SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1143

Analiza DNK metodama obrade signala u vremenskoj i frekvencijskoj domeni

Nika Parađina

Zagreb, rujan 2008.

Hvala mentoru prof.dr.sc.D.Seršiću na pomoći, kolegama na konstruktivnim savjetima i svima ostalima koji su se trudili shvatiti o čemu pišem.

Sažetak

Pod „predikcijom gena“ ovdje se podrazumijeva određivanje položaja i granica nekodirajućih i kodirajućih regija unutar gena. DNK lanac se dijeli na gene i međugenski prostor. Geni eukariota, organizama koji unutar stanice imaju jezgru, se sastoje od eksona i introna. Eksoni upravljaju sintezom proteina,a proteini reguliraju biološke funkcije organizma.

Neki se autori oslanjaju na tvrdnju da kodirajuća područja tj. eksoni pokazuju svojstvo perioda tri dok introni tj. nekodirajuće regije nemaju niti jedan period izraženiji od ostalih.[1] [3] Na temelju tog svojstva razvio se niz tehnika koji, koristeći to predznanje, pokušava izdvojiti eksone od nekodirajućih regija. Ovdje su prezentirane četiri metode koje se temelje na periodu tri. Sve četiri su izrasle iz usko pojasnopropusnog filtra tzv. antinotch filtra s centralnom frekvencijom $\frac{2\pi}{3}$. Zato što geni nisu savršeno pravilni niti svi pokazuju jednaka svojstva, uspješnost ovih metoda zavisi o analiziranom uzorku.

I introni imaju svoje pravilnosti. Te pravilnosti odnose se na konsenzus sekvenci nukleotida na početku i na kraju introna. Otkrivanje tih sekvenci omogućava otkrivanje granica između kodirajućih i nekodirajućih područja. Neuronska mreža, prethodno trenirana na uzorcima iz javnih baza genoma, odlučuje da li je slijed baza dio tražene sekvence. Na taj način se pronalaskom introna određuju pozicije eksone koji se nalaze između njih.

Niti metode bazirane na eksonima niti metode koje se oslanjaju na svojstva introna ne rade savršeno svaka za sebe. Njihova točnost ovisi o genu koji se analizira. Spajanjem tih dviju metoda u jedan sustav mogle bi se iskoristiti njihove dobre karakteristike i na taj način dobiti preciznija detekcija kodirajućih i nekodirajućih regija. Upravo ta kombinacija svih mogućih znanja o građi gena je i osnova današnjih sustava za detekciju eksona i u tom smjeru treba ići daljnje istraživanje ovo teme.

Sadržaj

Popis oznaka i kratica

A	adenin
G	guanin
C	citozin
T	timin
U	uracil
DNK	deoksiribonukleinska kiselina
RNK	ribonukleinska kiselina
Ala	alanin
Arg	arginin
Asn	asparagin
Asp	asparaginska kiselina
Cys	cistein
Phe	fenilalanin
Gln	glutamin
Glu	glutaminska kiselina
Gly	glicin
His	histidin
Ile	izoleucin
Leu	leucin
Lys	lizin
Met	metionin
Pro	prolin
Ser	serin
Tyr	tirozin
Thr	treonin
Trp	triptofan
Val	valin

Popis tablica i slika

1.Uvod

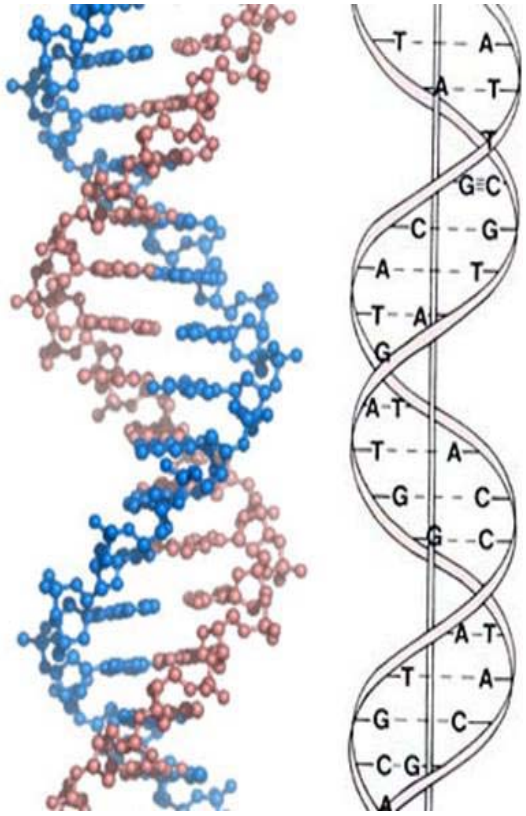
Broj javno dostupnih potpunih genoma je znatno narastao u zadnjih nekoliko godina. Razumijevanje prirode tih informacija i utvrđivanje uloge gena u određivanju bioloških funkcija postao je multidisciplinarni istraživački problem. Identifikacija gena podrazumijeva pronalaženje kompletne strukture gena, posebno preciziranje granica kodirajućih i nekodirajućih regija. DNK sekvenca matematički se može reprezentirati nizom od četiri karaktera A,T,C i G. Takav tip podataka omogućuje korištenje metoda digitalne obradbe signala u predikciji gena.

Uvijek će biti neophodno provođenje molekularnih eksperimenata da bi se dokazala lokacija eksona, učestalost transkripcije, uzorak izrezivanja i stupanj ekspresije nekog gena, ali računalne metode mogu unaprijed ponuditi pouzdanu indikaciju položaja eksona unutar gena te time smanjiti potreban opseg eksperimenata i njihove troškove.

Bez sumnje, područje istraživanja genoma je u interesu cijelog ljudskog roda jer osim što nam nudi informacije o nama samima otvara i put liječenja i prevencije bolesti na razini gena. Zato i postoji potreba za razmjenom znanja i rezultata iz svih znanstvenih grana.

2. Biološke osnove

DNK molekula se sastoji od dva spiralno povezana komplementarna DNK lanca. Lanac je sastavljen od četiri nukleotida ili baza: adenina, timina, citozina i guanina, redom A,T,C i G. Povezani su tako da lijevi kraj jednog nukleotida tvori čvrstu kovalentnu vezu s desnim krajem sljedećeg na taj način tvoreći lanac. Unutar DNK molekule lanci se povezuju na način da se adenin spaja s timinom, a guanin veže s citozinom slabim hidrogenskim vezama. Iako su te veze slabe, ima ih mnogo te je dobivena spiralna struktura DNK molekule stabilna.



Slika 1: DNK molekula

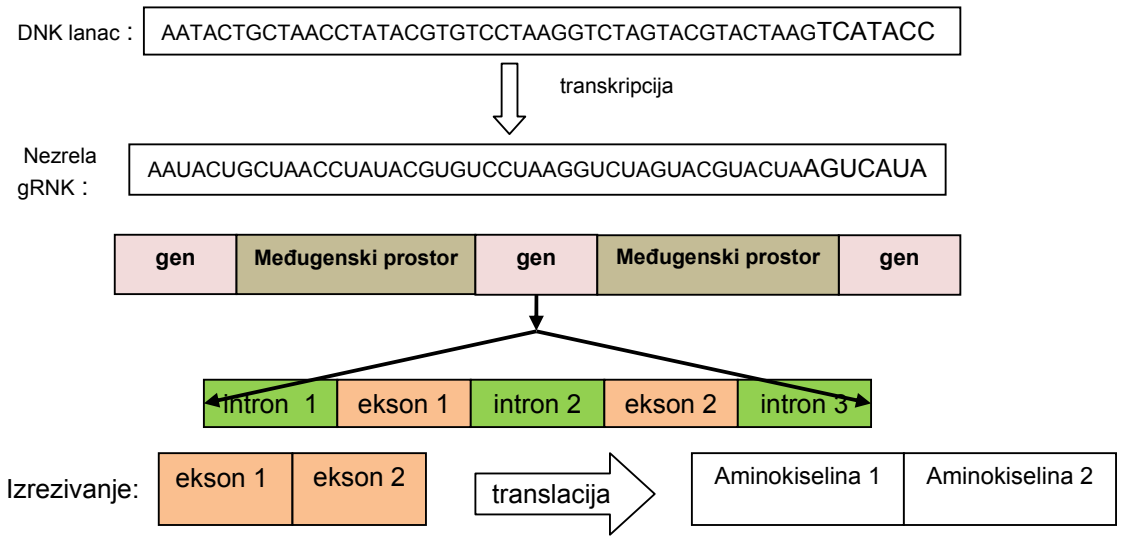
Protein je biomolekula koju tvori niz aminokiselina. Proteini su odgovorni za odvijanje i/ili ubrzavanje gotovo svih bioloških funkcija. Čak i sintezu proteina potiču sami proteini. Proteini i njihova funkcija određeni su nukleotidima u DNK lancu jer redoslijed baza diktira redoslijed aminokiselina.

Ako promatramo izolirano područje DNK molekule u nekom trenutku, u sintezi proteina biti će aktivan samo jedan lanac, dok su slučajevi kada su oba lanca istovremeno aktivna vrlo rijetki. [1] Svaki lanac može se podijeliti na gene i međugenski prostor. Samo geni sudjeluju u sintezi proteina. Iako sve stanice jednog organizma imaju jednake gene, za svaku skupinu stanica aktivan je samo određeni podskup gena, što im uz funkciju određuje i oblik. [2].

Geni eukariota, organizama čije stanice sadrže jezgru, građeni su od eksona i introna. Eksoni kodiraju proteine dok se smatra da introni imaju posredničku i upravljačku funkciju, makar postoje teorije da bi i introni mogli biti odgovorni za

ekspresiju proteina te time utjecati na razvoj bolesti. Određivanje biološke funkcije gena zapravo se svodi na detekciju eksona i određivanje intron-ekson granice. Prokarioti, organizmi čije stanice ne sadrže jezgru, nemaju introne, već samo eksone, zbog čega je jednostavnije odrediti kodirajuće dijelove takve DNK molekule.

Proces sinteze proteina može se podijeliti na tri glavna koraka: transkripciju, izrezivanje (*eng. splicing*) te translaciju.
Transkripcija je proces stvaranja nezrele glasničke RNK molekule (*eng. immature messanger RNA* ili *mRNA*) koja je preslika DNK lanca uz zamjenu timina uracilom. Takva mRNK ulazi u proces izrezivanja, gdje se posredstvom malih jezgrenih ribonukleinskih proteina (*eng. small nucleo ribonucleoproteins* ili *snRNP*) uklanjaju introni, a eksoni se spajaju jedan na drugi. Tako dobivena glasnička RNK je sada zrela.
Ona putuje izvan jezgre u citoplazmu gdje se u ribosomu na temelju nje sintetiziraju lanci aminokiselina tvoreći protein. U sintezi proteina u ribosomu sudjeluje i prijenosna RNK (*eng. transfer RNA* ili *tRNA*). Taj korak naziva se translacija.



Slika 2: Prikaz procesa sinteze proteina

Metode otkrivanja kodirajućih područja (eksona) unutar gena mogu se podijeliti na one koje se baziraju na svojstvima samih eksona i na one koje se oslanjaju na specifičnosti građe introna.

3. Detektori temeljeni na svojstvima eksona

Eksoni, kao i cijeli DNK lanac, se sastoje od četiri baze, adenina, guanina, citozina i timina. Po tri baze čine jedan kodon koji odgovara jednoj aminokiselini. To daje 4^3 tj. 64 moguća kodona, od kojih 3 otpadaju na stop kodone, a jedan na start kodon. Isti kodon koji označava početak procesa translacije ukoliko se nalazi unutar eksona označava aminokiselinu metionin. Drugim riječima, 61 mogući kodon određuje 20 aminokiselina, što vodi na zaključak da više kodona označava jednu istu aminokiselinu.

Pokazalo se da su sekvence nukleotida u intronima slučajno raspoređene dok u eksonima postoji primjetna periodičnost s tri. Pretpostavlja se da je uzrok te periodičnost u učestalosti pojave pojedinih kodona, koji se obično javljaju u uzorku *RNY*, gdje je *R* A ili G, *N* C ili T, a *Y* može biti bilo koja od baza.[4] Zašto priroda preferira upotrebu nekih kodona u odnosu na druge, do danas ostaje neriješeno pitanje.

Da bi za detekciju eksona mogli koristiti standardne postupke obradbe signala mora se ulazni niz podataka koji je sam gen zapisan kao niz od četiri karaktera, A,C, T i G pretvoriti u prikladniji oblik. To znači da se jedan gen zapisuje kao četiri indikatorske binarne sekvence, takve da se za niz nukleotida AATCTGCTACTCT.... dobije niz $x_A(n)$=1100000010000...., gdje 1 označuje prisutnost adenina u genu, a 0 označava bilo koju od preostalih baza. Na analogan način se stvaraju binarni nizovi $x_T(n)$, $x_C(n)$ i $x_G(n)$, a n označava položaj baze u genu gledano s lijeva na desno.

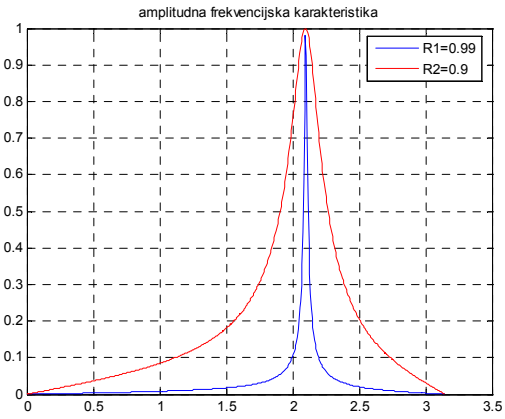
Za proučavanje periodičnosti s tri najprije je upotrjebljena FFT na pomičnom otvoru. Makar je i tom metodom dokazano postojanje periodičnosti i time dobivena dobra pretpostavka o položaju eksona unutar gena, metode koje su se dalje razvijale iz FFT daju preciznije rezultate.

3.1. Metoda s antinotch filtrom [1]

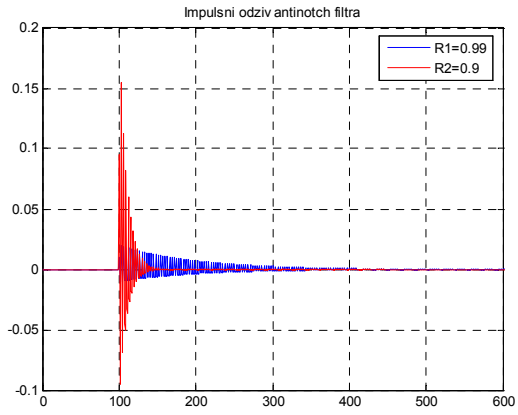
Antinotch filtar je usko pojasnopropusni filtar koji propušta samo odabranu frekvenciju. Njegova prijenosna funkcija je:

$$H(z) = \frac{1}{2}\frac{(1-R^2)+(R^2-1)z^{-2}}{1-2R\cos\theta z^{-1}+R^2 z^{-2}}, \quad (1)$$

θ je $\frac{2\pi}{3}$ jer se na taj način iz sekvence gena izvlače regije s karakterističnim periodom 3, tj. eksoni. Propusnost filtra regulira se odabirom parametra R. Što je R bliži 1, to će pojas propuštanja biti uži, ali će i vrijeme istitravanja impulsnog odziva filtra biti duže. Na slikama 3 i 4 prikazane su amplitudno frekvencijska karakteristika filtra i njegov impulsni odziv.



Slika 3: Amplitudno frekvencijska karakteristika antinotch filtra za R=0.99 i R=0.9

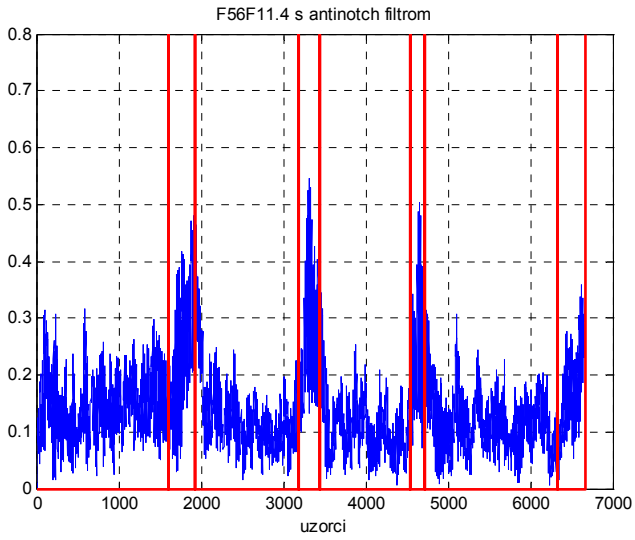


Slika 4: Impulsni odziv antinotch filtra za R=0.99 i R=0.9

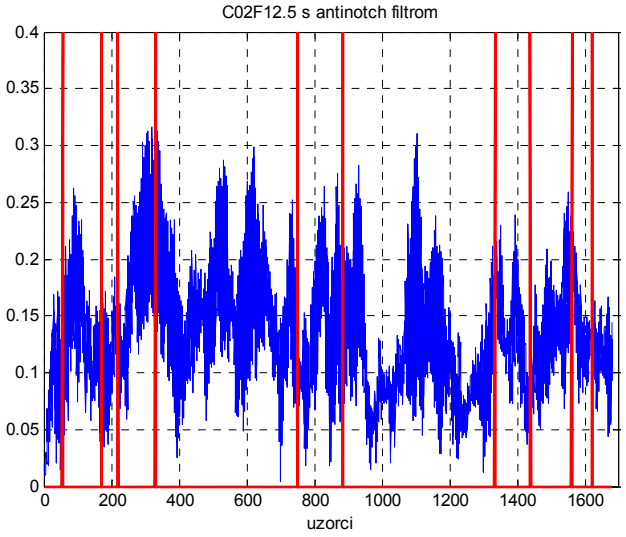
Da bi mogli što preciznije izolirati područja koja pripadaju eksonima, u interesu nam je napraviti što selektivniji filtar, ali selektivnost filtra u frekvencijskom području umanjuje rezoluciju u vremenskoj domeni.

Antinotch filtar treba primijeniti na svaku od četiri binarne sekvence dobivene iz gena. Njihovom linearnom kombinacijom dobivamo raspodjelu energije po cijelom genu. Područja veće energije odgovaraju eksonima, dok introni imaju znatno nižu energiju.

Slika 5 prikazuje rezultat primjene antinotch filtra na gen F56F11.4 uz R=0.992. Crveni kvadrati na slikama označavaju točna područja eksona. Odstupanje se događa zbog dugog impulsnog odziva antinotch filtra koji „razmazuje“ bridove. Problem koji se nameće je primjenjivost ove metode. Naime, zbog već spomenutog dugog impulsnog odziva antinotch filtra, razmak introna i eksona unutar gena mora biti takav da se filtar uspije istitrati prije nailaska na drugi ekson. Pretpostavka je da unutar eksona ne postoje područja izraženoga perioda tri. Slika 6 prikazuje primjer gene C02F12.5 koji ne odgovara navedenim uvjetima i za takav slučaj je ova metoda neprimjenjiva.

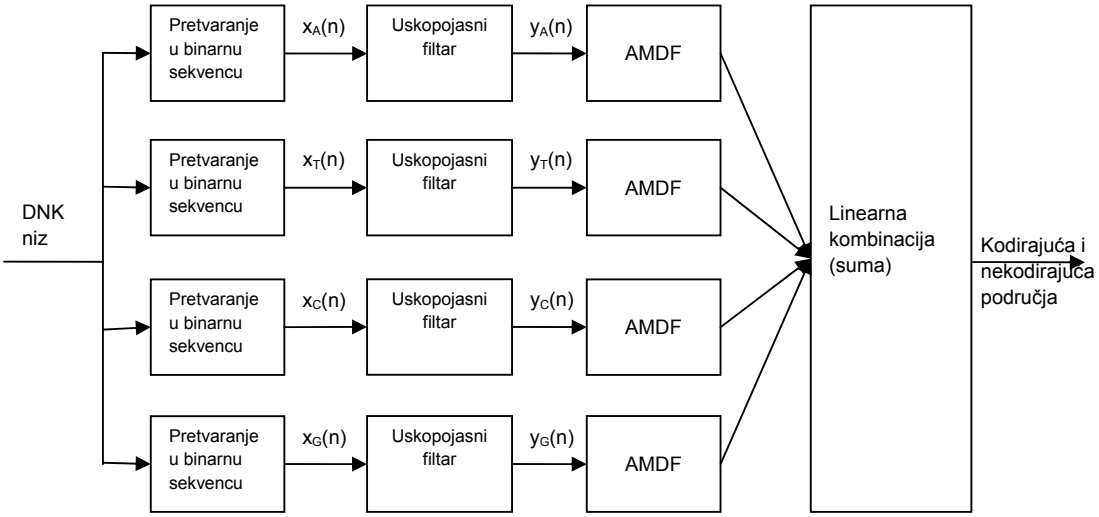


Slika 5: Raspored regija gena F56F11.4 dobiven antinotch filtrom



Slika 6: Raspored regija gena C02F12.5 dobiven antinotch filtrom

3.2. Funkcija srednje razlike magnitude AMDF (eng. Average Magnitude Diference Function) [4]



Slika 7: Blok shema AMDF metode

Ova metoda je nadogradnja na prethodno spomenutu metodu. Sekvenca gena ponovno se mora rastaviti na iste četiri binarne indikatorske sekvence na već opisan način. Ti binarni nizovi zatim prolaze kroz uskopojasni filtar s centralnom frekvencijom $\frac{2\pi}{3}$ čime se naglašavaju kodirajuće regije zbog periodičnosti sa tri, a umanjuje značaj nekodirajućih regija. Prijenosna funkcija filtra te parametri jednaki su kao u izrazu (1) .

Funkcija srednje razlike magnituda dugo se vremena koristila u obradi govornih signala prije nego što je njezina uloga zamijenjena drugim metodama. Pokazuje se korisnom za obradu periodičkih i neperiodičkih signala koji su duži od perioda koji promatramo. Signali koji se dobiju propuštanjem binarnih nizova kroz usko pojasnopropusni filtar obrađuju se po segmentima AMDF funkcijom koja je definirana na način:

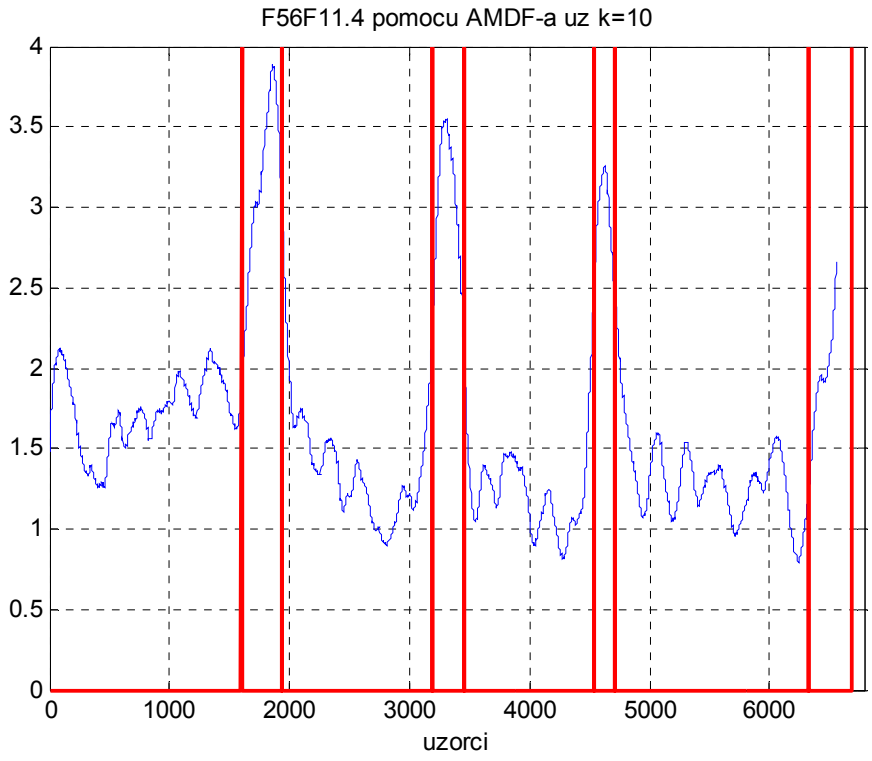
$$AMDF(k) = \frac{1}{N}\sum_{n=1}^{N}\left|y(n) - y(n-k)\right|, \quad (2)$$

N određuje duljinu segmenta. Pokazalo se da je najbolje uzeti broj manji od očekivane veličine eksona, a koji je višekratnik broja 3. Zato je odabran N=117.

Ako se upotrijebi k=3, u idealnom slučaju, očekivano je da će područja eksona, nakon obrade AMDF-om, imati vrijednosti jednake nuli dok će ostali nekodirajući dijelovi gena imati vrijednosti različite od nule. Međutim, u realnom slučaju, periodičnost s tri može se javiti i u nekodirajućim područjima, dok se u područjima eksona nalaze i periodi različiti od tri. Zato nije za očekivati da će na kodirajućim područjima vrijednosti biti jednake nuli.

Bolji način korištenja ovog sustava jer postavljanje parametra k na period različit i veći od 3. U tom slučaju će se u regijama eksona dobiti puno veće vrijednosti nego u preostalim nekodirajućim dijelovima gena. Iako unutar eksona postoje i drugi periodi, period tri je dominantan.

Na slici 8 prikazani su rezultati dobiveni za gen F56F11.4 uz k=10.



Slika 8: Raspored regija gena F56F11.4 dobiven AMDF metodom

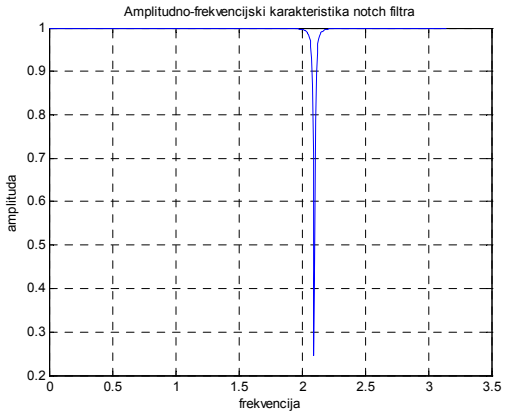
3.3. Detektor s kvocijentom ovojnica

Sve dosad prikazane metode su, uz nezaobilazno postavljena ograničenja ispitivanih gena, imala veliki problem vezan uz loše performanse u vremenskoj domeni usko pojasnopropusnog filtra. Ova metoda bavi se razvojem sustava koji se u osnovi temelji na antinotch i notch filtru. Oni zajedno sa svojim antikauzalnim varijantama čine sustav koji je precizan i u vremenskoj i u frekvencijskoj domeni.

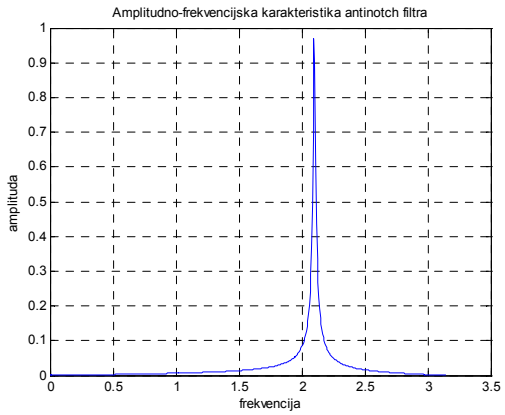
Antinotch i notch filtri su komplementi, frekvenciju koju antinotch propušta, notch guši. Transfer funkcija antinotch filtra navedena je u dijelu 3.1, a transfer funkcija notch filtra je:

$$G(z) = \frac{1}{2}\frac{1 + R^2 - 4R\cos\theta z^{-1} + (1 + R^2)z^{-2}}{1 - 2R\cos\theta z^{-1} + R^2 z^{-2}} \quad (3)$$

Amplitudno-frekvencijske karakteristike antinotch i notch filtara prikazani su na slikama 9 i 10.



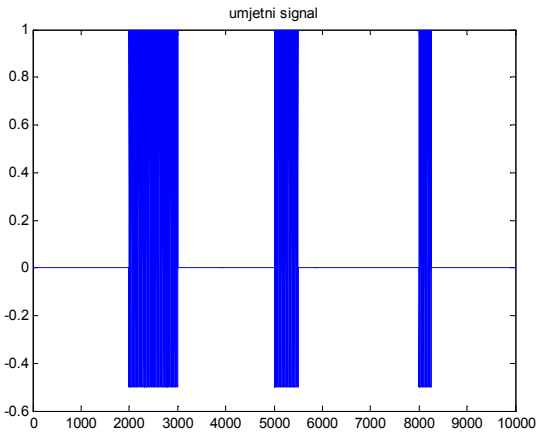
Slika 9: Amplitudno frekvencijska karakteristika notch filtra



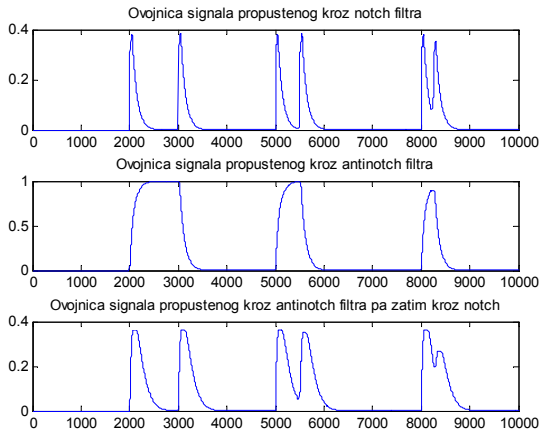
Slika 10: Amplitudno frekvencijska karakteristika antinotch filtra

3.3.1 Konstruirani signal

Za demonstraciju ideje ovoga sustava, poslužiti će idealni konstruirani signal bez šuma. Slika 11 prikazuje taj signal. Signal je u svim trenucima jednak nuli osim kada se u njemu pojavljuju tri segmenta sinusoidnih signala frekvencije $\frac{2\pi}{3}$ koji su lokalizirani u različitim vremenskih trenutcima te su različitog trajanja.



Slika 11: Konstruirani idealni signal bez dodanog šuma



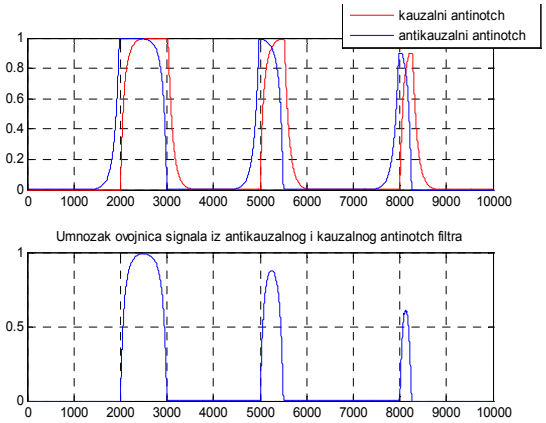
Slika 12: gore: ovojnica signala nakon prolaza kroz notch, sredina: ovojnica signala nakon prolaza kroz antinotch, dolje: ovojnica signala nakon prvo prolaska kroz antinotch, a zatim kroz notch

Nakon prolaza kroz antinotch filtar, iz gornjih slika može se primijetiti da je lijevi brid dobivene amplitudne ovojnice signala točno određen, dok je desni zbog već navedenog dugog istitravanja antinotch i notch filtara neprecizan. Ovojnica signala određuje se usrednjavanjem apsolutnih vrijednosti signala na izlazu iz filtara pomoću median funkcije preko 99 uzoraka.

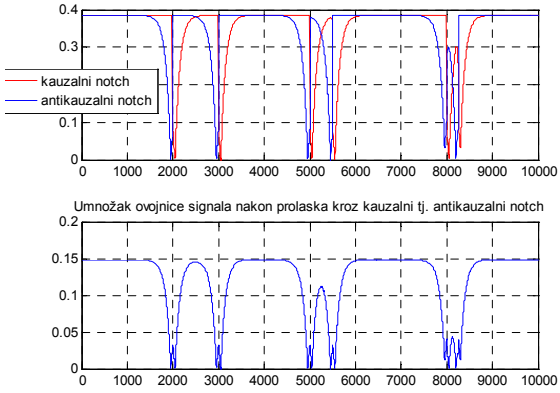
Nameće se ideje da bi se za antikauzalni filtar dobila obrnuta situacija te da bi umnožak amplitudnih ovojnica signala koji su prošli kroz kauzalni i antikauzalni antinotch filtar rezultirao preciznije određenim signalom, što se može primijetiti na slici 13. Neka se tako dobiveni signal zove $y_1(n)$.

Na slici 14 prikazan je signal, $y_2(n)$, dobiven množenjem ovojnice signala koji je prošao kroz kauzalni notch filtar i koji je zatim komplementiran, s ovojnicom signala koji se dobije prolaskom kroz antikauzalni notch filtar i komplementiranjem. Komplement ovojnice signala dobije se oduzimanjem iste od maksimalne vrijednosti te ovojnice.

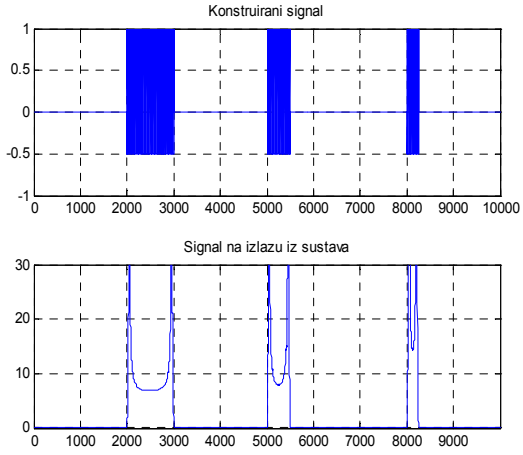
Dijeljenje signala $y_1(n)$ s $y_2(n)$ daje precizno određene regije pojave frekvencije $\frac{2\pi}{3}$ u konstruiranom signalu, što je prikazano na slici 15.



Slika 13: gore: prikaz ovojnice signala nakon prolaza kroz kauzalni i antikauzalni antinotch filtar, dolje: signal nakon umnoška gornjih ovojnica



Slika 14: gore: prikaz ovojnice signala nakon prolaza kroz kauzalni i antikauzalni notch filtar, dolje: signal nakon umnoška gornjih ovojnica



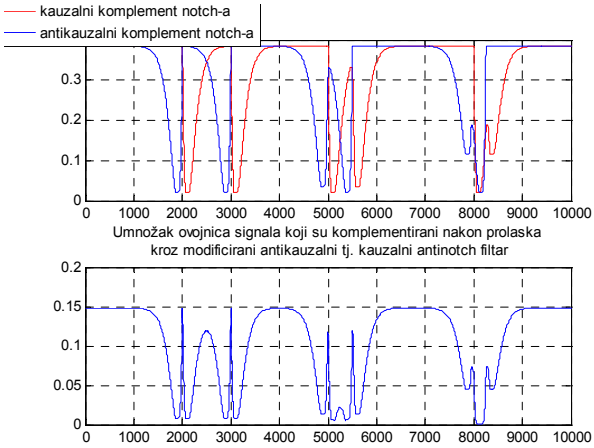
Slika 15: gore: konstruirani idealni signal, dolje: raspored regija dobiven $y_1(n)/y_2(n)$

Problem kod realnih signal je što unutar eksona postoje i ostale frekvencije osim $\frac{2\pi}{3}$.

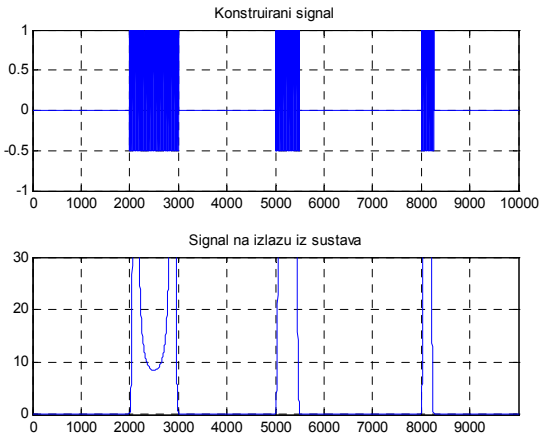
Sukladno tome notch filtar će u području eksona ugušiti periodičnost sa tri, ali zbog prisutnosti drugih frekvencija amplituda ovojnice signala na izlazu iz filtra neće ići u nulu unutar željenog područja. Iz tog razloga se u notch filtar dovodi signal koji je prethodno prošao kroz antinotch filtar jer antinotch filtar guši sve frekvencije različite od $\frac{2\pi}{3}$. Iako se time kvare karakteristike notch filtra, preinake su nužne da bi on bio primjenjiv za analizu realnih signala.

Zatim se od ovojnice ovako dobivenog signala radi svojevrstan komplement na način da se ona oduzme od maksimalne vrijednosti ovojnice signala koji je prošao samo kroz notch filtar jer se na taj način izbjegava dijeljenje s nulom u daljnjem postupku. Analogan postupak se primjenjuje i za antikauzalne filtre.

Neka se umnožak amplitudnih ovojnica komplementiranih signala koji su dobiveni prolaskom tih signala kroz na gornji način modificiran antikauzalni odnosno kauzalni filtar zove $y_3(n)$ (slika 16). Rezultat dijeljenja signala $y_1(n)$ s $y_3(n)$ prikazan je na slici 17 zajedno s originalnim signalom i to je ujedno i rezultat detektora s kvocijentom ovojnica.



Slika 16: gore: prikaz ovojnice signala nakon prolaza kroz komplement kauzalnog i antikauzalnog notch filtra, dolje: signal nakon umnoška gornjih ovojnica

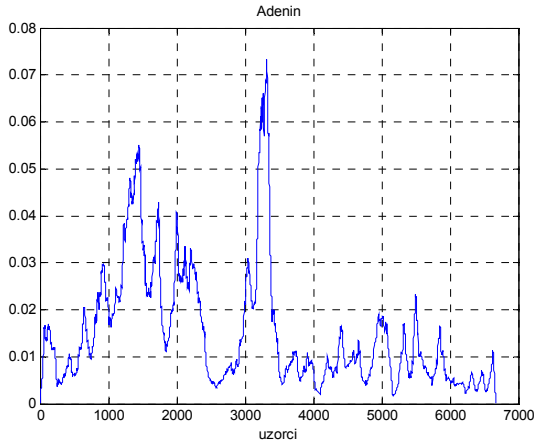


Slika 17: gore: konstruirani idealni signal, dolje: raspored regija dobiven detektorom s kvocijentom ovojnica

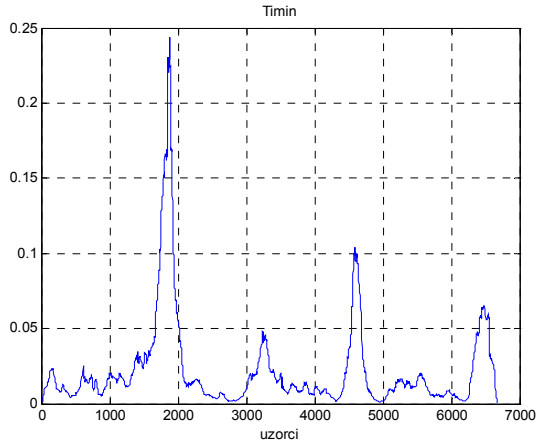
3.3.2 Realni signal

Za realni signal navedeni sustav se prvo primjeni na svaki od četiri binarna niza $x_A(n)$, $x_T(n)$, $x_C(n)$ i $x_G(n)$ koji predstavljaju odgovarajuće baze, a zatim se dobiveni rezultati zbrajaju da bi se dobila konačna slika raspodjele kodirajućih i nekodirajućih područja unutar gena. Za pojedine gene možda će neka baza samostalno dati bolju detekciju nego zbroj svih baza, ali se za nepoznati gen ne može unaprijed znati koja od baza je najindikativnija te se zato vrši superpozicija svih rezultata.

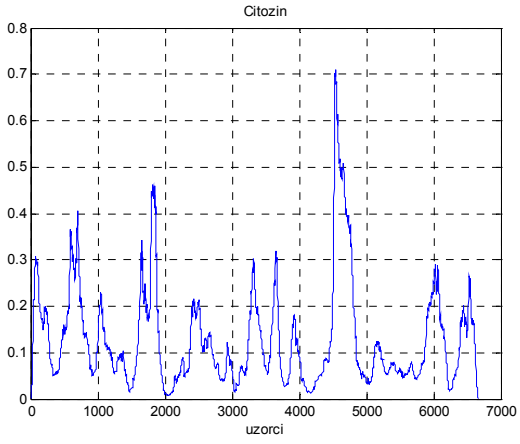
Na donjim slikama (slike 18 do 22) prikazani su signali na izlazu iz gore opisanog sustava za svaku od binarnih baznih sekvenci te finalni rezultat na kojemu su crvenim pravokutnicima određena točna područja eksona.



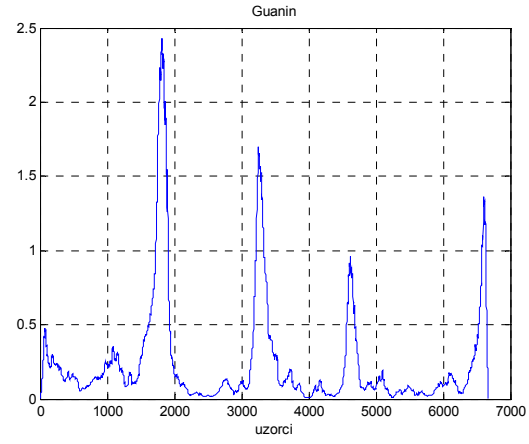
Slika 18: Baza adenin F56F11.4- raspored regija



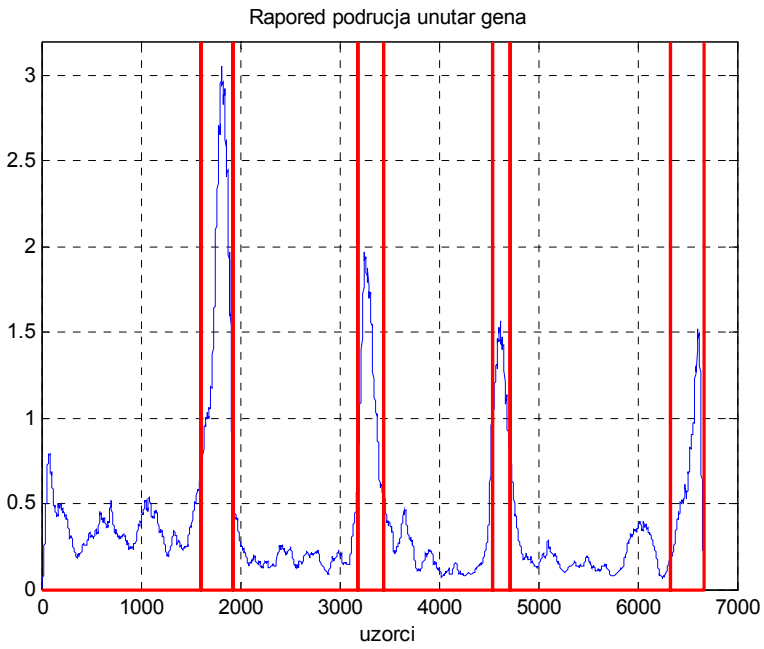
Slika 19: Baza timin F56F11.4- raspored regija



Slika 20: Baza citozin F56F11.4- raspored regija



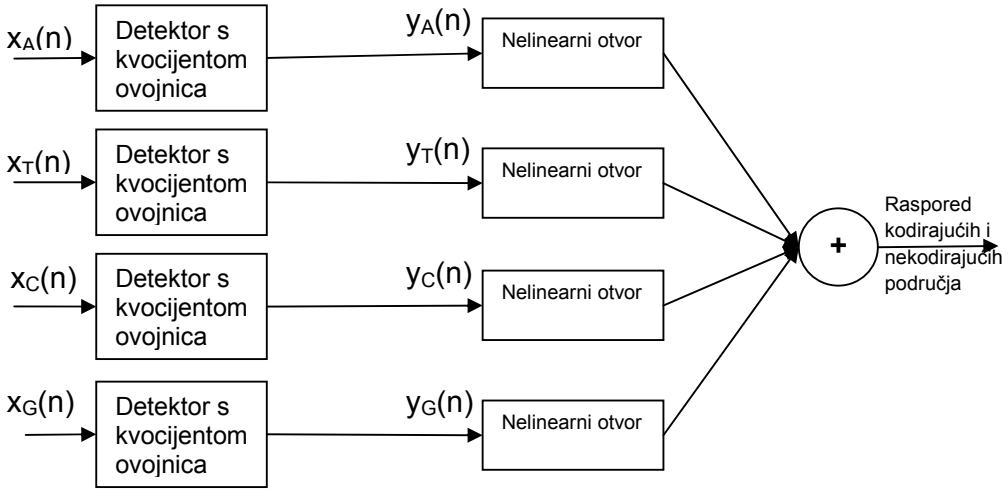
Slika 21: Baza guanin F56F11.4-raspored regija



Slika 22: Raspored regija za gen F56F11.4 za linearnu kombinaciju adenina, timina, citozina i guanina

3.4. Nadograđeni detektor s kvocijentom ovojnica

Pregledniji rezultati dobiju se ako se u gore opisani sustav nakon svakog koraka obrade doda nelinearni pomični otvor koji služi potiskivanju nekodirajućih područja [7].Ta ideja prikazana je na slici 23.



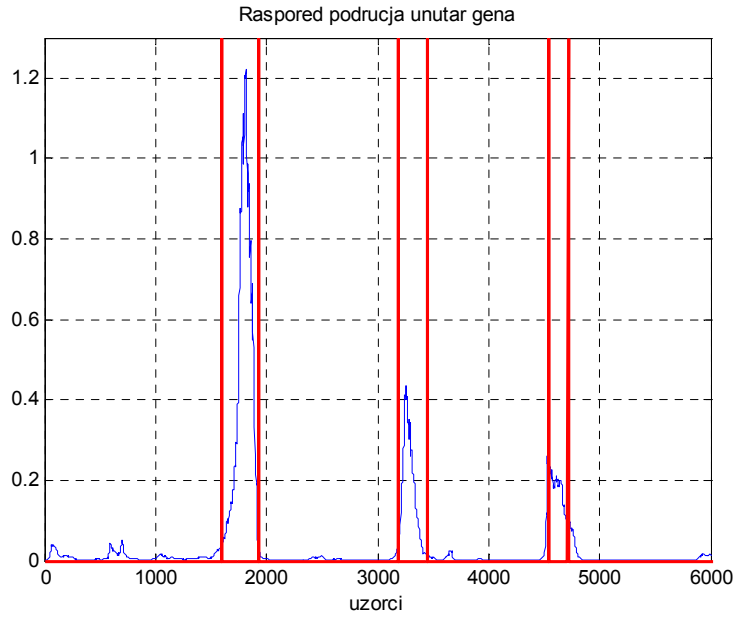
Slika 23: Blok shema nadopunjene antinotch-notch metode

Nelinearni otvor može se definirati izrazom (4) :

$$y_{novi}(n) = \left(\frac{y_{stari}(n)}{M} \right)^2 \square y_{stari}(n), \quad 1 \le n \le N \quad (4)$$

gdje je M najveća vrijednost uzorka unutar prozora, n je redni broj uzorka, a N broj uzoraka u jednom prozoru. Idealno bi bilo odabrati N veći od najvećega eksona, a manji od najmanje nekodirajuće regije. Nažalost, nije uvijek moguće ispuniti takve uvjete. Zadovoljava N koji je veličine najmanjeg eksona.

Prozor radi tako da naglašava velike amplitude, koje su očekivane u područjima eksona, a potiskuje ionako male vrijednosti unutar nekodirajućih regija. Ukoliko se unutar nekodirajućeg područja javi kratkotrajna periodičnost sa tri, javit će se i porast amplitude, ali manji nego unutar kodirajuće regije. Tada nelinearni prozor djeluje tako da taj mali porast eliminira. Prema shemi na slici 22 prozor se nalazi grani, pa se time umanjuje i mogućnost dobivanja regije velike amplitude unutar introna kao posljedice sumiranja rezultata. Ako se nakon cijelog sustava dobije krivo registrirano kodirajuće područje, to znači da testirani uzorak ne odgovara pretpostavkama sa samog početka. Rezultat primjene ovog nadopunjenog sustava na gen F56F11.4 prikazan je na slici 24.



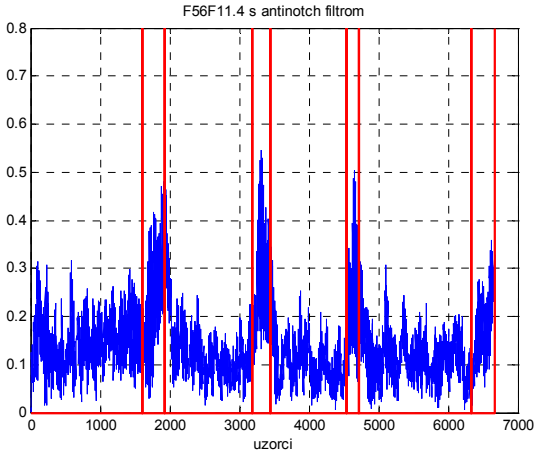
Slika 24: Raspored regija gena F56F11.4 dobiven nadograđenim detektorom s kvocijentom ovojnica

4. Opis i usporedba rezultata

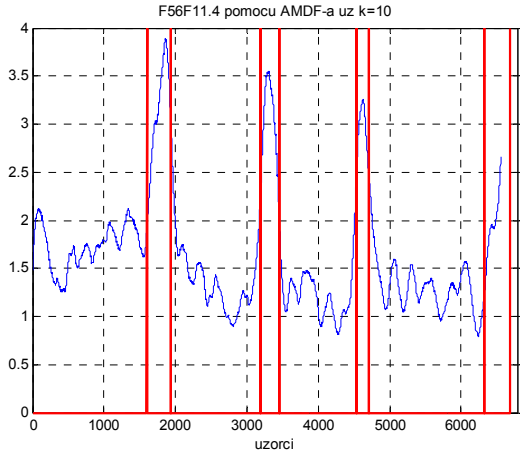
1. Gen F56F11.4

Gen F56F11.4 ima 4 eksona čija je najveća duljina manja od minimalne duljine nekodirajućih regija oko eksona. Nekodirajuće regije ne pokazuju svojstvo periodičnosti sa tri, dok je isto svojstvo jako izraženo unutar eksona. Zbog ovih osobina očekivano je da će i najgrublja metoda pružiti rezultate pomoću kojih će se moći pretpostaviti područje eksona.

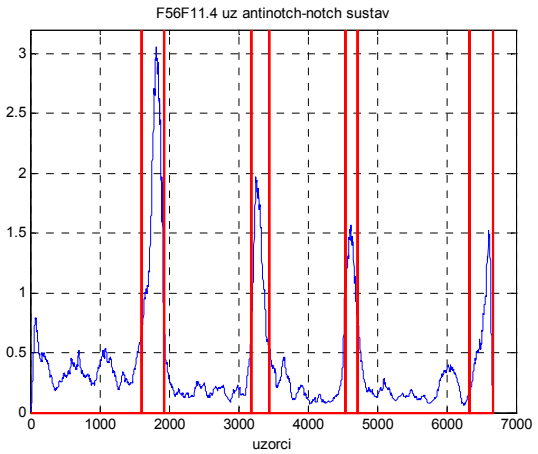
Najlošije pozicioniranje dobiva se korištenjem antinotch filtra u njegovom izvornom obliku, a najbolje rješenje koristeći detektor s kvocijentom ovojnica sa izglađivanjem pomoću nelinearnog prozora. (slike 25-28)



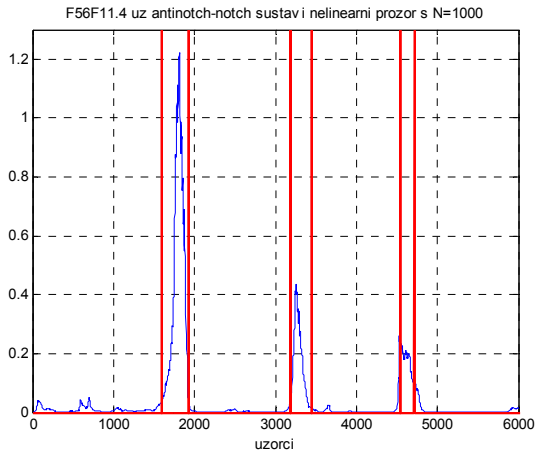
Slika 25: Raspored regija gena F56F11.4 dobiven antinotch filtrom



Slika 26: Raspored regija gena F56F11.4 dobiven AMDF metodom



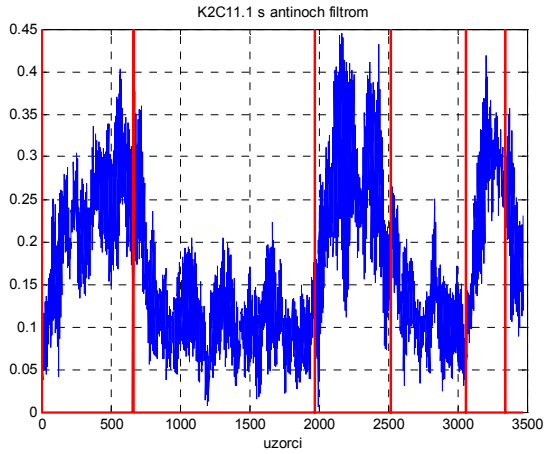
Slika 27: Raspored regija gena F56F11.4 dobiven detektorom s kvocijentom ovojnica



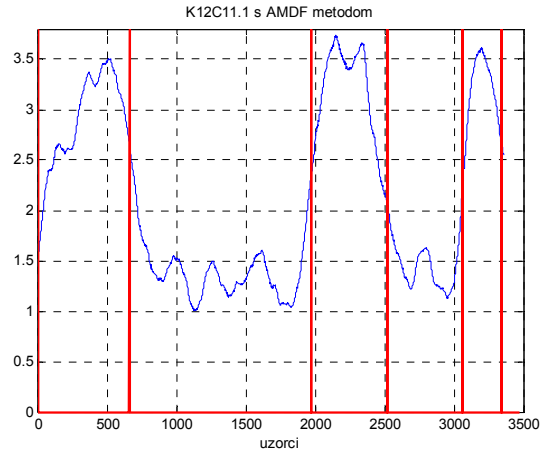
Slika 28: Raspored regija gena F56F11.4 dobiven nadograđenim detektorom s kvocijentom ovojnica

2. Gen K12C11.1

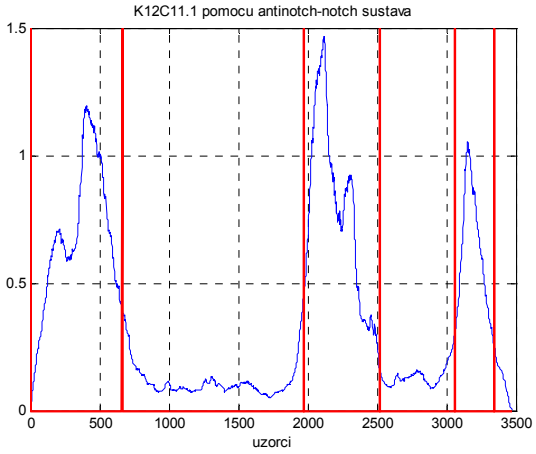
Gen K12C11.1 ima jednake osobine kao i gore opisani F56F11.4 te su iz tog razloga i očekivanja i rezultati jednaki. (slike 29 do 32)



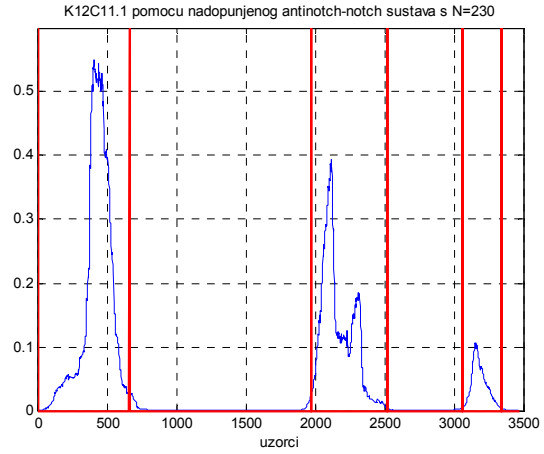
Slika 29: Raspored regija gena K12C11.1 dobiven antinotch filtrom



Slika 30: Raspored regija gena K12C11.1 dobiven AMDF metodom



Slika 31: Raspored regija gena F56F11.4 dobiven detektorom s kvocijentom ovojnica



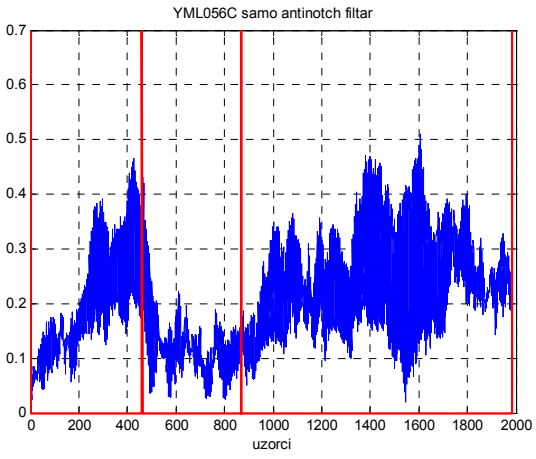
Slika 32: Raspored regija gena F56F11.4 dobiven nadograđenim detektorom s kvocijentom ovojnica

3. Gen YML056C

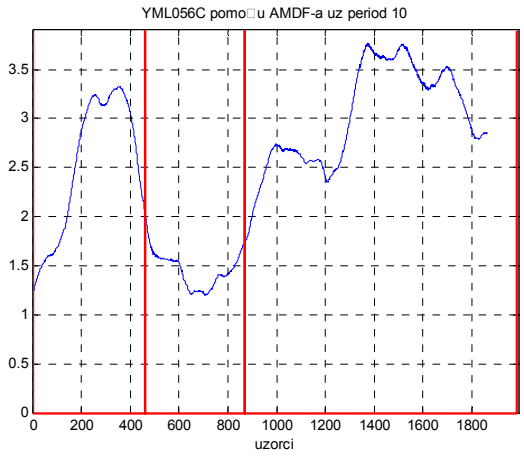
Gen YML056C pokazuje nepotpunu pravilnost. Sastoji se od 1983 nukleotida. Ima dva dugačka eksona između kojih se nalazi kraći intron koji nema svojstvo perioda 3. Odstupa od uvjeta za minimalnu i maksimalnu dužinu koje su postavljeni u početku da bi jednostavan antinotch filtar mogao odrediti kodirajuće regije, ali ipak zadržava dva glavna svojstva:

- period tri je tipičan samo za eksone
- eksoni su dovoljno dugi da bi se, zavisno o metodi, mogli detektirati

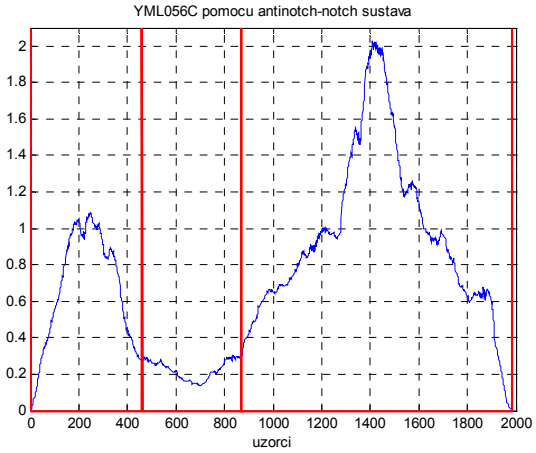
U ovom slučaju može se primijetiti da se i iz samog antinotch filtra naslućuje koje bi regije mogle pripadati eksonima, ali se zbog velike količine šuma ne može niti nagađati mogući početak eksona čak ni u okolini nekoliko stotina nukleotida. AMDF i detektor s kvocijentom ovojnica daju puno bolje rezultate. Jasno su uočljive kodirajuće i nekodirajuće regije te je moguće pretpostaviti intron-ekson granicu u okolini 50 nukleotida. Nadograđeni detektor s kvocijentom ovojnica daje daleko najbolje rezultate. Granice među područjima mogu se odrediti na razini najviše desetak nukleotida što je puno bolje od bilo kojeg drugog rezultata. (slike 33 do 36)



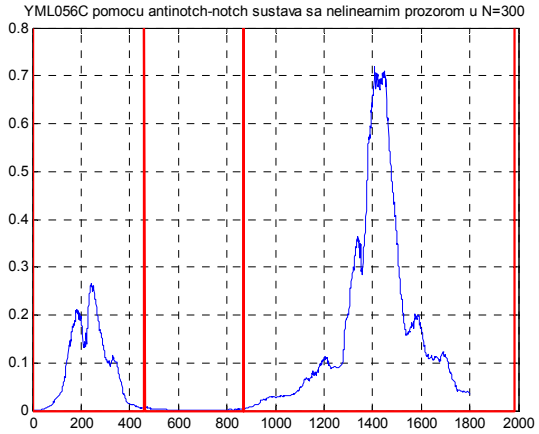
Slika 33: Raspored regija gena YML056C dobiven antinotch filtrom



Slika 34: Raspored regija gena YML056C dobiven AMDF metodom



Slika 35: Raspored regija gena YML056C dobiven detektorom s kvocijentom ovojnica



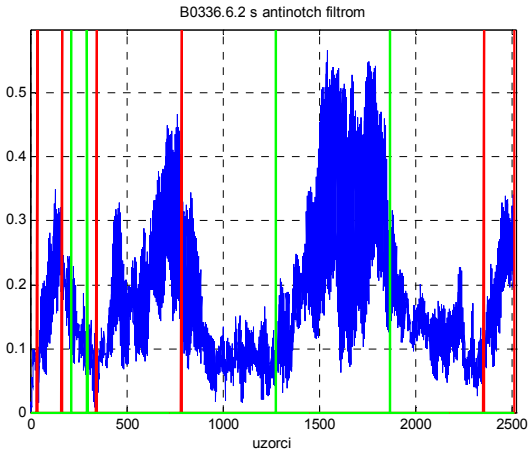
Slika 36: Raspored regija gena YML056C dobiven nadograđenim detektorom s kvocijentom ovojnica

4. Gen B0336.6.2

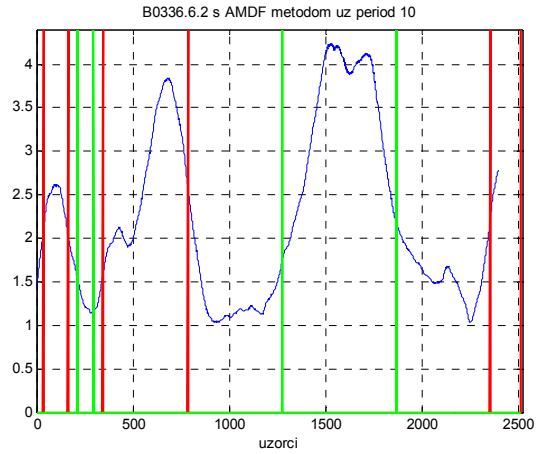
Gen B0336.6.2 se sastoji od 2512 nukleotida i sadrži pet eksona. Prva dva eksona su duljine stotinjak baza i razdvojeni su intronom iste duljine. Na prvi pogled mogli bi zaključiti da u ovom slučaju osnovni antinotch filtar i na njemu bazirana funkcija srednje razlike magnituda daju preciznije rezultate nego sustav s kvocijentom ovojnica, no to nije slučaj. Te dvije metode daju povišenje amplitude na mjestu prva dva eksona jer se filtar, koji doista naiđe na određenu periodičnost sa tri unutar prvog eksona, ne stigne istitrati do dolaska drugog eksona. Zato i pozicija tih eksona, koju je moguće procijeniti gledajući povišenje amplitude, ne odgovara njihovom pravom položaju unutar gena.

Sve četiri metode dobro detektiraju regije koje odgovaraju preostalim trima eksonima zato što oni posjeduju ranije navedena poželjna svojstva.

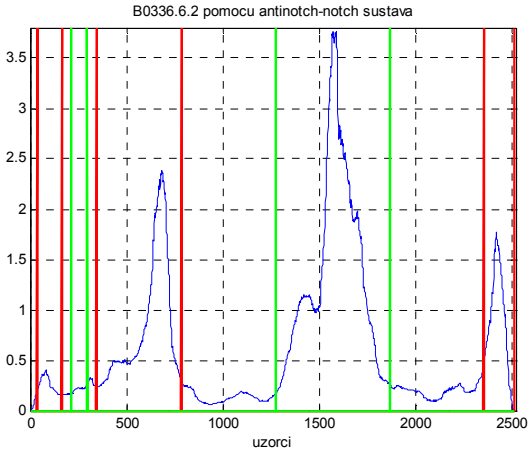
(slike 37 do 40)



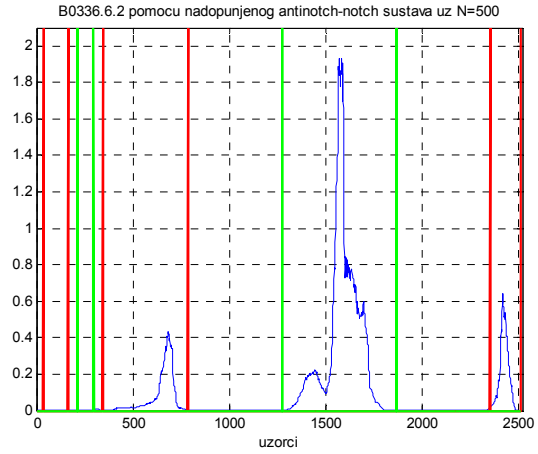
Slika 37: Raspored regija gena B0336.6.2 dobiven antinotch filtrom



Slika 38: Raspored regija gena B0336.6.2 dobiven AMDF metodom



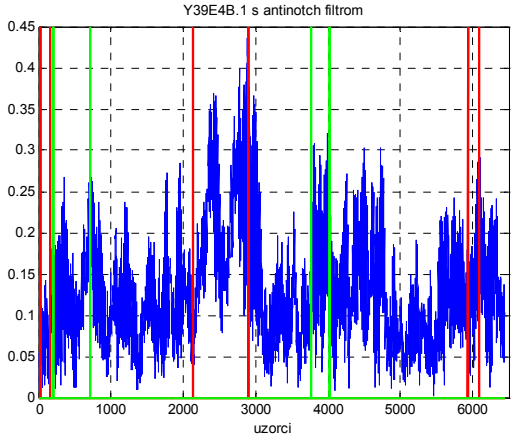
Slika 39: Raspored regija gena B0336.6.2 dobiven detektorom s kvocijentom ovojnica



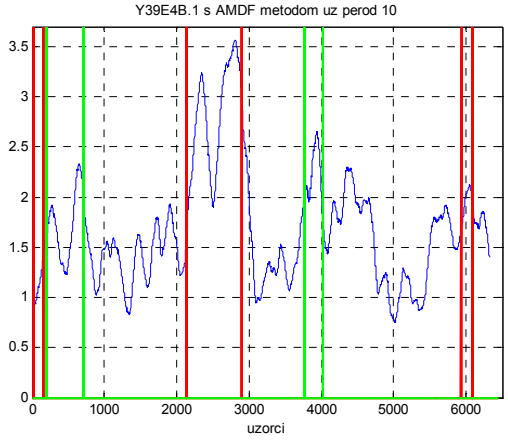
Slika 40: Raspored regija gena B0336.6.2 dobiven nadograđenim detektorom s kvocijentom ovojnica

5. Gen Y39E4B.1

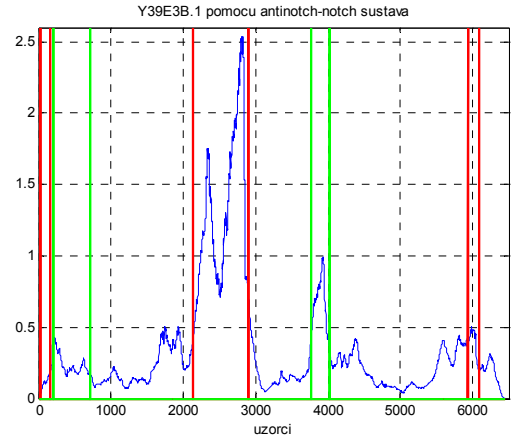
Gen Y39E4B.1 se sastoji od 6441nukleotida i 5 eksona. Osim prvog eksona, svi ostali su dovoljno dugi da bi period tri bio uočljiv, ali kraći od introna između njih. Ipak, kao što se može uočiti iz slika 41 do 44, detekcija eksona nije dobra. Razlog leži u činjenici da period tri na koji se oslanjamo nije dominantno izražen u eksonima niti je potpuno nepostojeći unutar nekodirajućih područja. Iz ovog se primjera može najbolje uočiti koliki utjecaj ima preciznost filtra u obje domene. Prve dvije metode ne uspijevaju napraviti razliku među područjima, dok druge dvije metode pronalaze 3 od 5 područja koja odgovaraju eksonima.



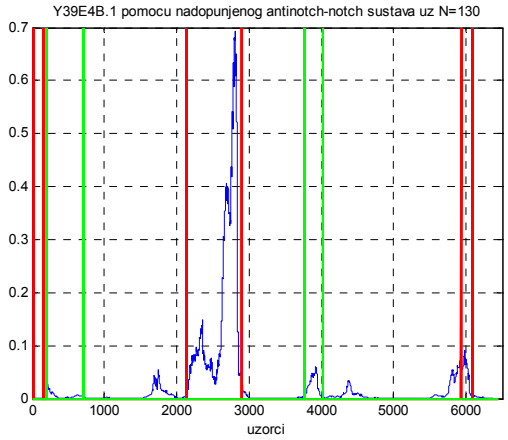
Slika 41: Raspored regija gena Y39E4B.1 dobiven antinotch filtrom



Slika 42: Raspored regija gena Y39E4B.1 dobiven AMDF metodom



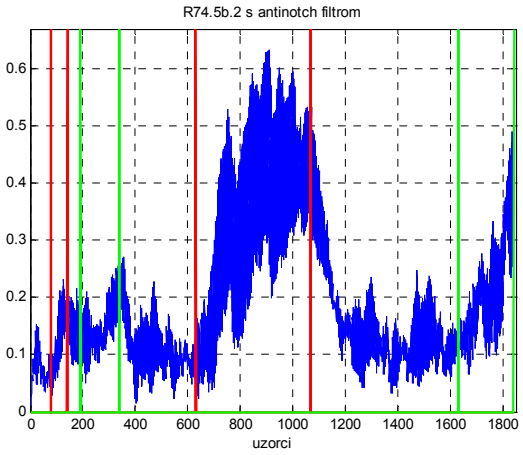
Slika 43: Raspored regija gena Y39E4B.1 dobiven detektorom s kvocijentom ovojnica



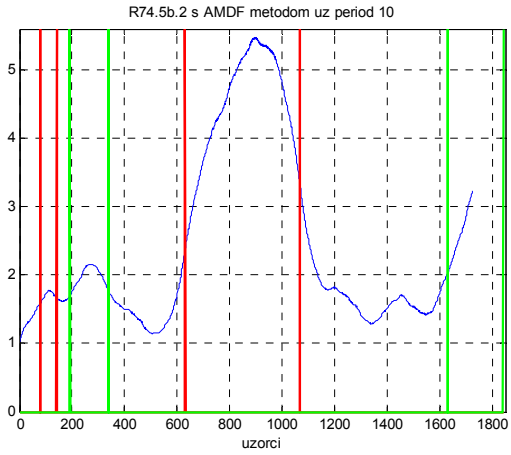
Slika 44: Raspored regija gena Y39E4B.1 dobiven nadograđenim detektorom s kvocijentom ovojnica

6. Gen R74.5b.2

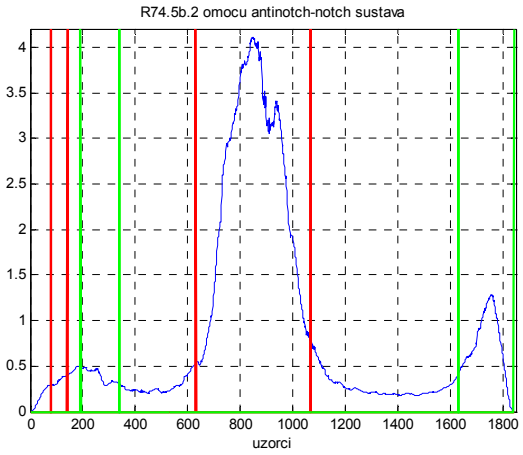
Gen R74.5b.2 sadrži 1842 nukleotida koji čine četiri eksona. Prvi ekson ne pokazuje periodičnost sa tri i niti jedna metoda ga ne uspijeva detektirati. Drugi ekson se sastoji od 147 baza koje pokazuju djelomičnu periodičnost sa tri i reakcija svih detektora nije izostala, ali nije niti naglašena. Za razliku od prva dva eksona, druga dva su dulji i period tri je u njima izraženiji, te se kod svih metoda, uz postavljanje dobro određenog praga, područje koje odgovara tim eksonima može predvidjeti u okolini 50-tak nukleotida. (slike 45 do 48)



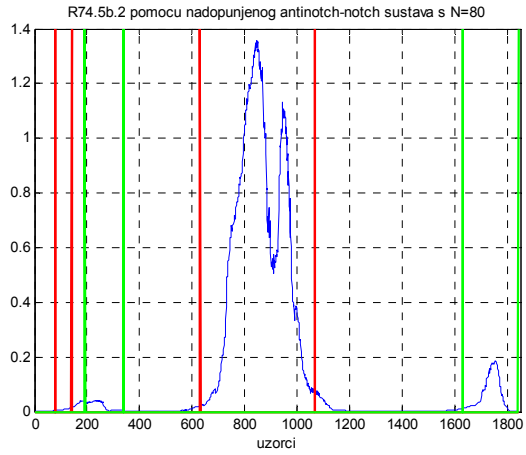
Slika 45: Raspored regija gena R74.5b.2 dobiven antinotch filtrom



Slika 46: Raspored regija gena R74.5b.2 dobiven AMDF metodom



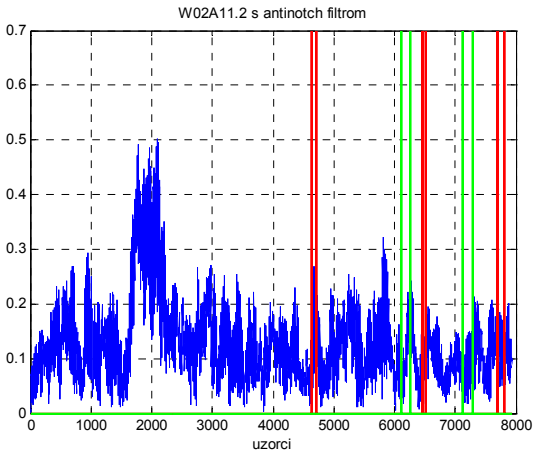
Slika 47: Raspored regija gena R74.5b.2 dobiven detektorom s kvocijentom ovojnica



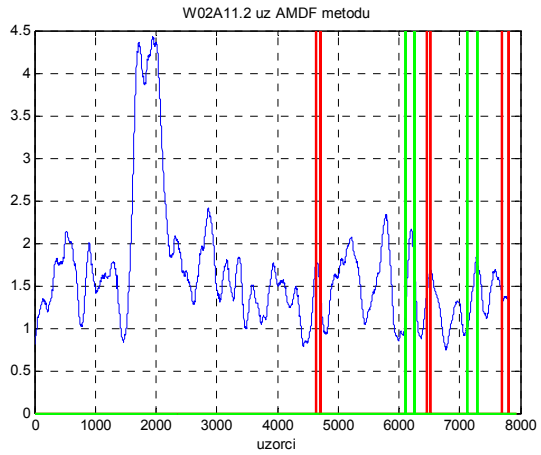
Slika 48: Raspored regija gena R74.5b.2 dobiven nadograđenim detektorom s kvocijentom ovojnica

7. Gen W02A11.2

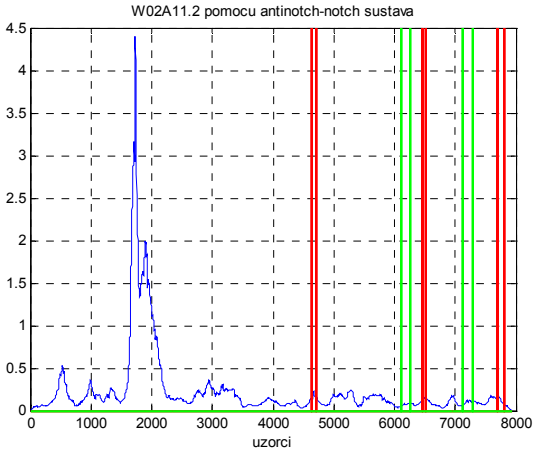
Gen W02A11.2 je primjer gena koji je po rasporedu i duljini kodirajućih i nekodirajućih područja idealan za testiranje navedenih detekcijskih metoda. Međutim, sve su metode su pri pokušaju pozicioniranja eksona u potpunosti podbacile. Također sve metode su detektirale ekson od 1600 do 2200 nukleotida, a to je nekodirajuće područje. Pet eksona duljine 100-tinjak baza nalaze se od 4641. nukleotida do 7809. nukleotida. Manjak reakcije detektora na eksone, a istovremena aktivnost u nekodirajućem području može se objasniti specifičnosti ovog uzoraka da kodirajuća područja ne pokazuju svojstvo periodičnost sa tri dok je isto svojstvo prisutno u nekodirajućem području. (slike 49 do 52)



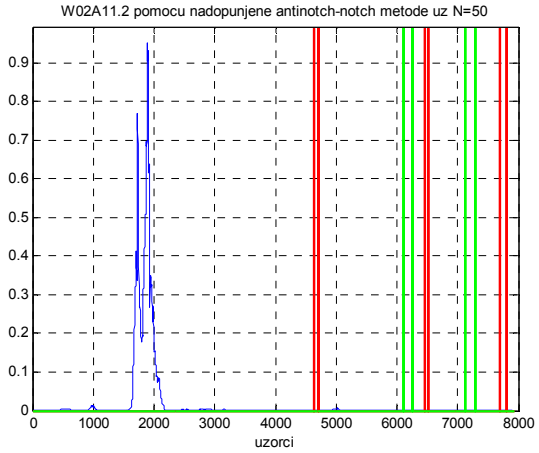
Slika 49: Raspored regija gena W02A11.2 dobiven antinotch filtrom



Slika 50: Raspored regija gena W02A11.2 dobiven AMDF metodom



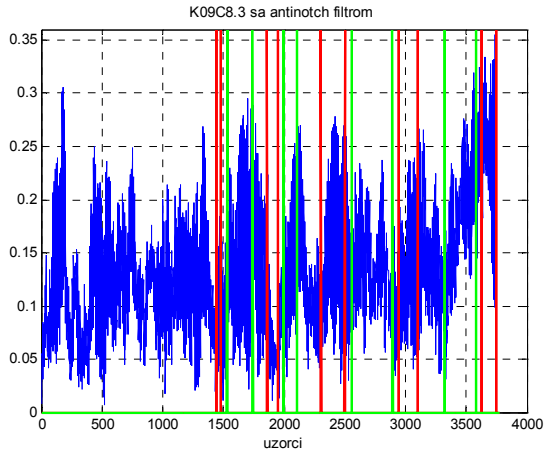
Slika 51: Raspored regija gena W02A11.2 dobiven detektorom s kvocijentom ovojnica



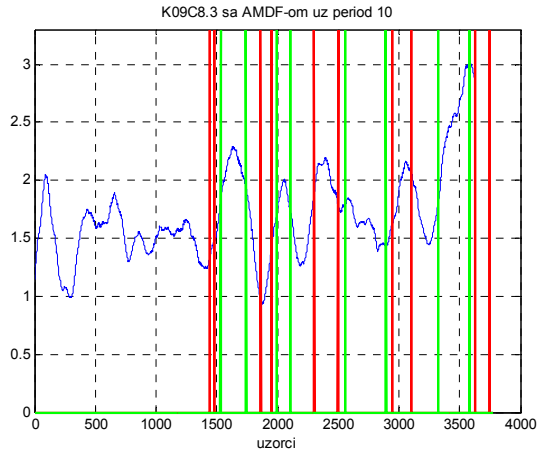
Slika 52: Raspored regija gena W02A11.2 dobiven nadograđenim detektorom s kvocijentom ovojnica

8. Gen K09C8.3

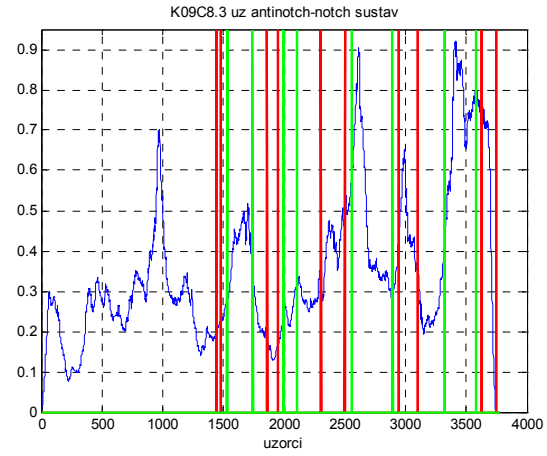
Gen K09C8.3 je primjer gena koji ne slijedi pravila koja su navedena ranije. Sastoji se od 3774 nukleotida unutar kojih se nalaze 9 eksona. Razlika u duljini eksona i nekodirajućih regija ne postoji, eksoni su različite duljine te se nalaze vrlo blizu jedan drugome. U ovakvom uzorku jako je teško detektirati pojedina područja, pogotovo ako i introni pokazuju periodičnost sa tri. Sam antinotch filtar ne uspijeva razlučiti regije uopće. Funkcija srednje razlike magnituda daje malo bolje rezultate, ali niti tu ne možemo sa sigurnošću utvrditi regije. Detektor s kvocijentom ovojnica bolje pronalazi i jače reagira na područja koja pokazuju periodičnost sa 3 od prethodnih metoda, ali detektira i pogrešno točna područja. Ne može se zbog toga reći da sustav radi krivo, već da za ovaj gen period tri postoji i u nekodirajućoj regiji. Detektor s kvocijentom ovojnica s nelinearnim prozorom učinkovitije od ostalih potiskuje pogrešno točno detektirani ekson od 910-tog do 1000-tog uzorka, ali nažalost isto tako umanjuje i eksone koje je prvobitno dobro detektirao. Za slučaj K09C8.3 mora se priznati da su sva četiri dosad obrađena modela podbacila. Niti jedan ne omogućava procjenu položaj eksona. (slike 53 do 56)



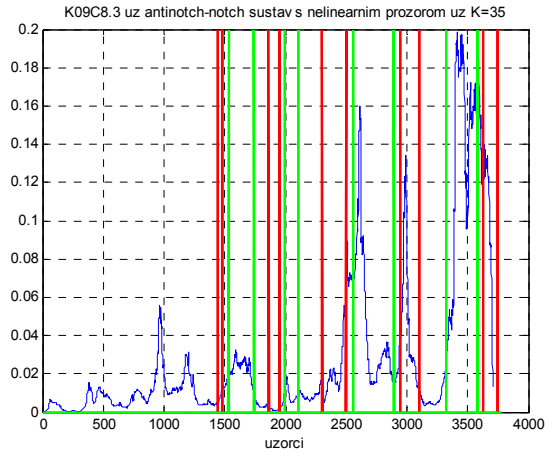
Slika 53: Raspored regija gena K09C8.3 dobiven antinotch filtrom



Slika 54: Raspored regija gena K09C8.3 dobiven AMDF metodom



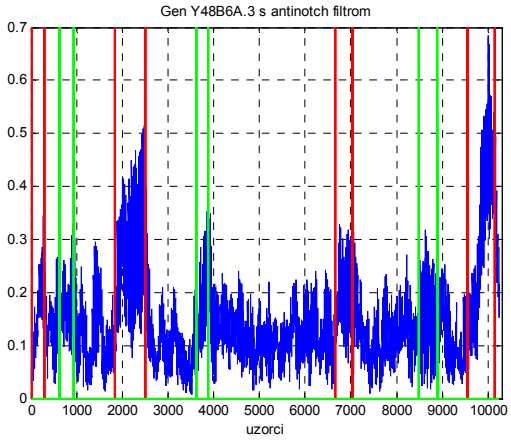
Slika 55: Raspored regija gena K09C8.3 dobiven detektorom s kvocijentom ovojnica



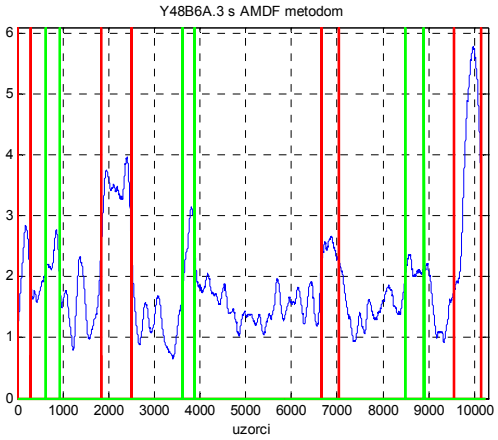
Slika 56: Raspored regija gena K09C8.3 dobiven nadograđenim detektorom s kvocijentom ovojnica

9. Gen Y48B6A.3

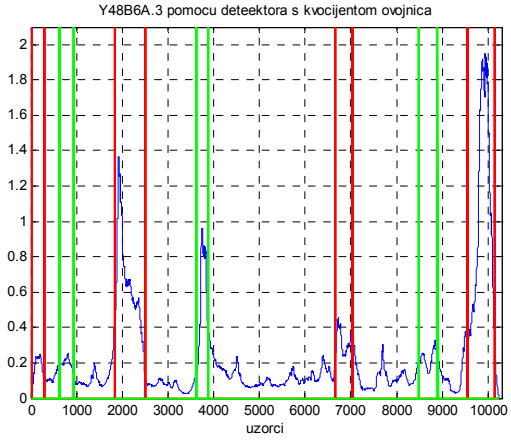
Gen Y48B6A.3 sadrži 10251 nukleotid i ima 7 eksona. Na slikama 57do 60 prikazani su rasporedi regija za sve četiri opisane metode. Sve četiri metode dobro su izolirale četiri od sedam eksona, međutim samo nadograđeni detektor s kvocijentom ovojnica (slika 60) uspijeva napraviti razliku između nekodirajućih regija i svih sedam eksona.



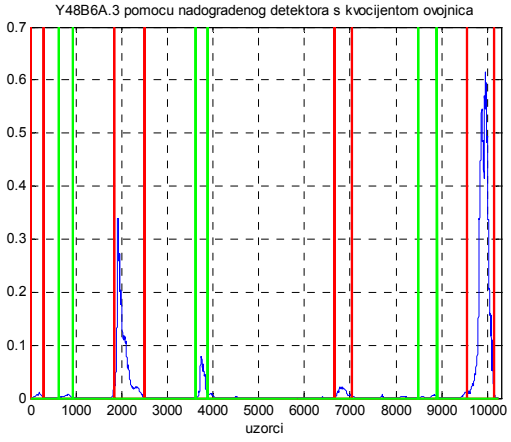
Slika 57: Raspored regija gena Y48B6A.3 dobiven antinotch filtrom



Slika 58: Raspored regija gena Y48B6A.3 dobiven AMDF metodom



Slika 59: Raspored regija gena Y48B6A.3 dobiven detektorom s kvocijentom ovojnica



Slika 60: Raspored regija gena Y48B6A.3 dobiven nadograđenim detektorom s kvocijentom ovojnica

5. Pseudogen

U citoplazmi stanice, unutar ribosoma, vrši se proces translacije kodona sastavljenih od tri nukleotida u aminokiselinu. Niz tako dobivenih aminokiselina tvori protein. Tablica 1 prikazuje način pridjeljivanja aminokiselina kodonima.

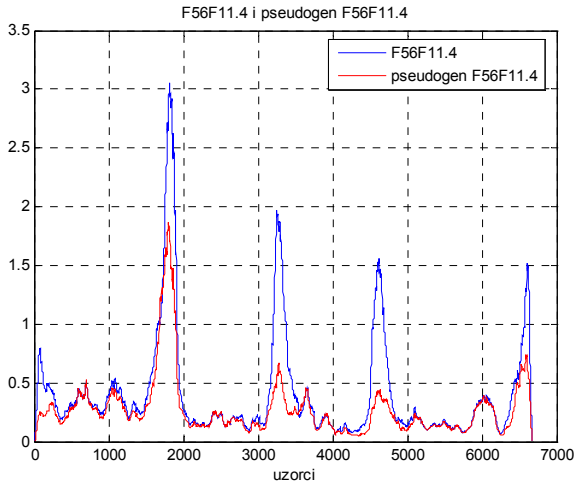
AMINOKISELINE	TRIPLETI NUKLEOTIDA-KODONI
Lys	AAA AAG
Asn	AAT AAC
Arg	AGA AGG CGA CGG CGT CGC
Ser	AGT AGC TCA TCG TCT TCC
Ile	ATA ATT ATC
START/Met	ATG
Thr	ACA ACG ACT ACC
Glu	GAA GAG
Asp	GAT GAC
Gly	GGA GGG GGT GGC
Val	GTA GTG GTT GTC
Ala	GCA GCG GCT GCC
Tyr	TAT TAC
Cys	TGT TGC
Leu	TTA TTG CTA CTG CTT CTC
Phe	TTT TTC
Gln	CAA CAG
His	CAT CAC
Pro	CCA CCG CCT CCC
Trp	TGG
STOP	TAA TAG TGA

Tablica 1: Aminokiseline i odgovarajući kodoni [3]

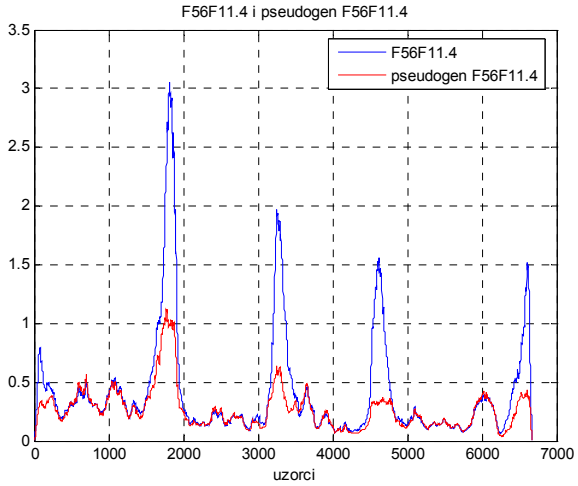
Ako se zanemare stop kodone, ostaje 61 kodon koji kodira samo 20 aminokiselina. Očito je da više kodona odgovara jednoj aminokiselini. Nije poznato po kojemu ključu se odlučuje koji će triplet nukleotida sudjelovati u sintezi proteina, ali se pretpostavlja da je razlog zašto nalazimo period tri unutar eksona, upravo u preferiranju nekih kodona u odnosu na druge. [4]

Pseudogen je umjetno stvoren niz nukleotida koji od pravog gena zadržava nekodirajuća područja, a na temelju poznatog niza aminokiselina rekonstruira eksone. Pritom se svakoj aminokiselini odgovarajući kodoni dodjeljuju po uniformnoj razdiobi, čime se eliminira pristranost prema kodonima. Dobiveni pseudogen sugerira da gornje objašnjenje nastanka perioda tri nije kompletno.
Takav umjetno stvoreni gen pokazuje periodičnost sa tri, ako i originalni gen, čija smo nekodirajuća područja iskoristili, također pokazuje periodičnost.
Za svaki gen vrijedi da je zbroj nukleotida u svim eksonima višekratnik broja tri. To je razumljivo jer samo eksoni kodiraju proteine, a svaka aminokiselina u proteinu odgovara tripletu baza iz eksona.
Problem kod kreiranja pseudogena se javlja jer unutar jednog gena postoji više eksona, ali broj baza po eksonu ne mora biti višekratnik broja tri.
Pri kreiranju pseudogena kreće se od niza aminokiselina iz kojega se ne vidi koliko je eksona sudjelovalo u njegovom stvaranju. Paralelno promatrajući poznatu gensku sekvencu mora se odrediti od koliko se eksona ona sastoji te baze unutar eksona premjestiti tako da ukupan broj nukleotida u eksonu bude djeljiv sa tri. Time se zapravo mijenja struktura gena, ali se ne utječe na finalni protein koji se iz tog gena proizvede.

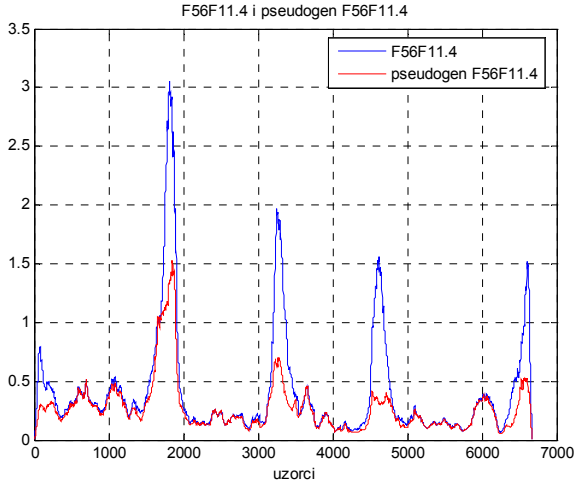
Na slikama 61 do 63 prikazani su, redom, originalni gen F56F11.4 i njegove tri slučajno dobivene varijante. Analiza regija dobivena je uz korištenje detektora s kvocijentom ovojnica.



Slika 61: Gen F56F11.4 i prva varijanta pseudogena F56F11.4



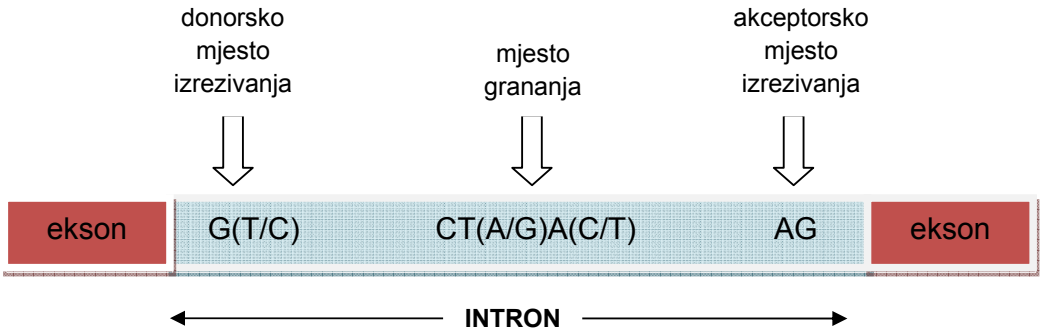
Slika 62: Gen F56F11.4 i druga varijanta pseudogena F56F11.4



Slika 63: Gen F56F11.4 i treća varijanta pseudogena F56F11.4

6. Detektor temeljen na građi introna

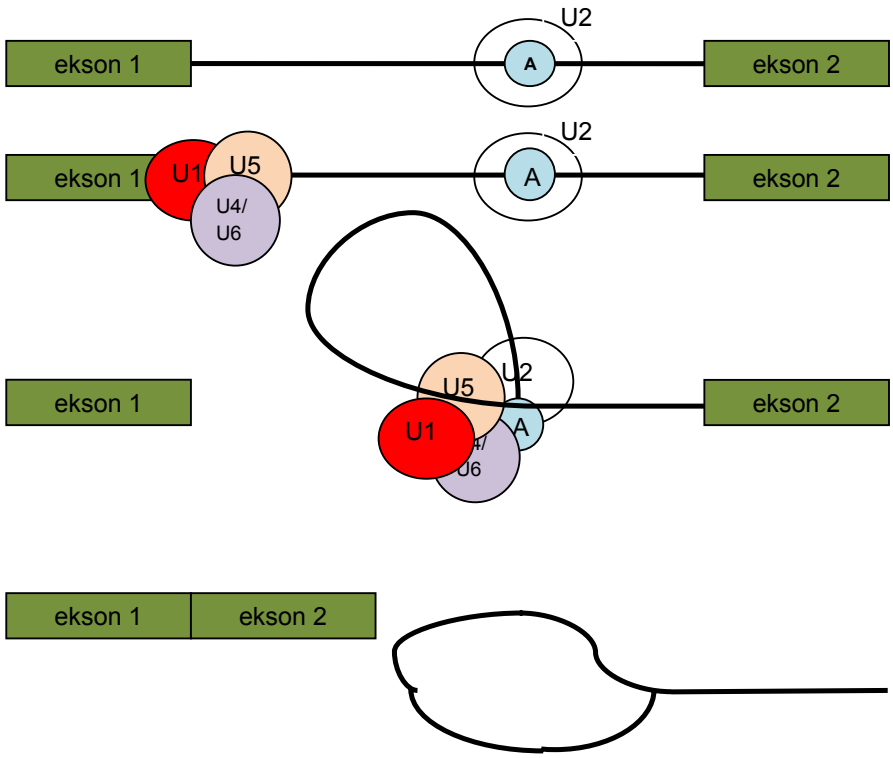
Broj introna i njihova duljina razlikuju se od gena do gena. Pokazuje se da što je viši stupanj razvoja organizma to je broj introna veći. Međutim, svi introni posjeduju slijedeće jednake značajke. Početak introna naziva se donorsko mjesto izrezivanja (eng. donor splice site), a kraj svakog introna zove se akceptorsko mjesto izrezivanja (eng. acceptor splice site). Donorsko mjesto izrezivanja uvijek počinje sa G, a u 99% slučajeva ga slijedi T. U preostalih 1% slučajeva umjesto T dolazi C. Akceptorsko mjesto izrezivanja je uvijek obilježeno sa AG.[5] Između te dvije točke nalazi se mjesto grananja (eng. branch site), na koje se spaja snRNP koji pokreče izrezivanje. Mjesto grananja je oblika CT(A ili G)A(C ili T).



Slika 64: Prikaz specifičnosti eksona i granica između eksona i introna

Te tri specifične točke su, zajedno sa malim jezgrenim ribonukleoproteinima, odgovorne za proces izrezivanja koji je obavezan za sintezu proteina. Na slici 65 je prikazan pojednostavljen proces izrezivanja. Osnovna ideja je da se intron izdvoji iz lanca na donorskom i na akceptorskom mjestu izrezivanja, te se zatim slobodni eksoni povežu tvoreći zrelu glasničku RNK. Ovaj proces ubrzava i nadzire makromolekula koja se zove spliceosome koja se sastoji od pet malih jezgrenih ribonukleinskih proteina (snRNP-a) U1, U2, U4, U5 i U6 te od brojnih drugih proteina. Signal za izrezivanje daje U2 koja se spoji na mjesto grananja i tada se prvo odvaja donorsko mjesto izrezivanja te se preko mjesta grananja spaja sa akceptorskim

mjestom izrezivanja. Time je intron potpuno spojen „sam na sebe“, a eksoni se povežu tako da se lijevi kraj jednog eksona poveže na desni kraj drugoga.[10]

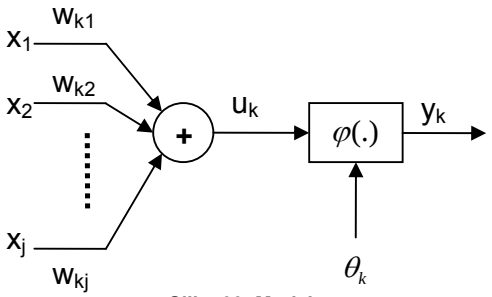


Slika 65: Pojednostavljeni prikaz procesa izrezivanja

Postavlja se pitanje ako su poznati početci i krajevi introna zašto se jednostavno njihovim određivanje ne odredi položaj eksona. Problem je u tome što je prosječan gen sastavljen od 1500 do 6000 nukleotida, ali ima znatno dužih te je broj mogućih kombinacija koje odgovaraju intronu na temelju njegove tri karakteristične točke (akceptorsko i donorsko mjesto izrezivanja, te mjesto grananja) iznimno velik. Samo za gen F56F11.4 duljine 6677 nukleotida dobije se 190218 mogućih kombinacija.

6.1. Uvod u neuronske mreže

U ovom kontekstu neuronske mreže podrazumijevaju umjetne neuronske mreže. Njihov cilj je što vjernije imitirati rad biološke neuronske mreže. Neuronske mreže se implementiraju na digitalnim računalima opće namjene ili pomoću specijaliziranih sklopova. Osnovna građevna jedinica neuronske mreže je neuron. Model neurona prikazan je na slici 66.



Slika 66: Model neurona

Skup ulaza $x_{1.....}x_j$, od kojih svaki ima svoju težinu w_{jk} ulaze u sumator k-tog neurona. Sumator zbraja otežane ulaze i na taj načina računa njihovu linearnu kombinaciju $u_k = \sum_{i=1}^{j} w_{ki} x_i$.

Linearna kombinacija ulaza se uvodi u nelinearnu funkciju φ koja, ovisno o zadanom pragu θ_k, ograničava izlaz neurona na interval [0,1], tj. vrijedi da je $y_k = \varphi(u_k - \theta_k)$.

Neuronska mreža je slična mozgu jer znanje stiče kroz proces učenja i koristi veze među neuronima za pohranu znanja.

Za razvoj detektora upotrebljena je jednoslojna neuronska mreža s povratnom propagacijom pogreške. Kao aktivacijska funkcija u ulaznom sloju neurona se koristi tangencijalno sigmoidna aktivacijska funkcija (tansig), a u neuronu izlaznog sloja koristi se funkcija purelin.

6.2. Detekcija intron-ekson i ekson- intron granica pomoću neuronske mreže

Te-Ming Chen et al. [11] su dokazali da su baze raspoređene u okolini donorskog i akceptorskog mjesta međusobno zavisne i to 8 nukleotida lijevo i desno od G(T/C) te 8 nukleotida desno i 26 nukleotida lijevo od AG. Njihova zavisnost proizlazi iz samog procesa izrezivanja. Uklanjanje introna vrši se posredstvom ribonukleoproteina iz spliceosoma. Oni, kako sam naziv kaže, u svom sastavu sadrže nukleotide koje se u okolini mjesta izrezivanja spajaju sa nukleotidima koji čine gen nepoštivajući načine uparivanja koji vrijede za tvorbu spiralne strukture DNK molekule. To znači da se sve se baze mogu međusobno uparivati tvoreći veze različite jačine.

Ta povezanost nije očita, pa se nameće ideja upotrebe neuronske mreže koja bi na temelju učenja na primjerima točnih i netočnih mjesta izrezivanja vraćala 1 ukoliko je nepoznato mjesto izrezivanja točno, a 0 ako nije. Svi primjeri gena i odgovarajućih mjesta izrezivanja kojima je stvorena baza za treniranje, validaciju i testiranje uzeti su sa sljedećih internetskih baza www.wormbase.org i www.genedb.org. Baza je stvorena od 205 različitih gena.

Umjesto karaktera a, c, t i g kojima je opisan gen, svakom je karakteru pridružen jedan cijeli broj prije ulaska u neuronsku mrežu, redom 1, 2, 3 i 4.

Učinkovitost mreže ocjenjuje se iz postotka krivih mjesta izrezivanja koja su proglašena točnima (eng. false positive FP) te postotka ispravnih mjesta izrezivanja koja su proglašena netočnim (eng. false negative FN). Postotak FN i FP se izračunava na način :

$$FN[\%] = \frac{broj\_FN}{broj\_FN + broj\_TP}, \quad (5)$$

$$FP[\%] = \frac{broj\_FP}{broj\_FP + broj\_TN}, \quad (6)$$

gdje je TP broj ispravno detektiranih točnih mjesta izrezivanja, a TN broj ispravno detektiranih netočnih mjesta izrezivanja.

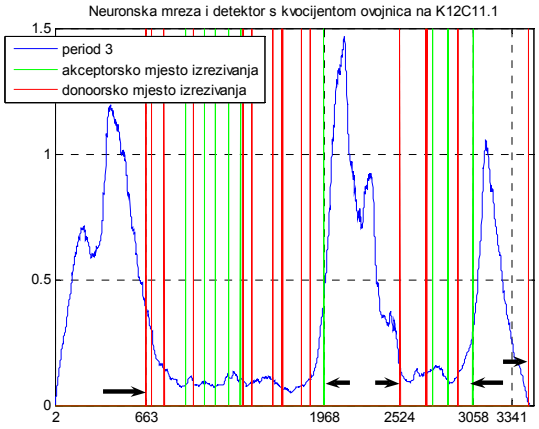
6.3. Opis rezultata

Skupovi značajki se razlikuju za akceptorska i za donorska mjesta izrezivanja.

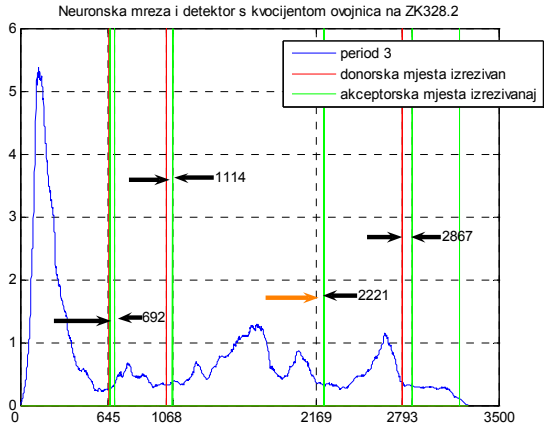
Bazu donorskih mjesta izrezivanja čini 1462 točnih mjesta izrezivanja, 36476 netočnih mjesta izrezivanja. Broj netočnih je veći od broja točnih, jer to odgovara situaciji u prirodi. Međutim, takva velika razlika između broja točnih i netočnih sprječava neuronsku mrežu da pravilno nauči. Punjenje baze s još više uzoraka, rezultiralo bi samo daljnjim neproporcionalnim rastom razlike između točnih i netočnih. Zato, iako se time zaobilazi prirodni raspored, bazu treba umjetno nadopuniti točnim uzorcima i to ponavljanjem onih koji se već u bazi nalaze. Treba ih nadodati toliko da omjer broja točnih i netočnih mjesta izrezivanje bude otprilike 1:1. To za gornji konkretan slučaj znači da će se baza točnih mjesta izrezivanja ponoviti 25 puta unutar cijelog skupa donorskih mjesta izrezivanja koji služi za treniranje, validaciju i testiranje mreže. Tako dobivena baza donorskih mjesta izrezivanja sadrži 73026 uzoraka dimenzije 16. Upotrijebljena neuronska mreža za određivanje donorskih mjesta izrezivanja se sastoji od 16 neurona u ulaznom sloju i 1 neuron u izlaznom sloju. Prag osjetljivosti postavljen je na vrijednost 0.5. Dobivena vrijednost FN je 4.8 %, a srednja vrijednost FP je 10.2 %. Uz ovako odabran prag postotak FP-a je veći od postotka FN-a, ali to je u redu je jer se pretpostavlja da će se slučaj lažno točno proglašenog mjesta izrezivanja moći lakše riješiti nekom naknadnom obradom.

Bazu akceptorskih mjesta izrezivanja čini 1476 točnih mjesta izrezivanja, 46589 netočnih mjesta izrezivanja. Broj točnih mjesta treba ponoviti 32 puta da bi broj točnih i netočnih akceptorskih mjesta izrezivanja postao ujednačen. Time se dobije ukupna baza od 95927 uzoraka dimenzije 34. Neuronsku mrežu čine 17 neurona u ulaznom sloju i 1 neuron u izlaznom sloju. Prag osjetljivosti postavljen je na vrijednost 0.5. Uz tako odabrani prag srednja vrijednost FN na testiranim uzorcima je 3.6%, a srednja vrijednost FP je 6.9 %.

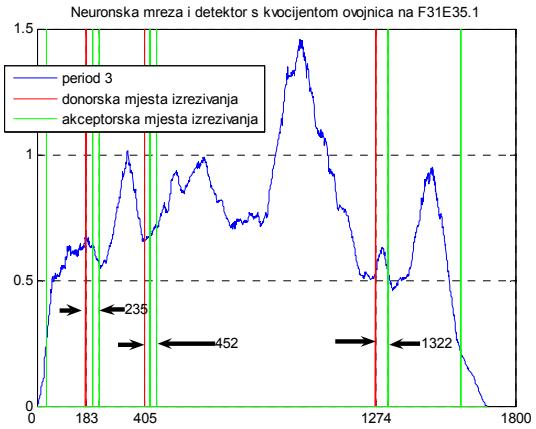
Na donjim slikama 67 do 70 prikazani su zajednički rezultati detektora s kvocijentom ovojnica i neuronske mreže. Iako niti jedan od njih samostalno ne određuje precizno regije, njihovi izlazi se nadopunjuju i može se dobiti dobra predodžba o poziciji eksona.



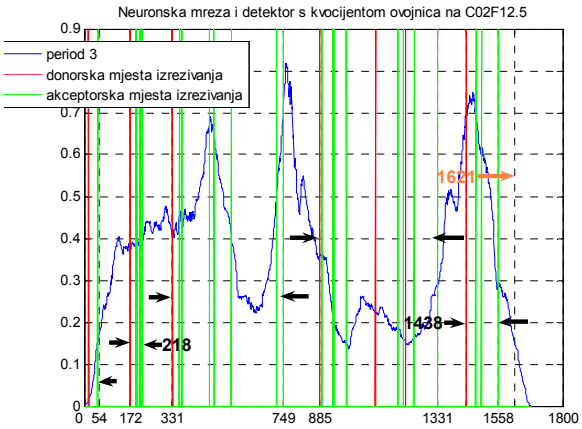
Slika 67: Neuronska mreža i detektor s kvocijentom ovojnica primijenjeni na gen K12C11.1



Slika 68: Neuronska mreža i detektor s kvocijentom ovojnica primijenjeni na gen ZK328.2



Slika 69: Neuronska mreža i detektor s kvocijentom ovojnica primijenjeni na gen F31E35.1



Slika 70: Neuronska mreža i detektor s kvocijentom ovojnica primijenjeni na gen C02F12.5

6. Zaključak

Opisane metode su uspješne u svojim zadanim okvirima i uz realna ograničenja koja su im postavljena. Nije realno očekivati da će besprijekorno raditi za svaki gen niti da će rezultati biti nedvojbeno točni. Njihov cilj je bio dati granice unutar kojih će se vjerojatno nalazi kodirajuće područje, bilo to ispitivanjem samih eksona ili određivanjem mjesta izrezivanja. Bitno je primijetiti da metode bazirane na eksonima i one bazirane na svojstvima nekodirajućih područja nisu međusobno isključive, već da bi se naprotiv mogle uspješno spojiti u sustav koji bi tako iskoristio njihove prednosti i vrlo vjerojatno davao bolje rezultate.

Ovako dizajnirana neuronska mreža rezultat je potpuno eksperimentalnog pristupa pri određivanju broja uzoraka koji čine bazu za treniranje, validaciju i testiranje. Povečanjem broja različitih uzoraka u bazi, rezultati bi se trebali još poboljšavati. Također trebalo bi se posvetiti smanjivanju dimenzionalnosti vektora značajki, što u ovom radu nije obrađeno, a također bi moglo donijeti određeni napredak, barem što se tiče smanjivanja vremena potrebnog za obradu rezultata.

Detektor temeljen na pronalaženju regija s periodom 3 ima realno ograničenje ako testirani gen ne pokazuje takvo svojstvo i tu nema pomoći. Međutim, stvar na kojoj treba raditi je slučaj u kojemu periodičnost sa tri pokazuju i nekodirajuće i kodirajuće regije. Taj problem bi se mogao riješiti naknadnom obradom rezultata. Da je to ispravan način može se zaključiti iz slučajeva gena analiziranih nadograđenim detektorom s kvocijentom ovojnica. Ugradnja nelinearnog prozora nikako nije idealna metoda naknadne obrade rezultata jer podrazumijeva predznanje o veličini eksona kojega pokušavamo otkriti. Ona služi za ilustraciju da je moguće na razini binarnih sekvenci baza eliminirati periodičnost koja nastaje zbog introna i time riješiti problem periodičnost u nekodirajućim regijama.

Upravo na kombiniranju svih dostupnih znanja se u današnje doba i baziraju alati za analiziranje i predikciju gena, poput GENSCAN-a. [13] Stvaranje takvog sustava uz poboljšanja i razvoj gore obrađenih metoda, trebalo bi biti osnova daljnjeg istraživanja.

7. Literatura

[1] P.P.Vaidyanathan,Byung-Jun Yoon, *The role of signal-processing concepts in genomic and proteomics,* Journal of the Franklin Institute, 2004, vol.341, pp.111-135

[2] P.P.Vaidyanathan, *A Signal Processor's Tour*, www.systems.caltech.edu/EE/Groups/dsp/ppv/papers/CASGeneGalley.pdf

[3] Dimitris Anastassiou, *Genomic Signal Processing*, IEEE Signal Processing Magazine, 2001., vol. 18, pp. 8-20

[4] Eliathamby Ambikairajah, Julien Epps, Mahmood Akhtar*, Gene and exon prediction using time dopmain algorithms*, IEEE Signal Processing and Its Applications, 2005, vol.1, pp. 199-202

[5] Yuliya Sarkisyan, *Multiple Sequence Alignments*, http://ai.stanford.edu/~serafim/CS262_2007/notes/lecture15.pdf

[6] Yuan Xin Tian et al., *Fourier Power Spectrum Analysis of Exons for the Perod-3 Behavior*, Chinese Chemical Letters, 2005, vol. 16, pp. 939-942

[7] Trevor W. Fox, Alex Carreira, *A Digital Signal Processing Method for Gene Prediction with Improved Noise Suppression*, EURASIP Journal on Applied Signal Processing , vol.2004, pp.108-114

[8] Xabi Abad et al, *Requirements for gene splicing mediated by U1 snRNA binding to a target sequence*, Nucleic Acids Research, 2008, vol. 36, pp. 2338-2352

[9] P.P. Vaidyanathan, Byung-Jun Yoon, *Gene and Exon Prediction using allpass-based Filters,*http://www.systems.caltech.edu/dsp/students/bjyoon/conf/gensips_2002.pdf

[10] Joanne C. McGrail, Raymond T. O'Keefe, *The U1,U2 and U5 snRNAs crosslink to the 5'exon during yeast pre-mRNA splicing,*Nucleic Acid Research, 2007, vol.36, pp. 814-825

[11] Te-Ming Chen, Chung-Chin Lu, Wen-Hsiung Li, *Prediction of splice siteswith dependency graphs and their expanded bayesian networks,*Bioinformatics, 2005, vol.21, pp.471-482

[12] Sven Lončarić, Predavanja iz kolegija Neuronske mreže

[13] Chris Burge, Samuel Karlin*, Prediction of Complete Gene Structures in Human Genomic DNA*, Journal of Molecular Biology, 1997, vol.268, pp. 78-94

[14] Christopher B. Burge, Samuel Karlin, *Finding the genes in genomic DNA*, Current Opinion in Structural Biology 1998, vol.8, pp.346-354