno od mjesta početka transkripcije. Vezanje TFIID uzrokuje veliku deformaciju TATA sloga . Distorzija služi kao fizički znak lokacije aktivnog promotora u jako velikom genomu.



Slika 7. Promotor i terminator u DNA

Nakon što je polimeraza II našla promotor, ona na početku transkripcije mora dobiti pristup kalupu, u tome joj pomaže TFII. Polimeraza II ostaje na promotoru sintetizirajući male RNA sve dok ne dođe do promjena u građi i tada se oslobađa i počinje prepisivanje gena.

Nakon što je polimeraza II započela transkripciju većina transkripcijskih faktora se oslobađa od DNA kako bi bili slobodni za inicijaciju transkripcije na nekom drugom mjestu.

Nakon početka transkripcije RNA polimeraza II ne prolazi glatko duž lanca DNA molekule već neke dijelove prolazi brže, a neke sporije. Produljenje lanca RNA uz pomoć polimeraze potpomognuto je i mnogim elongacijskim faktorima, proteinima koji smanjuju mogućnost da će RNA polimeraza disocirati prije nego što dođe do kraja gena. Ti se faktori udružuju s RNA polimerazom kratko nakon početka procesa i pomažu polimerazi u kretanju duž različite DNA sekvence. Eukariotske polimerate također se moraju boriti i sa strukturom kromatina dok se kreću duž lanca DNA.

Svaki protein koji se kreće duž DNA lanca duple uzvojnice ima tendenciju stvaranja napetosti. U eukariotskim stanicama topoizomeraze brzo uklanjaju tu napetost.

4. Obrada RNA

U bakterijama ribosomi imaju izravan pristup molekuli mRNA u nastajanju i translacija kod bakterija započinje na nastajućem lancu mRNA dok još traje transkripcija. U eukariotima, mRNA sintetizirana u jezgri mora se prvo transportirati u citoplazmu, prije nego se može upotrijebiti kao kalup za sintezu proteina. Početni produkti transkripcije u eukariotima prvo se moraju modificirati prije izlaska iz jezgre. Dorada obuhvaća modifikaciju obaju krajeva promatrnog transkripta i uklanjanje intorna. Reakcije dorade povezane su s transkripcijom.

Modifikacije kraja eukariotske mRNA su dodavanje kape na 5' i na 3' kraj. Dodavanje tih kapa omogućava stanici ocjenjivanje jesu li prisutna oba kraja molekule prije izvoza RNA iz jezgre u citoplazmu, gdje dolazi do translacije u protein. Procesom izrezivanja RNA dolazi do povezivanja različith sekvenci koje kodiraju protein, tako se omogućava i sinteza različitih proteina iz istog gena.

4.1. Dodavanje kape na RNA

Nakon što RNA polimeraza II proizvede oko 25 nukleotida RNA 5' kraj se modificira dodavanjem 7-metilgvanozinske kape koja sadrži modificirani guanin. Dodavanje kape je reakcija koju omogućavaju tri enzima: jedan pomiče jedan fosfat s 5' nastajuće RNA , jedan dodaje GMP u obrnutom smijeru, a treći dodaje metilnu skupinu na G bazu i na ribazu jednog ili dva nukleotida na 5' kraj RNA lanca. Zato što se sva tri enzima vežu za fosforizirani rep polimeraze II, oni mogu modificirati 5' kraj nastajućeg transkripta čim se on odvoji od polimeraze. 5' kapa stabilizira RNA i poravna eukariotske mRNA s ribosomima tijekom translacije.

Metalna kapa signalizira 5' kraj eukariotske mRNA. Taj signal pomaže stanici da razlikuje mRNA od drugih RNA u stanici, jer RNA polimeraza I i III transkripcijom proizvode RNA bez kapa, jer im nedostaju repovi. U jezgri kapa veže proteinski kompleks CBC, koji pomaže RNA da bude dobro obrađena i da izađe iz jezgre u citoplazmu.

3' kraj većine eukariotskih mRNA nije definiran zaustavljanjem transkripcije, nego kidanjem primarnog transkripta i dodavanjem poli-A-repa-reakcijom poliadenacije. Signali za poliadenilaciju obuhvaćaju visoko konzervirane heksanukleotide (AAUAAA u stanicama sisavaca), smještene 10 do 30 nukleotida uzvodno od mjesta poliadenilacije, i G-U bogati nizvodni sljedni element. Neki geni imaju U-bogate sljedne elemente uzvodno od AAUAAA. Te sljedove prepoznaje proteinski kompleks koji obuhvaća endonukleaze što kidaju lanac RNA i odvojenu poli-A-polimerazu i mogu putovati s polimerazom skroz do mjesta započinjanja transkripcije. Kidanje i poliadenilacija signaliziraju zaustavljanje transkripcije do kojeg obično dolazi nekoliko stotina nukleotida nizvodno od mjesta dodavanja poli-A.

Gotovo su sve mRNA u eukariotima poliadenilirane, a poli-A rep, pokazano je, regulira i translaciju i stabilnost mRNA. Poliadenilacija igra važnu regulatornu ulogu u ranom razvoju, gdje promjene duljine poli-A repova kontroliraju translaciju mRNA.

4.2. Izrezivanje introna prekrajanjem iz pre-mRNA

Kodirajuće sekvence koje kodiraju proteine kod eukariota uglavnom su ispresjecane nekodirajućim sekvencama. Kodirajući egzoni uglavnom su manji od nekodirajućih introna. I egzoni i introni se prevode u RNA. Introni se izrezivanjem uklanjaju iz novonastalih pre-mRNA. Najvažnija stvar kod izrezivanje RNA je to što time nastaje mRNA. Do izrezivanja dolazi odmah nakon dodavanja kapa na 5' i 3' kraj pre-mRNA.

U svakom procesu izrezivanja uklanja se samo jedan intron, nakon čega se međusobno povezuju dva egzona. Mehanizam koji katalizira izrezivanje RNA je složen. Sastoji se od 5 dodatnih mRNA molekula i preko 50 proteina, a pri tome hidrolizira monogo ATP-a. Ta složenost je potrebana da bi se osiguralo da se izrezivanje vrši točno, dok je u isto vrijeme dovoljno fleksibilno, zbog mnoštva različitih introna koji se nalaze u stanici. Česte pogreške u izrezivanju RNA jako bi štetile stanici i rezlutirale bi pogrešnim radom proteina. Ako i dođe do pogrešaka u stanici se nalaze mehanizmi koji uklanjaju te pogreške.

Možda se čini beskorisnim uklanjati velike količine introna izrezivanjem RNA. Veze intron-ekson omogućavaju nastanak novih, korisnih proteina. Prisustvo mnogih introna omogućava genetsku rekombinaciju povezivanjem egzona različitih gena. Transkripti monogih eukariotskih gena se izrezuju na mnoštvo načina i tako proizvode različite RNA , tako omogućuju nastanak različitih proteina iz istog gena, alternativno prekrajanje. Izrezivanje omogućava eukariotima da povećaju već i tako enormno velik potencijal kodiranja.

4.3. Nukleotidne sekvence signaliziraju mjesto početka izrezivanja

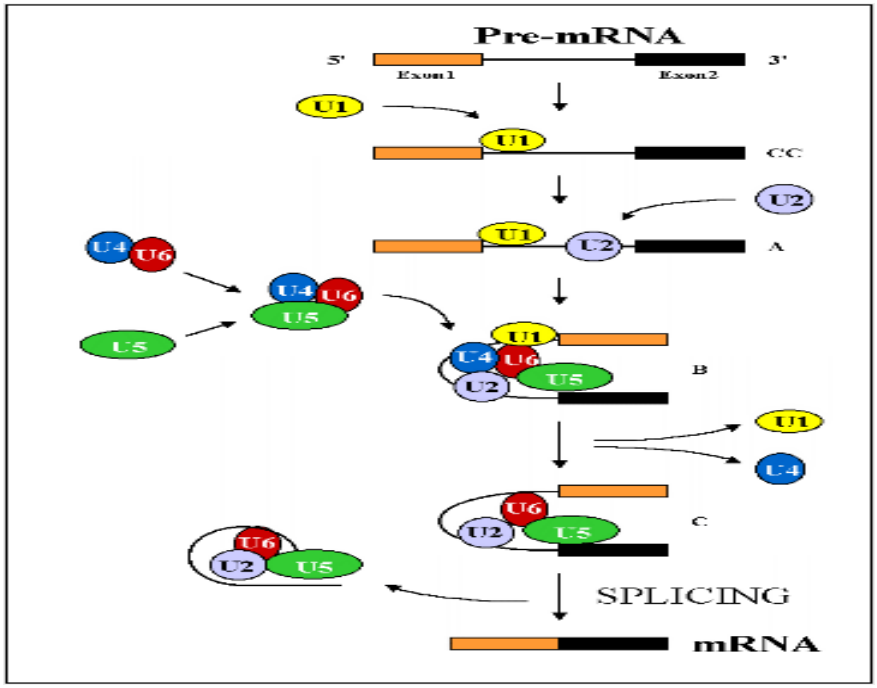
Introni imaju raspon količine nukleotida od 10 do 100000. Određivanje granica između introna i eksona predstavlja veliki problem, koji postaje još veći kada se uzme u obzir i mogućnost alternativnog prekrajanja. Izrezivanje intorna iz RNA uključuje tri pozicije u RNA: 5' mjesto izrezivanja, 3' mjesto izrezivanja i 2' koji čini bazu lasa. U izrezvanju pre-mRNA svaki od ta tri mjesta ima nukleotidnu sekvencu koja je ista od introna do introna, koja govori stanici gdje izrezivanje mora početi. Ali postoje mnoge varijacije te sekvence u svakoj stanici, koje nam tako otežavaju pronalazak točnog mjesta izrezivanja.

Prekrajanje mRNA se odvija u dva koraka. Prvo se mRNA kida na 5' mjestu za prekrajanje i 5' kraj introna spaja se s adeninskim nukleotidom unutar introna (blizu 3' kraj). U tom koraku nastaje veza između 5' kraja introna i 2' hidroksilne skupine adeninskog nukleotida. Nastali intermedijar je sličan lasu u kojem intron pravi omču. Drugi korak u prekrajanju odvija se istodobnim kidanjem na 3' mjestu za prekrajanje i sljepljivanjem dvaju egzona. Intron je tako isječen u obliku lasa, potom je lineariziran i razgrađen u jezgri cjelovite stanice.

4.4. Izrezivanje RNA vrši se u tjelešcima za prekrajanje

Tjelešca za prekrajanje su građena od proteina i malih nuklearnih RNA. Za razliku od drugih koraka nastanka RNA, izrezivanje u velikom dijelu vrše RNA molekule, a ne proteini. RNA molekule prepoznaju vezu intron-ekson i sudjeluju u izrezivanju. Te RNA molekule su relativno kratke (50-200 nukleotida) i mogu se svrstati u pet skupina malih nuklearnih RNA (snRNA) nazvanih: U1, U2, U4, U5 i U6.

Zovu se snRNA (male nuklearne RNA) i povezju se s najmanje sedam, a najviše deset proteinskih podjedinica i tako tvore snRNP (male nuklearne ribonukleoproteinske čestice). Te snRNP tvore jezgru tjelešca za prekrajanje, veliki skup RNA molekula i proteina koje vrše izrezivanje pre-mRNA u stanici. U1, U2 i U5 snRNP svaka sadrži jednu snRNA molekulu, a U4 i U6 su međusobno združene i jednu snRNP.

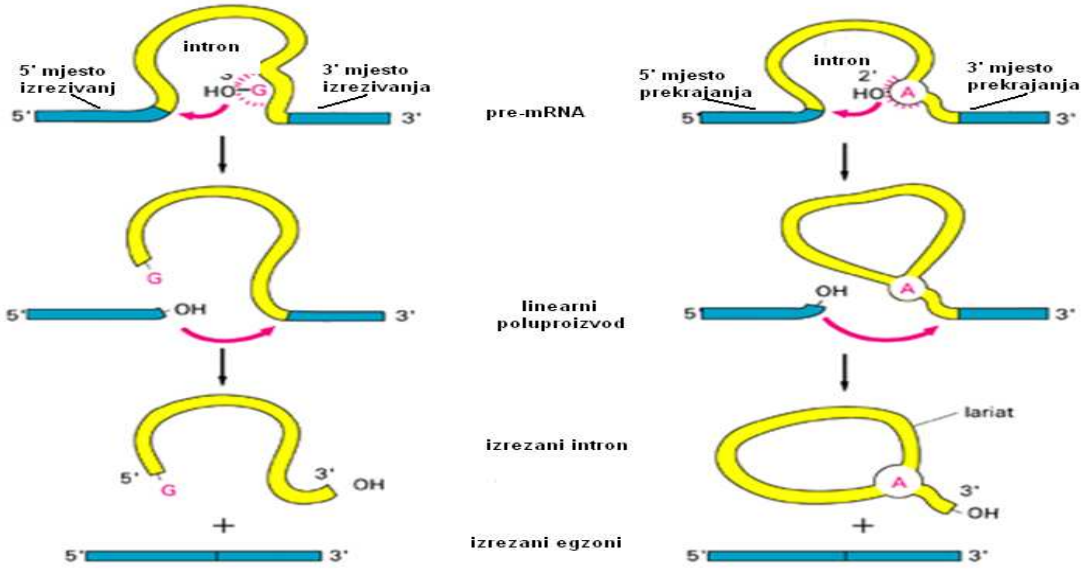


Slika 8. Proces prekrajanja

Izrezivanje introna počinje vezanjem U1 snRNP na 5' mjesto za prekrajanje pre-mRNA. Prepoznavanje 5' mjesta za prekrajanje obuhvaća sparivanje baza između 5' usaglašenog slijeda za prekrajanje i komplementarnog slijeda na 5' kraju U1 snRNA. Nakon toga dolazi U2 koja se veže na točku grananja komplementarnim sparivanjem između U2 i točke grananja. Već nastali kompleks, koji čine U4 /U6 i U5 snRNP, ugrađuje se u tjelešca za prekrajanje, pri tome se U5 veže na slijed uzvodno od mjesta prekrajanja. Reakcija prekrajanja povezana je s preslagivanjem snRNA. Disocijacija U6 od U4 i premještanje U1 na 5' mjesto za prekrajanje predhode prvom koraku u procesu prekrajanja (nastaje intermedijar sličan lasu). Zatim slijedi vezanje

U5 na 3' mjesto za prekrajanje, a zatim slijedi izrezivanje introna i povezivanje egzona.

snRNA prepoznaju mjesta prekrajanja i izravno kataliziraju reakciju prekrajanja. Katalitička uloga RNA vidi se u rekaciji samoprekrajanja nekih RNA. Takve RNA mogu katalizirati uklanjanje vlastitih introna u odsutnosti drugih proteina ili RNA faktora. Prekrajanje je katalizirano intronom, koji djeluje kao ribozim upravljajući izrezivanjem samog sebe iz pre-mRNA.



Slika 9. Samoprekrajajući introni I i II

Samoprekrajajuće RNA mogu se naći u mitohondrijima, kloroplastima i bakterijama. Samoprekrajajuće RNA mogu se svrstati u dvije skupine na osnovi reakcijskog mehanizma samoprekrajanja. Prvi korak u prekrajanju skupine I introna je kidanje na 5' mjestu za prekrajanje posredovano gvanozinskim kofaktorom. 3' kraj slobodnog egzona potom reagira s 3' mjestom za prekrajanje da bi se izrezao intron kao linearna RNA. Samoprekrajajuća reakcija skupine II slična je nuklearnom prekajanju mRNA. Do kidanja 5' mjesta dolazi zbog napada adenozinskog nukleotida u intronu, pri tome nastaje proizvod sličan lasu koji se izrezuje.

Može se zaključiti da su aktivne katalitičke komponente tjelešca za prekrajanje male RNA, a ne proteini. Istraživanja su pokazala da U2 i U6 snRNA mogu katalizirati prvi korak prekrajanja u odsutnosti proteina.

Mnogi proteinski faktori prekrajanja, koji nisu snRNP, igraju važnu ulogu u povezivanju tjelešca za prekrajanje, posebno u pronalasku pravog mjesta za prekrajanje. Introni često sadrže više slijedova koji odgovaraju mjestu za prekranje. Zbog toga sustav za prekrajanje mora biti sposoban pronaći pravo mjesto prekrajanja. Prekrajajući faktori služe za usmjeravanje tjelešca za prekrajanje na korektno mjesto prekrajanja vezanjem na specifične RNA slijedove unutar egzona, pridružujući U1 i U2 na odgovarajuće mjesto na pre-mRNA interakcijama protein-potein. Oni povezuju i prekrajanje s transkripcijom udružujući se s fosforiliranom CTD RNA-polimerazom II.

4.5. Alternativno prekrajanje

Većina pre-mRNA sadržava brojne introne, različite mRNA mogu nastati iz istog gena različitim kombinacijama 5' i 3' mjesta za prekrajanje. Mogućnost da se egzoni spoje međusobno u raznim kombinacijama omogućava da se brojne mRNA mogu generirati iz iste pre-mRNA. Taj se proces zove alternativno prekrajanje, a događa se često u genima složenih eukariota. Alternativno prekrajanje značajno povećava raznolikost proteina.

Alternativno prekrajanje može biti različito u različitim tkivima, ono osigurava važan mehanizam za specifičnu tkivnu i razvojno reguliranu ekspresiju. Regulira se aktivatorima. Aktivatori pridružuju faktore prekrajanja mjestu prekrajanja.

4.6. Uređivanje RNA

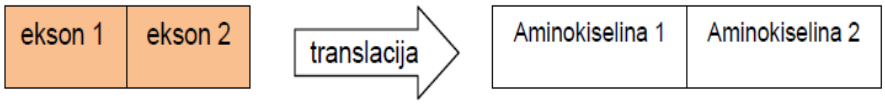
Neke se mRNA modificiraju u procesu dorade tako da se mijenja slijed aminokiselina u proteinu koji je njima kodiran. Uređivanje mitohondrijskih mRNA u nekim praživotinjam obuhvaća dodavanje i brisanje uridinskih ostataka na više mjesta u molekuli. Drugi oblici uređivanja mRNA u biljkama i stanicam sisavaca obuhvaćaju modifikaciju specifične baze.

4.7. Razgradnja RNA

Introni se razgrađuju u jezgri, a nenormalne mRNA, kojima nedostaje potpun otvoreni okvir čitanja, eliminiraju se raspadom posredovanim nesmislenim mRNA. Funkcionalne mRNA u eukariotskim stanicama razgrađuju se na više načina, osiguravajući tako dodatne mehanizme za kontrolu genske ekspresije. U nekim slučajevima, brzina razgradnje mRNA regulirana je signalima izvan stanice.

5. Translacija

Nakon transkripcije i dorade RNA slijedi proces translacije. Translacija predstavlja proces sinteze proteina koja se odvija prema kalupu mRNA. Sinteza proteina se smatra krajnjim korakom ekspresije gena, a translacija je tek prvi korak nastanka funkcionalnog proteina. Nakon sinteze polipeptidni se lanac mora smotati u odgovarajuću trodimenzionalnu konfiguraciju, a pri tome često podliježe različitim obicima dorade.



Slika 10. Proces tranaslacije

Mehanizmi koji kontroliraju aktivnosti proteina u stanici su jako važni za rad stanice. Jednom sintetizirani , mnogi proteini u odgovoru na izvanstanične signale mogu biti regulirani kovalentnim regulacijama i povezivanjem s drugim molekulama. Razine proteina se reguliraju različitom brzinom razgradnje proteina.

5.1. Translacija mRNA

Transportna RNA služi kao posrednik koji smješta aminokisleine na kalup mRNA. Aminoacil-tRNA-sinteze vežu aminokisleine na odgovarajuće tRNA. Koje se zatim putem komplementaronog sparivanja baza vežu na kodone mRNA.

Ribosom se sastoji od dviju podjedinica, koje su izgrađene od proteina i ribosomnih RNA. Stvaranje peptidne veze primarno je katalizirano ribosomnom 23S RNA.

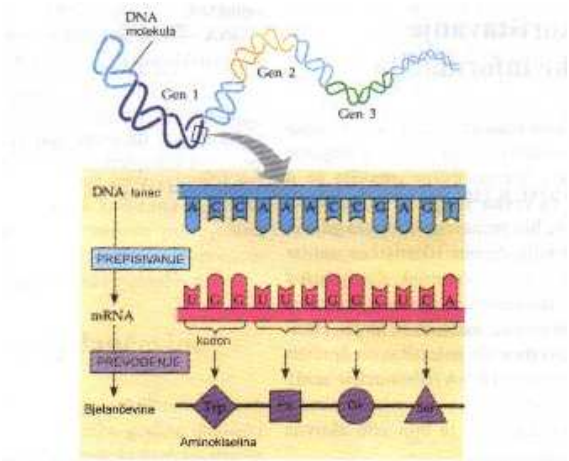
Translacija prokariotskih i eukariotskih mRNA započinje metioninskim ostatkom. Kod bakterija, inicijacijskom kodonu prethodi slijed koji smješta mRNA na ribosom putem sparivanja baza sa 16S rRNA. Kod eukariota, inicijacijski kodon se

pronalazi pretraživanjem mRNA s 5' kraja, a prepoznaje se na temelju njegove 7-metilgvanozinske kape.

Translacija započinje vezanjem metionil-tRNA i mRNA na malu ribosomnu podjedinicu. Zatim se komplesku pridružuje velika ribosomna podjedinica, te se polipeptidni lanac produžuje sve dok ribosom ne stigne do terminacijskog kodona na mRNA. Za odvijanje inicijacije, elongacije i terminacije translacije i kod prokariota i kod eukariota nužna je prisutnost različitih neribosomnih faktora.

Regulacija specifičnih mRNA može se postići vezanjem represorskih proteina te proteinima koji usmjeravaju mRAN u specifično područje u stanici. Kontrolirana poliadenilacija mRNA također je važan mehanizam regulacije translacije tijekom rane faze razvoja. Translacija nekih mRNA kontrolirana je nekodirajućim RNA koje RNA interferencijom dovode do razgradnje homolognih mRNA. Konačno, aktivnosti translacije u stanici može općenito biti regulirana modifikacijom inicijacijskih fakora.

Na kraju možemo zaključiti da je proces prenošenje genetičke informacije od DNA do gena složen i zahtjeva mnogo enzima, katalizatora, proteina i drugih molekula koje ga potiču i ubrzavaju. Sve te radnje mogu se na kraju svesti na slijedeći crtež:



Slika 11. Prenošenje genetičke informacije od DNA do proteina

6. Genom Caemprhabditis Elegansa

Genom C. Elegansa relativno je jednostavan animalni genom, koji je važan za analizu genoma općenito. Genom C. Elegansa se koristi pri proučavanju animalnog razvoja. Dug je $97x10^6$ parova baza, a sadrži oko 19 000 kodirajućih slijedova za proteine. Geni mu se protežu na oko 5 kb i sadržavaju prosjećno 5 intorna. Slijedovi za kodiranje proteina tako iznose samo oko 25% genoma.

Proteini predviđeni u C. Elegansu pokazuju značajnu sličnost s poznatim proteinima drugih organizama. Postoje značajne sličnosti između proteina C. Elegansa i čovjeka, koje su čak značajno veće nego između C. Elegansa i bilo kojeg kvasca ili bakterije. Proteini koji su isti C. Elegansu i bakterijama djeluju u osnovnim staničnim funkcijama, poput metabolizma, udvostručavanja DNA, prepisivanja , prevođenja i razvrstavanja proteina. Postoji mogućnost da te gene dijele sve eukariotske stanice. Razumijevanje ovih gena posebno je zanimljivo, jer ako ih razumijemo kod crva, lakše ćemo ih moći naći i razumjeti kod drugih eukariotskih organizama, pa čak i čovjeka.

U genomu C. Elegansa mnogi su geni duplicirani, tako da je broj jedinstvenih gena između 8000 i 9000. Dakle samo 8000 do 9000 gena nosi genetički bitne informacije, dok su ostali replike tih gena ili neki drugi genetički nebitni ponavljajući sljedovi.

Poznavanje i proučavanje genoma C. Elegansa je od velikog značaja, pa se njime osim biologa bave i bioinformatičari. Otkriven je cijeli genom C. Elegansa. Neki se geni još provjeravaju i onda ispravaljaju, ali uglavnom možemo govoriti o znanstevenicima potpuno poznatom genomu. Biolozi su svoja istraživanja prilagodili informatičarima, tako da su genom zapisali u obliku vektora slova. Vektore čine samo četiri slova koja predstavljaju četiri nukleotida koji izgrađuju genom crva. Ta slova su: A, T, G i T. Informacije u tom obliku pogodne su za intormatička istraživanja.

Kako u procesu prekrajanja mRNA točno prepoznati mjesta gdje se vrši izrezivanje? To je jedno od najvažnijih pitanja kojima se bavimo u ovome radu. Poznato je da svaki intron kod crva počinje s slovima GT, koja prestavljaju 5' mjesto izrezivanja, odnosto donorsko mjesto, a završava sa sekvencom AG, koja je 3' mjesto izrezivanja, odnostno akceptorsko mjesto. Mnogi eukariotski organizmi posjeduju još jedno mjesto koje pomaže pri pronalasku pravog mjesta izrezivanja, 2', odnosno točku grananja. Istraživanja su pokazala da crv ne posjeduje mjesto grananja, već intronu u izrezivanju pomažu samo dva mjesta: 5' i 3'.

Slika 12. Donorsko i akceptorsko mjesto u intronu

Poznato je da svaki intron posjeduje nekoliko mjesta na kojima se može izrezati, to ovisi o tome koji je protein stanici u tom trenutku potreban i koje proteine ona tada proizvodi da bi katalizirala nastanak točno određenog proteina. Ta činjenica nam otežava istraživanje, jer govori da se svaki intron i ekson mogu izrezati na više načina i tvoriti različite proteine.

Kako prepoznati pravo mjesto izrezivanja? Ulazni podaci koje smo korsitili su genom crva koji je potpuno određen, dakle dobili smo informacije točnih mjesta izrezivanja, na temelju kojih smo trenirali Random Forest.

7. Važnost pojedinih nukleotida u izrezivanju introna C. Elegansa

7.1. Donorsko mjesto

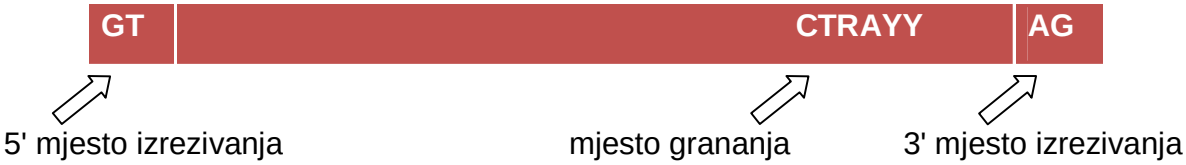
Svaki gen crva sastoji se od egzona i introna, pri čemu su introni nekodirajuće regije, koje se izrezuju iz pre-mRNA, što dovodi do nastanka zrele mRNA.

Poznata je činjenica da svaki intron započinje s dva nukleotida: GT. Ta činjenica nam omogućava da bioinformatičkom obradom podataka, koji predstavljaju cijeli genom crva, a zapisani su u obliku nukleotida, pronađemo početak introna. Početak introna se još zove i donorsko mjesto. Podatci koje smo koristili su nizovi nukleotida. Svaki niz je duljine 398 nukleotida, a na kraju niza još je dodana informacija radi li se stvarno o mjestu izrezivanja ili ne. Ta činjenica je poznata iz bioloških istraživanja. Poznavajući tu činjenicu možemo vršiti istraživanja, odnosno obrađivati podatke. Ulazni podatci su posloženi tako da se kombinacija GT nalazi uvijek na istom mjestu u nizu, dakle oni su poravnani. Donorsko mjesto u nizovima uvijek je na poziciji 200 guanin, a timin je uvijek 201. nukleotid u nizu.

Poznavajući te činjenici može se započeti obrada podataka. Prvo se podatci pripremaju u oblik potreban statističkim metodama, nakon toga se dobiveni podatci pripremaju za crtanje i na kraju se crtaju krivulje, koje pokazuju uspješnosti naših detektora.

7.2. Mjesto grananja

Drugo važno mjesto za izrezivanje introna iz pre-mRNA je mjesto grananja. To mjesto se može pronaći kod većine eukariotskih organizama. Kod većine eukariotskih organizama mjesto grananja se nalazi u intronu 10 do 30 nukleotida od akceptorskog mjesta, dakle ono je bliže 3' mjestu izrezivanja nego što je 5' mjestu izrezivanja.



Slika 13. Položaji mjesta izrezivanja kod većine sisavaca

Crv se razlikuje od većine životinjskih organizama, jer za razliku od njih, introni crva nemaju mjesto grananja, iako imaju proteine koji se inaće pojavljuju oko mjesta grananja.

Ta činjenica nam otežava bioinformatička istraživanja i predviđanja mjesta izrezivanja introna, jer nam smanjuje broj podataka s kojima možemo vršiti statističke metode. Statističke metode daju točnije podatke ako im dajemo veći broj poznatih podataka.

Slika 14. Izgled mjesta bitnih za izrezivanje intron kod crva

7.3. Pronalazak akceptorskog mjesta I

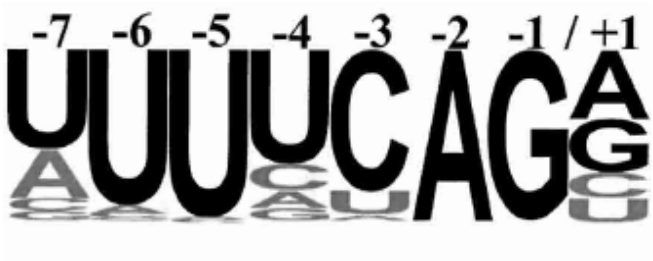
Crv je ipak zanimljiv organizam koji se razlikuje od većine drugih animalnih organizama. Naime on ima odeđenu jedinstvenost koja nam omogućava da pronađemo točno mjesto izrezivanja akceptora. Istraživanja su pokazala da većina introna kod crva završava s grupom nukleotida: UUUUCAG. Poznavanje te činjenice znatno nam je pomoglo u istraživanju, jer ta činjenica povećava broj poznatih sekvenci koje su bitne pri izrezivanju.

Slika 15. Izgled 5' i 3' mjesta izrezivanja kod crva

Naravno ne završva svaki intron kod crva točno s UUUUCAG, već postoje i razlike, ali ipak u većini slučajeva možemo tvrditi da intron završava baš s tom sekvencom.

Prolaskom kroz sve gene crva i brojanjem slova koja se nalaze u okolini AG može se provjeriti ta pretpostavka. Svaki ulazni vektor na kojem smo vršili istraživanje ima ukupono 398 nukleotida. Posebno smo vršili istraživanja za akceptore i donore. Kod akceptora 3' mjesto izrezivanja nalazili se na pozicijama 198 i 199. Poznavanjem te činjenice i provjerom iste, mogli smo krenuti u istraživanje.

Provjeravajući tvrdnju da svaki intron kod crva završava s UUUUCAG stvorili smo pet različitih vekora. Svaki vekor je imao jedan redak i četiri stupca. Svaki stupac predstavljao je jedan nukleotid i to redom: A, T, C, G. Prolaskom kroz cijeli genom brojali smo koliko se puta na određenoj poziciji pojavilo određeno slovo i dobivene vrijednosti spremali u vektore. Pozicije u početnom vektoru koje smo istraživali su 193, 194, 195, 196 i 197. U isto vrijme smo izbrojali i koliko se puta na poziciji 198 i 199 pojavljuje AG.



Slika 13. Sedam zadnjih nukleotida u intronu crva

Dokazali smo da svi vektori za koje se tvrdilo da na 198. i 199. poziciji imaju AG, zaista i imaju AG na toj poziciji.

Provjeravali smo tvrdnje da se na poziciji:

- 193 (odnosno -7) u najvećoj količini pojavljuje timin, a zatim slijede: adenin, citozin i guanin

- 194 (odnosno -6) u najvećoj količini pojavljuje timin, a zatim adenin, koji se pojavljuje u puno više slučajeva od citozina i guanina, koji loše utječu na izrezivanje
- 195 (odnosno -5) je pozicija koja najviše, izuzev naravno AG, utječe na izrezivanje. Na toj poziciji T se pojavljuje u 97% slučajeva, u malim količinama možemo pronaći i adenin i citozin, dok guanin ima jako malu frekvenciju pojavljivanja i jako loše utječe na izrezivanje, odnosno vezanje snRNA i proteina na to mjesto izrezivanja
- 196 (odnosno -4) mnogo je tolerantnija na pojavljianje drugih nukleotida, u odnosu na -5 poziciju. Ipak u najvećoj količini i kod nje možemo pronaći timin, pa zatim slijede adenin, citozin i guanin, koji se pojavljuje čak u 8% slučajeva
- 197 (odnosno -3) u najviše slučajeva može pronaći citozin, timin se nalazi samo u 15% slučajeva, adenin u 2%, a guanin se nikada ne pojavljuje na toj poziciji
- 198 (odnosno -2) pojavljuje uvijek A
- 199 (odnosno -1) pojavljuje uvijek G

pozicija	p(A) [%]	p(T) [%]	p(c) [%]	p(g) [%]
-7	28.57	57	8.01	6.42
-6	5.75	88.88	3.26	2.11
-5	0.83	97.43	1.41	0.33
-4	8.8	67.35	16.14	7.71
-3	3.03	13.33	83.43	0.21

Tablica 1. Prikaz udjela pojedinog nukleotida na kraju introna crva na pozicijama -7, -6, -5,-4, -3

Provedena istraživanja su u velikoj mjeri potvrdila početne predostavke, naravno uz male razlike. Rezultati isrtaživanja su:

- 193 (odnosno -7) u najvećoj količini pojavljuje timin, 57%, a zatim slijede: adenin u 28.57%, citozin u 8.01% i guanin u 6.42% slučajeva
- 194 (odnosno -6) u najvećoj količini pojavljuje timin, 88,88%, a zatim adenin u 5.75%, citozin u 3.26% i guanin u 2.11% slučajeva
- 195 (odnosno -5) je pozicija koja najviše, izuzev naravno AG, utječe na izrezivanje. Na toj poziciji T se pojavljuje u 97, 43% slučajeva, u malim količinama pojavljuje se: adenin u 0.83%, citozin u 1.41% a guanin se isto ipak pojavljuje u 0.33% slučajeva
- 196 (odnosno -4) u najviše slučajeva očekivano je pronađen timin, točnije u 67,35% slučajeva, zatim slijedi citozin s 16,14%, pa adenin s 8.8% i na kraju guanin s 7.71%
- 197 (odnosno -3) u najviše slučajeva pronađen je citozin i to u 83,4%, timin se nalazi samo u 13.33% slučajeva, adenin u 3.03%, a guanin se pojavljuje na toj poziciji u 0.21% slučajeva
- 198 (odnosno -2) pojavljuje uvijek A
- 199 (odnosno -1) pojavljuje uvijek G

Iz dobivenih podataka lako je zaključiti da su polazne pretpostavke bile približno jednake konačnim rezultatima. Naravno postoji par iznimki. Na poziciji -7 svi se nukleotidi pojavljuju u količinama koje odgovaraju uvjetu postavljenom u početnoj tvrdnji koji glasi: timin>adenin>citozin>guanin.

Na poziciji -6 kao što je i predpostavljeno najviše se pojavljuje timin, pa adenin i u najmanim količinama citozin i guanin.

Na poziciji -5 timin se pojavljuje u 97.43%, a početna pretpostavka je bila 97%. Citozin i adenin mogu se naći u mnogo manjim količinama od timina, a guanin u neznatno malim u odnosu na timin, a to odgovara početnim predpostavkama.

Na poziciji -4 očekivano se najviše pojavljuje timin, pa zatim adenin, citozin i guanin. Pretpostavljeno je pojavljivanje guanina u 8% slučajeva, dok se on zaista na toj poziciji nalazi u 7.71% slučajeva izrezivanja.

Konačno na poziciji -3 citozin se stvarno nalazi u najvećoj količini, zatim slijedi timin s postotkom 13.33, a ne očekivanih 15%, adenin smo pronašli na toj poziciji u 3.03% slučajeva, a ne očekivanih 2%, dok tvrdnja da se guanin nikada ne nalazi na -3 poziciji nije zadovoljena, jer se on ipak pojavljuje u 0.21% slučajeva.

Provjerivši početne pretpostavke o tome da se na kraju introna uglavnom nalazi niz nukleotida UUUUCAG može se početi stvaranje algoritama za učinkovitije određivanje mjesta prekrajanja, jer sada imamo više podataka koji nam to omogućuju. Dakle, sada više ne uzimamo samo AG kao važnu značajku, već i niz od 5 nukleotida UUUUC, što tako čini ukupono sedam značajnih nukleotida za određivanje akceptorskog mjesta.

Ulazni podatci su zapisani na način koji nije pogodan statstičkoj metodi Random Forest za obradu, tako da ulazne podatke prvo moramo prilagoditi. Podaci se prilagođavaju tako da se na početku datoteke koju stvaramo nalaze atributi koji predstavljaju nukleotide koji se mogu naći na toj poziciji vektora. Atributi su: A,T,G i C. Nakon atributa dodaje se i jedna klasa, koja može poprimiti vrijednost F ili T, ovisno o tome radi li se stvarno o akceptorskom mjestu ili ne. Ako je biološki potvrđeno da je taj vektor stvarno intron, tada se na kraj ulaznog vektora dodaje T, a ako nije intron, dodaje se F. Broj atributa jednak je broju nukleotida koji se nalaze u svakom vektoru. U načem slučaju postoji 140 atributa i jedana klasa.

Učitava se redak po redak iz ulazne datoteka koja predstavlja cijeli genom. Provjerava se labela u ulaznim podatcima. Ako je labela 1 tada se radi o stvarnim akceptorskim i donorskim mjestima. Nakon toga smanjujemo ulazne podatke tako da ulazni podatci moraju zadovoljiti još jedan uvjet.

Dokazali smo da se u većini slučajeva na poziciji 195 nalazi timin (odnosno uracil, samo što mi radimo s podatcima na kojima nije izvršena transkripcija, pa na mjestima gdje bi trebao biti uracil se i dalje nalazi timin), u 97,43%, a guanin samo u 0.33% slučajeva. Guanin se nalazi u jako malom postotku, tako da se on jako rijetko nalati na toj poziciji u intronu, pa smo ga zanemarili. Tako da je prvi uvjet bio da se na poziciji -5 ne može naći guanin. Tako samo smanjili skup podataka na kojima smo trenirali Random Forest.

Drugi uvijet je sličan, ali na drugoj poziciji. Promatrajući poziciju -3 u intronu zaključili smo da se u većini slučajeva na njoj nalazi citozin, ali u velikim količinama nalaze se i timin i adenin. Guanin se na toj poziciji jako rijetko pojavljuje. Pojavljuje se samo u 0.21% slučajeva. Tako da smo zanemarili i introne u kojima se na poziciji -3 nalazi guanin. Pri tome smo smanjili broj ulaznih podataka, što nam znatno ubrzava izvođenje algoritama, koji traju jako dugo zbog količine podataka koja se obrađuje.

Ako je labela u ulaznom vekoru podatak imala vrijednost 0, taj vektor ne predstavlja vekor koji sadrži akceptorsko mjesto izrezivanja. Na njemu smo isto vršili obradu kojom smo smanjili broj podataka koje štaljemo statističkom paketu Random Forest.

Smanjivanje podataka kod podataka koji siguruno ne predstavljaju mjesto izrezivanja obavljeno je tako da smo izbacili podatke koji sigurno ne mogu predstavljati mjesto izrezivanja. U novu datoteku prepisuju se samo podaci koji na poziciji -3 i -5 nemaju guanin.

Svi podaci zapisani u novoj datoteci zapisani u obliku vektora u kojima su nukleotidi odvojeni zarezima, a na kraju svakog vektora dodaje se još oznaka F ili T, ovisno o tome predstavlja li taj vektor akceptorsko mjesto ili ne. Sada su podaci spremni za obradu statističkom metodom Random Forest.

7.3.1. Statistička metoda 'Random Forest'

'RandomForest' je statistička metoda uzorkovanja s ponavljanjem. Iz ulaznog skupa se nasumično bira određeni broj podataka kreirajući tako jedno stablo. Istim principom dodjeljuju se drugim stablima njihovi podaci. Korisnik određuje broj stabala (porastom broja stabala smanjuje se vjerojatnost pogreške). RF radi na dva načina. U prvom koristi isti skup podataka i za treniranje i za validaciju, a u drugom korisnik može sam odrediti koji skup želi koristiti za jedno, a koji za drugo. U fazi treniranja stabla se uče na poznatim podacima, dok u fazi validacije (out-of-bag podaci - koriste se za procjenu pogreške stabala) moraju donijeti ispravnu odluku na temelju nepoznatih podataka. Svako stablo glasa za ili protiv neke ulazne pretpostavke te se na temelju većinskih glasova (broj glasova je veći od zadanog praga) donosi odluka.

Statistički paket 'Random Forest' kao ulaznu datoteku koristi 'C_elegans_akceptor.txt', na temelju kojih se stvaraju stabla odluke. Formiraju se tri izlazne datoteke:

- glasovi.txt
- izlmatrica.txt
- vaznost.txt

7.3.2. Crtanje krivulja

Nakon obrade podataka statističkom medom potrebno je te podatke preoblikovati u neki oblik iz kojega ćemo moći vidjeti preciznost našeg detektora. Oblik pogodan za promatranje rezultata su Precision-Recall krivulje. Za crtanje tih krivulja koristi se datoteka 'glasovi.txt', koja predstavlja jedan od izlaznih podataka statističke obrade, ali isto tako koristi se i datoteka koju smo kreirali prije statističke obrade, u kojoj se na kraju svakog retka nalaze F ili T, koji označavaju radi li se o akceptorima ili ne.

Obrada novih ulaznih podataka se vrši tako da se podaci smještaju u matrice s i redaka (točan broj redaka je broj redaka koji se nalazi u datoteci koja je dobivena obradom potrebnom za Random Forest) i 5 stupaca. U prva dva stupca upisuju se podaci zapisani u izlaznu datoteku Random Foresta, odnosno podaci iz 'glasovi.txt'. U

prvom stucu zapisani su borojevi koji predstavljaju udio stabala koja su glasovala protiv, a u drugom udio stabala koja su glasovala za. U zadnji stupac matrice upisuje se točnost, koja nam je bila poznata već u početnim podatcima istraživanja, odnosno točnost dobivena biološkim i kemijskim metodama. Znak točnosti je bio zapisan u obliku T ili F. Prije upisivanja u matricu ispituje se taj znak, ako se pojavljuje T u matricu se zapisuje 1, a ako je F u matricu se upisuje 0. Treći stupac matrice predstavlja postotak koliko je stabala glasalo za u odnosu na ukupan broj stabala. Treći stupac se dobiva formulom:

$$stupac\_3 = \frac{stupac\_2}{stupac\_1 + stupac\_2} \quad (1)$$

Prag se linerano povećava od 0 do 1. Četvrti stupac nastaje uspoređivanjem trećeg stupca i praga. Ukoliko je treći stupac veći od praga upisuje se jedinica, a u protivnom nula.

GLASOVI PROTIV	GLASOVI ZA	X=ZA/(PROTIV+ZA)	X >PRAG => 1 X<PRAG => 0	TOČNOST IZ UL. MAT
0.3621	0.6379	0.3621	0	0
0.9737	0.0263	0.9737	1	1
0.8122	0.1878	0.8122	1	0
0.8984	0.1016	0.8984	1	1

Tablica 2. Primjer dijela tablice koja se koristi za crtanje krivulja

Precision-Recall krivulja crta se na temelju četiri parametra:.

- TP – True positives → detektor (odluka RF-a) je rekao da je točno i bilo je točno (odluka donesena na temelju ulaznih podataka)
- TN – True negatives → detektor kaže da je netočno i njegova odluka je prema ulaznim podacima točna
- FP – False positives → detektor kaže da je točno i njegova odluka je prema ulaznim podacima netočna (bilo je netočno)
- FN – False negatives → detektor kaže da je netočno i njegova odluka je prema ulaznim podacima netočna (bilo je točno)

Formule za izračunavanje varijabli Precision i Recall su:

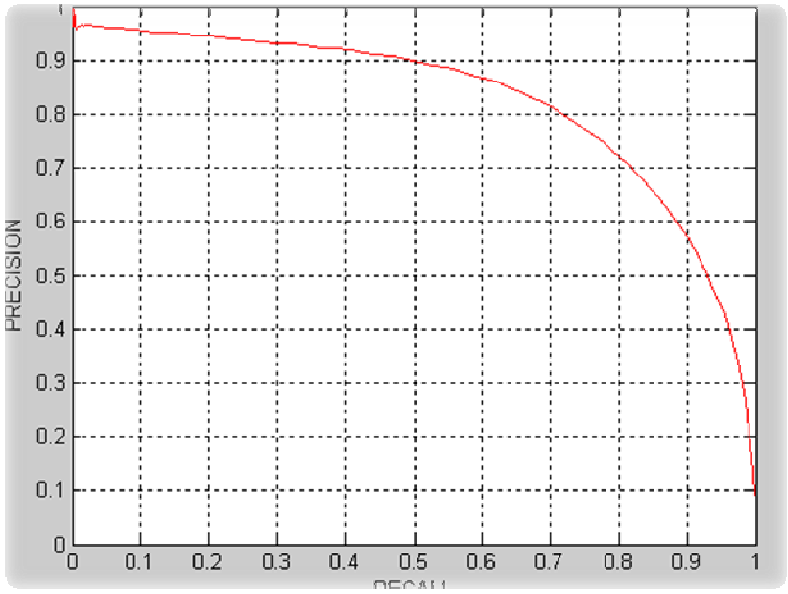
$$\Pr ecision = \frac{TP}{TP + FP} \tag{2}$$

$$\mathrm{Re}\, call = \frac{TP}{TP + FN} \tag{3}$$

Svakom parametru dodijeljene su kombinacije binarnih znakova dobivene usporedbom podataka dobivenih detektorom i točnih ulaznih podataka.

Navedenom obradom podataka, koje smo dobili kao izlazne podatke obrađene statističkom metodom Random Forest, dobivamo krivulje iz kojih možemo vidjeti točnost našeg detektora. Idealan detektor bi imao izgled pravca koji poprima vrijednosti na intervalu od [0,1]. Detektor ima vriejdnost y=1 na intervalu [0,1), a u točki x=1 y je element od [0,1].

Naš detektor nije idealan detektor, tako da on neće uvijek točno odrediti akceptorsko mjeto, ali ipak ima u velikom broju slučajeva on će ipak biti uspješan.



Slika14. Detektor s vektorima duljine 140 nukleotida i s predpostavkom da se na pozicijama -3 i -5 ne pojavljuje guanin

7.4. Pronalazak akceptorskog mjesta II

Za provjeravanje dobivenog detektora izvršena je još jedna analiza. Početni podatci su opet isti podaci, odnosno ista baza genoma crva, točnije njegovih akceptora. Ulazni podaci opet nisu bili u pogodnom obliku za statističku analizu. Otvorena je nova datoteka u koju se upisju obrađeni podaci. Na početku datoteke opet se navode atributi i klasa. Atributa ima onoliko koliko i slova u pojedinom retku, koja smo kopirali iz početnih podataka i na kojima je vršena daljnja analiza.

Ulazni vektori su duljine 398 nukleotida. Akceptorsko mjesto se nalazi na 198. i 199. nukleotidu. U novu datoteku se ne prepisuju svi nukleotidi, već samo 80 nukleotida od akceptorskog mjesta u lijevu stranu, odnosno 80 nukleotida koji su dio introna i 60 nukleotida u desnu stranu, koji su dio egzona. Ne prepisuju se nukleotidi na poziciji 198 i 199, jer se na toj poziciji u svim vektorima nalazi AG, pa nam ta pozicija ne predstavlja novu informaciju.

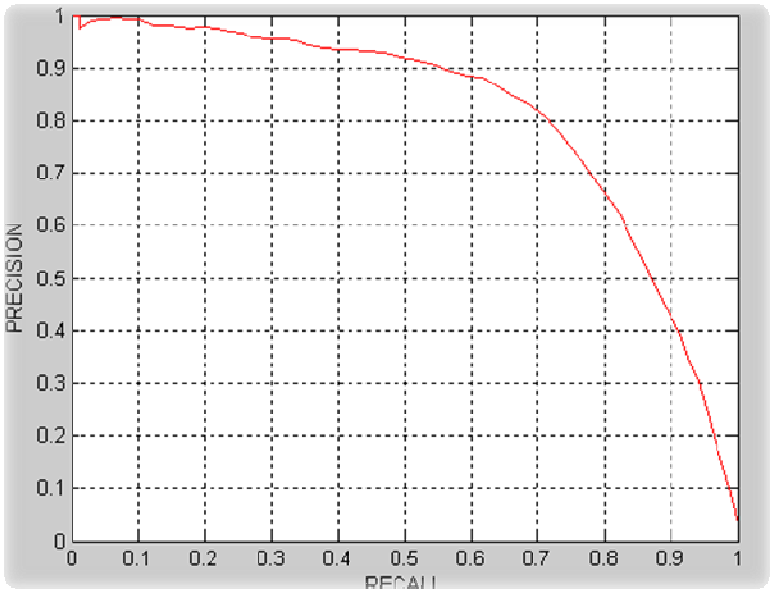
Provjerava se točnost uz pomoć ulazne značajke zapisane u varijabli labela. Ako je labela jednaka 1, tada taj niz predstavlja akceptorsko mjesto, a ako je labela jednaka 0, tada taj niz ne prestavlja akceptorsko mjesto.

Na kraju svakog obrađenog i skraćenog niza dodajemo klasu, odnosno točnost, koju čitamo iz labele, ako je labela 1 na kraj niza upisujemo T (true), a ako je 0 na kraj niza upisujemo F (folse).

Podatci su sad pretvoreni u oblik pogodan za Random Forest. Random Forest je statistički paket koji na njima vrši potrebnu analizu i kao rezultat daje tri datoteke. Datoteke su: glasovi_AG.txt, vaznost_AG.txt i izlmatrica_AG.txt.

Datoteka glasovi_AG služi za crtanje krivulje koja nam pokazje kvalitetu našeg detektora. Glasovi dobiveni statističkom obradom se obrađuju tako da se vidi koliki je postotak stabala glasao za u odnosu na ukupan broj stabala, dobivena vrijedost se uspoređuje s odabranim pragom i ako je veća od odabranog praga smatra se da je to

točno mjesto izrezivanja, a ako je manja od zadanog praga, tada to mjesto nije akceptorsko mjesto. Na kraju dobivamo krivulju koja pokazuje kvalitetu detektora.



Slika 15. Detektor s vektorima duljine 140 nukleotida

Važna činjenica je to da dobiveni detektor ima manju točnost od onoga u kojem smo uključili još neke značajke, odnosno u kojem smo izbacili mogućnost postojanja guanina na poziciji -3 i -5 od početka akceptorskog mjesta. To dokazuje važnost poznavanja što više značajki pri obradi podataka statističkim metodama, jer što više značajki znamo statističke metode imaju veću točnost.

8. Zaključak

DNA nosi genetičke informacije koje se procesom transkripcije prenose na novonastalu RNA. Novonastala RNA ne može izravno služiti kako kalup za sintezu proteina, već prvo mora proći kroz niz različitih promjena poput dodavanja repa, izrezivanja introna, međusobnog povezivanja egzona i na kraju nastanka proteina.

Zbog svega toga biologija je jako složena i biološki porces otkrivanja genoma i njegovih značajki je jako složen, dugotrajan i skup. Biolozi su otkrili genome nekih organizama, ali ipak genomi većine organizama su i dalje nepoznanica. Promatranje genoma kao niza od četiri nukleotida, čije pozicije imaju određene značajke, omogućava uvođenje novih pogleda na biološke procese.

Bioinformatičkim metodama obrađuju se podaci na jedan drugi način. Način koji ne zahtjeva puno novca, a omogućava predviđanja mnogo značajki i učenje detektora da prepoznaju mnoge značajeke u genomu. Učenja detektoda se vrše uz pomoć već biološki otkrivenih značajki. Statističe metode se treniraju na tim podacima i onda nakon toga nastaju detektori koji mogu sami, na nekom novom ulaznom skupu, određivati pojedine značajke genoma.

Statističke metode daju bolje rezultate ako su provedene na većem skupu podataka, pogotovo ako se treniraju sa što više poznatih činjenica.

Genom C. Elegana ima samo dva mjesta bitna za prekrajanje. Crv je ipak zanimljiv jer u 3' mjesto izrezivanja ima i niz od 5 nukleotida koji se u velikim postotcima nalaze da tom mjestu. Detektor treniran na tim podatcima, odnosno s više značajki daje bolje rezultate, u više posto slučajeva će točno prepoznati akceptorsko mjesto u odnosu na detektor koji koristi samo informaciju o tome nalaze li se na određenom mjestu nukleotidi AG i jesu li biolozi potvrdili za njih da su oni akceptorsko mjesto.

Provedena istraživanja omogućavaju ubrzanje razvoja bioinformatike, ali i same biologije, jer olakšvaju posao biolozima pri proučavanju nepoznaith genoma. Na taj način biolozi mogu znati što otprtilike mogu očekivati u svojim istraživanjima, pozivajući se na podatke dobivene statističim metodama. Time se smanjuje količina vremena i novaca utoršena u istraživanja, a povećava broj poznatih genoma i raznih drugih genetičkih činjenica.

9. Literatura

1. Alberts B., Johanson A., Lewis J., Raff M., Roberts K., Walter P., *The Cell*, New York, Garland Science, 2002.
2. Copper G. M., *Stanica*, Zagreb, Medicinska naklada, 2004.
3. Parađina N., Analiza DNK metodama obrade signala u vremenskoj i frekvencijskoj domeni, Diplomski rad br. 1143, FER, 2008.
4. Penović M., diplomski rad,FER, 2008.
5. Hollins C, Zorio D. A. R., Macmorris M., Blumenthal T., 20.01.2005., *U2AF binding selects for the high conservation of the* C. elegans *3′ splice site* , 15.5.2009.
6. Sören Sonnenburg, Gabriele Schweikert, Petra Philips, Jonas Behr, Gunnar Rätsch, *Accurate splice site prediction using support vector machines,* 8.12.2006., http://www.biomedcentral.com/content/pdf/1471-2105-8-S10-S7.pdf, 4.11.2008.
7. http://en.wikipedia.org/wiki/DNA, 8.3.2009.
8. http://en.wikipedia.org/wiki/Precision_and_recall, 9.3.2009.
9. http://en.wikipedia.org/wiki/RNA , 9.3.2009.
10. http://en.wikipedia.org/wiki/transkripcija , 12.3.2009.

10. Sažetak

Analiza kodirajućih regija u genomu

DNA molekula sastoji se od dva međusobno povezana spiralna lanca građena od četiri nukleotida: adenina (A), timina (T), citozina (C) i guvanina (G). Molekula je građena tako da se adenin i timin, citozin i guanin međusobno vežu. Nukleotidi čine baze lanaca koji su međusobno povezani kovalentnim vezama.

DNA molekula ne upravlja sama sintezom proteina. Sintezom proteina upravlja RNA koja je intermedijalna molekula. Kada je stanici potreban određeni protein, nukleotidna sekvenca, određenog dijela jako duge DAN, smještene u kromosomu, kopira se u RNA. Te kopije DNA služe kao kalupi za sintezu proteina.

Proces sinteze nije jednostavan već se sastoji od niza koraka koji počinju s obradom DNA, odnosno rastavljanjem dva lanca DNA, prepisivanjem jednog lanca u RNA, zatim obradom RNA, koja uključuje i izrezivanje nekodirajućih dijelova i na kraju sintezu aminokiselina u protein.

Genom C. Elegansa je prvi sekvencirani genom nekog višestaničnog organizma. Sadrži oko 19000 slijedova koji kodiraju proteine i zauzimaju oko 25% genoma. Svaki gen C. Elegansa sastoji se od kodirajućeg (egzona) i nekodirajućeg (introna) dijela, pri čemu svaki gen sadrži 5 introna.

Izrezivanje introna vrši se uz pomoć proteina i malih RNA molekula. Kod crva postoje samo dva važna mjesta za izrezivanje: 3' i 5', dok kod nekih organizama postoji i treće: mjesto grananja u intronu.

Koristeći se poznavanjem biologije i informatike, odnosno bioinformatike, uz pomoć programiranja i statističkih metoda moguće je provesti analize genoma i na temelju njih odrediti mjesta izrezivanja. Ulazne podatke pri analizi čini cijeli genom zapisan u obliku slova A, T, C i G. Ta slova predstavljaju signal koji se uređuje i

statistički obrađuje. Pri analizi i obradi podataka bitno je poznavanje činjenice da donorsko mjesto, odnosno 5', uvijek počinje s GT, a akceptorsko mjesto 3' s AG.

10.1. Ključne riječi

DNA, RNA, nukleotid, genom, gen, transkripcija, translacija, izrezivanje, mRNA, intron, egzon, mjesto grananja, akceptor, donor, programiranje, bioinformatika

10.2. Summary

Genome coding regions analysis

DNA molecule has two helical strands made from repeating units called nucleotides. These nucleotides are adenine (A), guanine (G), cytosine (C) and thymine (T). Adenine and thymine, and guanine and cytosine bind together.

The DNA in genomes does not direct protein synthesis itself, but instead uses RNA as an intermediary molecule. When the cell needs a particular protein, the nucleotide sequence of the appropriate protein of the immensely long DNA molecule in a chromosome is first copied in to RNA. It is these RNA copies of segments of the DNA that are used directly as templates to direct the synthesis of the protein.

Process of synthesis is not easy. Transcripts in eukaryotic cells are subject to a series of processing steps in nucleus, including opening and unwinding of a small portion of the DNA double helix to expose the bases of the each strand, transcription one strand in to RNA, RNA analysis, which includes RNA splicing and capping, and in the end protein synthesis.

C. elegans was the first eukaryotic organism to have its genome completely sequenced. It has about 19 000 protein coding sequences, which makes about 25% of hole genom. Each gen of C. Elegans has coding (exson) and noncoding part (intron). Each gen has 5 introns.

RNA splicing is preformed by proteins and small RNA molekules. Worm has only two inportant splicing sites: 3' and 5', while the most of the other eucaryotic organisam have third splice site too, branch point.

Using biology and informatics knowledge, that is bioinformatics, programming and statistic we can do genome analysis which can help us to find splice sites. Our data base is a whole genome written with four letters: A, T, C and G. These letters are signal which need to be statistically processed. Upon that, it is important to know that donors site, or we can say 5', begins with GT, and acceptors site with AG.

10.3. Key Worords

DNA, RNA, nucleotide, genome, gen, transcription, translation, splicing, mRNA, intron, exson, branch point, acceptor, donor, programming, bioinformatics

11. Privitak

Program 'sume_izrezivanje.m' služi za određivanje koliko se puta javlja pojedini nukleotid na pet bitnih pozicija u akceptorima. Program služi i za pripremu podataka u oblik potreban Random forestu. Otvara se nova datoteku na čiji se početak upisuju atributi (A, T, C i G), nakon kojih dolazi klasa (T ili F). Nizovi se obrađuju tako da se provjerava nalazi li se A na 198. mjestu niza i G na 199. mjestu. Zatim se ispituje labela kojom se određuje radili se zaista o akceptorskom mjestu ili ne. Ako je labela jedanka 1, ispitivani niz je stvarno akceptorsko mjesto. Nakon toga provjerava se nalazi li se na poziciji 195 i 197 G, ako se ne nalazi taj niz prestavlja stvarno mjesto akceptora i zapisuje se u novu datoteku, ali se prije toga nizovi smanjuju na duljinu od 140 nukleotida, tako da se uzeme 80 nukleotida lijevo od akceptorskog mjesta, a 60 njih desno od akceptorskog mjesta. Nukleotidi se međusobno odvajaju zarezima i na kraju se dodaje T ili F. Nakon toga se na isti način obrađuju podatci s labelom 0, u novu datoteku se zapisuju oni koji na pozicji 195 i 197 nemaju guanin. U treću datoteku upisuju se vrijednosti suma koje predstavljaju koliko se puta pojavljuje pojedini nukleotid na ispitivanoj poziciji. Na kraju toga podatci su spremni za obradu.

'sume_izrezivanje.m':

```
% Rezervacija prostora za:

    % podatke
    %podaci = char(zeros(1000, 140));
    %podaci(900000,:) = char(zeros(1, 140));

    % zastavice tocno / netocno
    tocnost = char(zeros(1000, 1));
    tocnost(900000,1) = char(0);

    % pozicije u genomu
    pozicija = zeros(900000,1);

%postavljanje suma na 0
sumaAGvani=0;
sumaAGunutra=0;
sumatri=zeros(1,4);
sumacetri=zeros(1,4);
sumapet=zeros(1,4);
sumasest=zeros(1,4);
```

```
sumasedam=zeros(1,4);

% Ucitavanje datoteka i pohrana podataka

% Open file
fid = fopen('C_elegans_acc_all_examples.fasta', 'r');
fid2= fopen('sume_izrezivanje.txt','w');
fid3= fopen('sume.txt','w');

% header
fprintf(fid2, '@relation C_elegans_branchpoint.txt\n');
for n=1:140
    fprintf(fid2, '@attribute NA%d {A,T,G,C}\n', n);
end
fprintf(fid2, '@attribute class {F,T}\n');
fprintf(fid2, '\n');
fprintf(fid2, '@data\n');


% Read input data from the FASTA file
l=1;
for k=1:Inf
     if (mod(k,10000)==0)
        a={'gegam se'}
        k/10000
     end
    % Read header
    head = fgets(fid);
    if isnumeric(head), break, end % end of the file

     % Chromosome number
    ind = regexp(head, 'chr_num\s*=\s*[+-\d]', 'end');
    len = regexp(head(ind:end), '[+-]?\s*\d*', 'once', 'end');
    chr_num = str2double(head(ind:ind+len-1));

     % Strand
    ind = regexp(head, 'strand\s*=\s*[+-]', 'end');
    strand = head(ind);

    % Position
    ind = regexp(head, 'position\s*=\s*[+-\d]', 'end');
    len = regexp(head(ind:end), '[+-]?\s*\d*', 'once', 'end');
    position = str2double(head(ind:ind+len-1));

    % Label
    ind = regexp(head, 'label\s*=\s*[+-\d]', 'end');
    len = regexp(head(ind:end), '[+-]?\s*\d*', 'once', 'end');
    label = str2double(head(ind:ind+len-1));

    % Read data
    data = fgets(fid);
    if isnumeric(head), break, end % end of the file
    sumaAGvani=sumaAGvani+1;

 % Do something
```

```
if (data(198) == 'A' && data(199) == 'G')
sumaAGunutra=sumaAGunutra+1;
  if (label==1)

     %sume  za pozicije -3
      if data(197)=='A'
          sumatri(1)=sumatri(1)+1;
      end
      if data(197)=='T'
          sumatri(2)=sumatri(2)+1;
      end
      if data(197)=='C'
          sumatri(3)=sumatri(3)+1;
      end
      if data(197)=='G'
          sumatri(4)=sumatri(4)+1;
      end

      %sume za pozicije-4
      if data(196)=='A'
          sumacetri(1)=sumacetri(1)+1;
      end
      if data(196)=='T'
          sumacetri(2)=sumacetri(2)+1;
      end
      if data(196)=='C'
          sumacetri(3)=sumacetri(3)+1;
      end
      if data(196)=='G'
          sumacetri(4)=sumacetri(4)+1;
      end

      %sume za pozicije-5
      if data(195)=='A'
          sumapet(1)=sumapet(1)+1;
      end
      if data(195)=='T'
          sumapet(2)=sumapet(2)+1;
      end
      if data(195)=='C'
          sumapet(3)=sumapet(3)+1;
      end
      if data(195)=='G'
          sumapet(4)=sumapet(4)+1;
      end

      %sume za pozicije-6
      if data(194)=='A'
          sumasest(1)=sumasest(1)+1;
      end
      if data(194)=='T'
          sumasest(2)=sumasest(2)+1;
      end
      if data(194)=='C'
          sumasest(3)=sumasest(3)+1;
      end
```

```
            if data(194)=='G'
                sumasest(4)=sumasest(4)+1;
            end

            %sume za pozicije-7
            if data(193)=='A'
                sumasedam(1)=sumasedam(1)+1;
            end
            if data(193)=='T'
                sumasedam(2)=sumasedam(2)+1;
            end
            if data(193)=='C'
                sumasedam(3)=sumasedam(3)+1;
            end
            if data(193)=='G'
                sumasedam(4)=sumasedam(4)+1;
            end



            if (data(195)~='G' && data(197) ~= 'G')
                tocnost(l) = 'T';
                fprintf(fid2,
'%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,', data([118:197,200:259]));
                fprintf(fid2, '%c\n', tocnost(l));
                pozicija(l) = position;
                l = l+1;
            end
        else
            if (data(195)~='G' && data(197) ~= 'G')
              tocnost(l) = 'F';
              fprintf(fid2,
'%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,', data([118:197,200:259]));
              fprintf(fid2, '%c\n', tocnost(l));
              pozicija(l) = position;
              l = l+1;
           end
        end
      end
end
%ispis suma
fprintf(fid3, '%d\n', sumaAGvani);
fprintf(fid3, '%d\n', sumaAGunutra);
fprintf(fid3, '%d %d %d %d\n', sumatri);
fprintf(fid3, '%d %d %d %d\n', sumacetri);
fprintf(fid3, '%d %d %d %d\n', sumapet);
fprintf(fid3, '%d %d %d %d\n', sumasest);
```

```
fprintf(fid3, '%d %d %d %d', sumasedam);


% Close file
fclose(fid);
fclose(fid2);
fclose(fid3);

% Uklanja visak prostora
tocnost (l:end,:)=[];
%podaci  (l:end,:)=[];
pozicija(l:end,:)=[];
```

Program 'izrezivanje_AG.m' priprema podatke u oblik pogodan za statističku obradu. Otvara se datoteka iz koje se čitaju podaci i nova u koju se upisuju obrađeni podaci. Na početak nove datoteke upisuju se atributi i klasa, a nakon toga i skraćeni vektori, duljne 140 nukleotida međusobno odvojenih zarezima. Na kraj vektora se upisuje F ili T. Pri ispitivanju se provjeravaju vrijednsti labele i ispituju se samo pozicije 189 i 199, onosno nalazi li se na tim pozicijama kombinacija AG.

'izrezivanje_AG.m':

```
% Rezervacija prostora za:

    % podatke
    %podaci = char(zeros(1000, 140));
    %podaci(900000,:) = char(zeros(1, 140));

    % zastavice tocno / netocno
    tocnost = char(zeros(1000, 1));
    tocnost(900000,1) = char(0);

    % pozicije u genomu
    pozicija = zeros(900000,1);

% Ucitavanje datoteka i pohrana podataka

% Open file
fid = fopen('C_elegans_don_all_examples.fasta', 'r');
fid2= fopen('C_elegans_stari.txt','w');

% header
fprintf(fid2, '@relation C_elegans_toc+netoc.txt\n');
for n=1:140
    fprintf(fid2, '@attribute NA%d {A,T,G,C}\n', n);
end
fprintf(fid2, '@attribute class {F,T}\n');
fprintf(fid2, '\n');
fprintf(fid2, '@data\n');
```

```
% Read input data from the FASTA file
l=1;
for k=1:Inf
     if (mod(k,100000)==0)
        a={'gegam se'}
        k/100000
    end
    % Read header
    head = fgets(fid);
    if isnumeric(head), break, end % end of the file

     % Chromosome number
    ind = regexp(head, 'chr_num\s*=\s*[+-\d]', 'end');
    len = regexp(head(ind:end), '[+-]?\s*\d*', 'once', 'end');
    chr_num = str2double(head(ind:ind+len-1));

     % Strand
    ind = regexp(head, 'strand\s*=\s*[+-]', 'end');
    strand = head(ind);

    % Position
    ind = regexp(head, 'position\s*=\s*[+-\d]', 'end');
    len = regexp(head(ind:end), '[+-]?\s*\d*', 'once', 'end');
    position = str2double(head(ind:ind+len-1));

    % Label
    ind = regexp(head, 'label\s*=\s*[+-\d]', 'end');
    len = regexp(head(ind:end), '[+-]?\s*\d*', 'once', 'end');
    label = str2double(head(ind:ind+len-1));

    % Read data
    data = fgets(fid);
    if isnumeric(head), break, end % end of the file


    % Do something
    if (data(198) == 'A' && data(199) == 'G')
      if (label==1)
          tocnost(l) = 'T';
          fprintf(fid2,
'%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,', data([118:197,200:259]));
          fprintf(fid2, '%c\n', tocnost(l));
          pozicija(l) = position;
          l = l+1;
      else
          tocnost(l) = 'F';
         fprintf(fid2,
'%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
```

```
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c
,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,%c,', data([118:197,200:259]));
          fprintf(fid2, '%c\n', tocnost(l));
          pozicija(l) = position;
          l = l+1;

       end
     end
end

% Close file
fclose(fid);
fclose(fid2);

% Uklanja visak prostora
tocnost (l:end,:)=[];
%podaci  (l:end,:)=[];
pozicija(l:end,:)=[];
```

Nakon što su podaci obrađeni, prvo pripremljeni u oblik pogodan za obradu statističkim metodama, a zatim obrađeni statističkom metodom 'Random Forest', potrebno je nacrtati krivulju koja grafički pokazuje kvalitetu našeg detektora.

‘crtamo.m’:

```
% Open file
fid = fopen('glasovi', 'r');      %('glasovi.txt','r');
fid2 = fopen('sume_izrezivanje.txt', 'r');

% upis 1. i 2. stupca matrice
br_redova=0;
mat=zeros(900000, 5);
for i=1:900000    %for i=1:Inf
    s=fgets(fid);
    if isnumeric(s), break, end % end of the file
    a=((s(16)-48)*10000+(s(18)-48)*1000+(s(19)-48)*100+(s(20)-
48)*10+(s(21)-48))/10000;
    mat(i,1)=a;
    a=((s(26)-48)*10000+(s(28)-48)*1000+(s(29)-48)*100+(s(30)-
48)*10+(s(31)-48))/10000;
    mat(i,2)=a;
end
br_redova=i;
mat(i+1:end,:)=[];
b={'upisan je prvi i drugi stupac'}

% upis 5. stupca matrice - iz originalne datoteke
i=1;
for k=1:(144+br_redova)             %slova pocinju od 145.reda (i=1)
```

```
   head = fgets(fid2);
   if isnumeric(head), break, end
   if (k>144)
       a=head(281);
       if (a=='T')
            mat(i,5)=1;
       else
           mat(i,5)=0;
       end
       i=i+1;
   end
end
b={'upisan je peti stupac'}

% upis u 3. i 4. stupac matrice
prag=0; precision=0; recall=0;
for br=1:100
    sumTP=0; sumTN=0; sumFP=0; sumFN=0;
    for i=1:br_redova
            mat(i,3)=mat(i,2)/(mat(i,1)+mat(i,2));
            if (prag<mat(i,3))
                mat(i,4)=1;
            else
                mat(i,4)=0;
            end

            if(mat(i,4)==1 && mat(i,5)==1)
               sumTP=sumTP+1;
            end

            if(mat(i,4)==0 && mat(i,5)==1)
                sumFN=sumFN+1;
            end

            if(mat(i,4)==1 && mat(i,5)==0)
                sumFP=sumFP+1;
            end

            if(mat(i,4)==0 && mat(i,5)==0)
                sumTN=sumTN+1;
            end
        end
    precision(br)=sumTP/(sumTP+(sumFP));
    recall(br)=sumTP/(sumTP+sumFN);
    prag=prag+0.01;
end
b={'upisan je treci i cetvrti stupac'}

% crtanje precision-recall krivulje
plot(recall,precision,'r')
xlabel('RECALL');
ylabel('PRECISION');

fclose(fid2);
fclose(fid);
```