SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 898

PREDVIĐANJE KODIRAJUĆIH REGIJA U GENOMU METODAMA DIGITALNE OBRADE SIGNALA

Vedrana Baličević

Zagreb, lipanj 2009.

Hvala prof. dr. sc. Damiru Seršiću na pomoći, idejama i optimizmu svih mjeseci trajanja projekta i završnog rada, svim asistentima i profesorima čija su vrata uvijek bila otvorena za sva pitanja, te ostatku tima s bio-info projekta na ugodnom društvu i moralnoj potpori tijekom svih uspona i padova u radu i raspoloženju.

Sadržaj

1. Uvod

Posljednjih godina područje bioinformatike, kao kombinacije informacijske tehnologije, matematičkih znanosti, biologije i biomedicine, postaje sve popularnije i raširenije u znanstvenim krugovima. Istraživanjima same biologije na molekularnoj razini do danas je dobivena velika količina podataka, koje je potrebno matematičkim i statističkim postupcima obraditi kako bi se iz njih dobile značajne informacije. Upravo zbog velikih količina podataka nužno je upotrijebiti računalne metode i algoritme za digitalnu obradu signala da bi se došlo do željenih rezultata.

Podaci su oblikovani tako da omogućuju informatičku obradu, primjerice geni, koji su građeni od 4 vrste nukleotida, oblikovani su kao sekvence 4 različita slova, a proteini, građeni od 21 vrste aminokiselina, oblikovani su kao sekvence od 21 različitog slova.

Geni su ti koji nose nasljednu informaciju svakog pojedinog organizma i predstavljaju upute za sintezu proteina odgovornih za odvijanje gotovo svih bioloških funkcija. No protein se generira na temelju samo određenih, kodirajućih dijelova gena, koji su biološkim metodama pronađeni, međutim nepoznate su značajke i pravilnosti po kojoj bi se oni mogli raspoznati ili predvidjeti iz same sekvence.

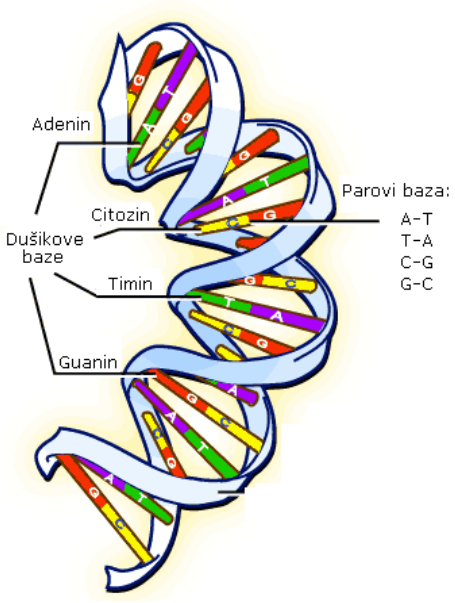
Do danas provedena brojna opsežna istraživanja u području bioinformatike rezultirala su manjom količinom korisnih informacija. Očekivano je da je nakon duge evolucije organizama pohrana genetičke informacije jednako tako napredovala i pronašla najbolje načine za njeno kodiranje, no potraga za svojstvima na temelju kojih će se moći razlikovati kodirajući eksoni od nekodirajućih introna u genu pokazuje da su biološki mehanizmi izrezivanja nekodirajućih dijelova ipak puno složeniji i da nije problem samo količina podataka već i načini prepoznavanja korisnog sadržaja iz njih i njihova interpretacija. S porastom korisnih informacija otkrivenih na bioinformatičkoj razini, rastu i mogućnosti razumijevanja i otkrivanja genetskih promjena i bolesti, a zahvaljujući pristupačnosti podataka pripremljenih za bioinformatička istraživanja, područje bioinformatike otvoreno je za svako istraživanje i sigurno će biti i dalje aktualno.

2. Biološke osnove

2.1 Osnove genetskog nasljeđivanja

Osnovna strukturna i funkcionalna jedinica svih poznatih organizama je stanica. Prema građi stanice, organizmi se dijele na jednostavnije prokariote i složenije eukariote. Prokarioti su organizmi koji u stanicama nemaju pravu jezgru (nemaju oblikovanu jezgrinu ovojnicu) niti organele, za razliku od eukariota u čijim se stanicama nalazi jezgra, koja je membranom odijeljena od citoplazme, te brojni drugi stanični organeli. Prokarioti su bakterije i modrozelene alge, a protisti, gljive, biljke, životinje i čovjek eukarioti.

Svaka vrsta ima svoju vlastitu DNA koja se razlikuje od DNA drugih vrsta i to je čini jedinstvenom. Dakle, kako će neki organizam izgledati i kako će funkcionirati, zapisano je u molekuli DNA, za koje kažemo da su molekule nasljeđa. Dijelovi molekule DNA koji određuju neku nasljednu osobinu nazivaju se geni i predstavljaju upute za upute za sintezu proteina - organskih spojeva koji sudjeluju u izgradnji stanice i odgovorni su za odvijanje i/ili ubrzavanje gotovo svih bioloških funkcija.



Slika 1. Građa DNA

DNA je složena molekula, polimer, sastavljen od mnogo malih jedinica – nukleotida, nanizanih jedan iza drugoga u dva lanca, omotana jedan oko drugoga (slika 1). Tu strukturu nazivamo dvostrukom zavojnicom.

Dvolančanost je ideja prirode kojom je postignuta povećana stabilnost molekule. Osim toga, komplementarnost lanaca i činjenica da oni definiraju jedan drugog, određuje princip semikonzervativne replikacije te omogućuje naknadni popravak eventualnih oštećenja.

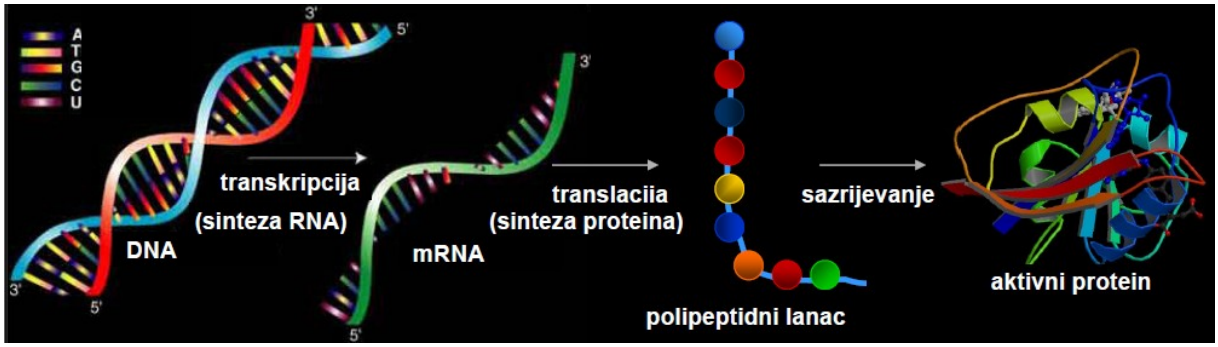
Jedinice od kojih je sastavljena DNA su nukleotidi, složene organske molekule građene od dušične baze (adenin, guanin, citozin i timin), šećera deoksiriboze i fosforne kiseline. DNA je oznaka za deoksiribonukleinsku kiselinu (eng. deoxyribonucleic acid*).*

U stanicama eukariota nalazimo dvije vrste DNA, koje razlikujemo prema smještaju unutar stanice. Jedna je smještena u jezgri i naziva se jezgrina ili nuklearna DNA, dok je druga, mitohondrijska DNA, smještena u mitohondrijima. Nuklearna DNA nasljeđuje se od majke i oca, a mitohondrijska isključivo od majke. Ukupna DNA organizma naziva se genom.

Osim DNA, u živom svijetu još nalazimo i ribonukleinsku kiselinu RNA *(eng.* ribonucleic acid*)* koja služi za prijenos nasljedne upute, ali ne i za njeno čuvanje. Ovo svojstvo omogućuje da DNA kao original genskog zapisa ne sudjeluje direktno u sintezi proteina, što bi bilo neprimjereno zbog njene veličine, ali i činjenice da stanica ne treba u svakom trenutku svaki protein za koji je šifra upisana u molekuli DNA. Osim toga, njeno direktno sudjelovanje u sintezi proteina dodatno bi ugrozilo njenu sigurnost i povećalo mogućnost oštećenja. Zbog toga je kratkovječnost molekule RNA prednost koja omogućava kontrolu ekspresije gena. Dugovječna RNA kodirala bi sintezu određenog proteina i onda kada on stanici više nije potreban, što ne bi bilo ekonomično.

Tri su tipa RNA: glasnička RNA (mRNA), koja sadrži prijepis nasljedne upute, transportna RNA (tRNA), koja donosi aminokiseline tijekom sinteze proteina i ribosomska RNA (rRNA) koja sudjeluje u građi ribosoma i aminokiseline povezuje u protein.

Cjelokupni proces nastanka funkcionalnog produkta (što je u najvećem broju slučajeva protein, a može biti i nešto drugo, ovisno u uputi sadržanoj u genu) naziva se ekspresija gena.



Slika 2. Ekspresija gena

2.2 Proces ekspresije gena

Prvi korak u ekspresiji gena je transkripcija ili prepisivanje DNA. U procesu transkripcije razlikujemo inicijaciju, elongaciju i terminaciju.

U fazi inicijacije RNA polimeraza (enzim koji katalizira sintezu RNA) veže se za mjesto na genu nazvano promotor, koji se u procesu transkripcije ne prepisuje. Promotor sadrži kratki niz naizmjenično vezanih T i A nukleotida uzvodno od mjesta početka transkripcije. Zbog ovog karakterističnog rasporeda parova AT, dio promotora se naziva i TATA blok, a sadrži ga 85% eukariotskih gena u svojim promotorima. Na inicijaciju djeluju i drugi elementi, kako udaljeni, tako i oni unutar gena. Pojačivači (eng. enhanceri) su sekvence koje stimuliraju transkripciju, mogu se nalaziti ispred (uzvodno) ili iza (nizvodno) od gena na udaljenosti od nekoliko tisuća baza (kb).

U fazi elongacije RNA polimeraza odvija jedan zavoj DNA pri čemu se lanci razdvajaju, a jedan od lanaca služi kao kalup. RNA polimeraza kreće se duž kalupa lanca i veže odgovarajuće nukleotide po principu komplementarnosti baza, uz zamjenu timina uracilom (U).

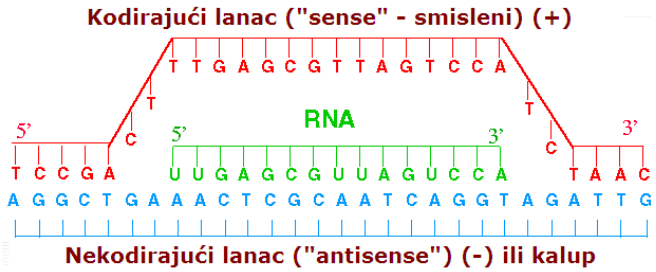
Posljednja je faza terminacije, u kojoj RNA polimeraza stiže do niza nukleotida koji se naziva terminacijski ili stop signal. Na tom mjestu se transkripcija

zaustavlja, a novonastala molekula RNA se odvaja od kalupa. Poznata su dva načina terminacije:

a) samoterminacija - ovisi samo o DNA sekvenci - najčešća je i obično se javlja pri sekvenci baza u lancu matrice DNA koja se sastoji od nekoliko uzastopnih adenina, a prethodi joj palindrom sekvenca.

b) terminacija pomoću proteina terminacije.

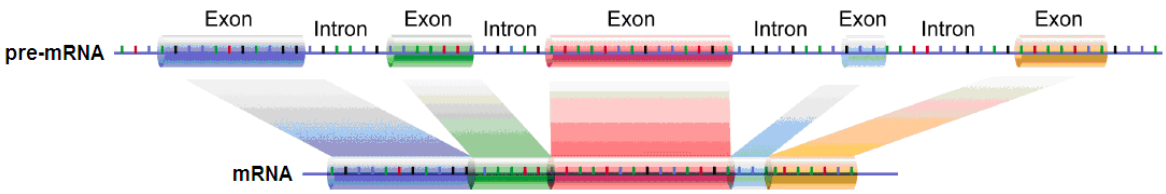
Novosintetizirana RNA jednaka je kodirajućem (smislenom) lancu, a komplementarna je kalupu.



Slika 3. pre-mRNA nastala procesom transkripcije

Proces transkripcije, kao i regulacija transkripcije, se znatno razlikuju kod prokariota i eukariota. Kod bakterija se transkripcija i translacija se odvijaju u citoplazmi (citosolu) te translacija započinje prije nego je potpuno završena transkripcija sinteza mRNA. Ova dva procesa su kod eukariota prostorno i vremenski odvojeni. Kod eukariota se proces transkripcije odvija u jezgri, pri čemu ne nastaje odmah zrela mRNA, već tzv. primarni transkript tj. nezrela RNA ili pre-mRNA (eng. pre-mature RNA).

Gen DNA se sastoji od eksona - kodirajućeg slijeda nukleotida koji se prevodi u proteine i introna – nekodirajućeg slijeda koji se ne prevodi u proteine, pa se prema tome i pre-mRNA nastala prepisivanjem DNA sastoji od eksona i introna. Prije izlaska iz jezgre nekodirajući introni se izrezuju, a eksoni spajaju u kontinuirani kodirajući segment. Proces se naziva obrada, izrezivanje ili prekrajanje RNA (eng. splicing processing), čime nastaje zrela mRNA.



Slika 4. Izrezivanje introna iz pre-mRNA

Poznato je da se na početku svakog introna u 99% slučajeva nalazi kombinacija nukleotida GT, a u 1% slučajeva kombinacija GC, dok je na kraju svakog introna kombinacija AG. Početak introna naziva se još i donorsko ili 5'-mjesto izrezivanja (eng. donor splice site), a kraj svakog introna zove se akceptorsko ili 3'- mjesto izrezivanja (eng. acceptor splice site).

Na početku procesa izrezivanja, na 5' kraj pre-mRNA dodaje se tzv. 5' Cap (RNA Cap) koju čine modificirani nukleotidi, koji štite mRNA od razgradnje i pomažu kod započinjanja translacije na ribosomima. Na 3' kraj dodaje se niz adenina (često i nekoliko stotina) koji štite mRNA od razgradnje i pomažu pri izlasku iz jezgre u citoplazmu (poli-A kraj). U proces izrezivanja introna uključene su male nuklearne ribonukleoproteinske čestice (čestice građene od proteina i male rRNA - snRNP).

Postoji 5 tipova snRNP: U1, U2, U4, U5 i U6, koji se udružuju s drugim proteinima u kompleks spliceosome. Spliceosome izrezuje introne na sljedeći način:

1. korak: nukleofilni napad 2-OH skupine na 5' kraj introna

2. korak: nukleofilni napad 3'OH skupine na 5' kraj eksona2

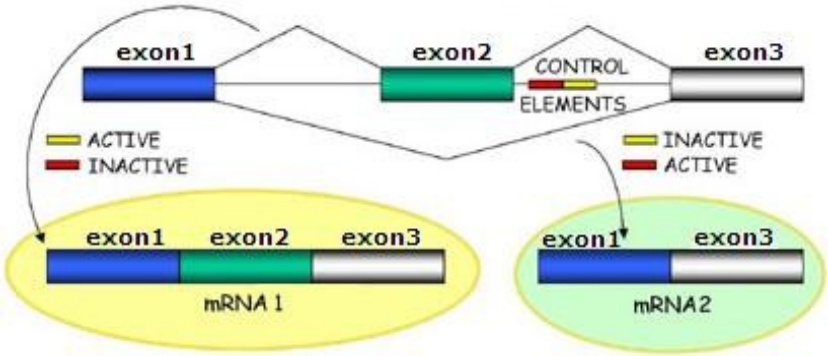
Nastala m-RNA se zatim transportira iz jezgre u citoplazmu gdje se odvija translacija. Translacija je sinteza proteina koja se odvija na ribosomima. Proteini su sastavljeni od niza aminokiselina, a koji protein će se sintetizirati translacijom, odnosno od kojih će aminokiselina biti građen i na koji će način one biti posložene u proteinu određeno je informacijama kodiranim u genima.

Svaka aminokiselina predstavljena je nizom od tri nukleotida. No, broj mogućih kombinacija nije jednak broju postojećih aminokiselina, tako da su neke aminokiseline određene s više kombinacija.

Tablica 1. Sve mogućnosti kombiniranja nukleotida u aminokiselinama

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

Pre-mRNA, produkt transkripcije eukariotskih gena može u nekim slučajevima dati više različitih mRNA, što je posljedica tzv. alternativnog procesiranja (prekrajanja), odnosno različitog izrezivanja introna i povezivanja eksona (eng. alternative splicing). Pretpostavlja se da čak 30% pre-mRNA u stanicama čovjeka podliježe alternativnom procesiranju. Prema tome ekspresija jednog gena rezultira sintezom dviju ili više različitih mRNA te, posljedično, sintezom dva ili više sličnih proteina.



Slika 5. Primjer nastajanja dviju različitih mRNA iz iste pre-mRNA alternativnim prekrajanjem

2.3 Genom organizma C.elegans

Crv Caenorhabditis elegans je prvi višestanični eukariotski organizam čiji je genom potpuno sekvenciran. Sekvenca je objavljena 1998. godine, s manjim pogreškama koje su do danas u većoj mjeri uklonjene, čime je dobivena vjerojatnost pojave pogreške pojedinog nukleotida u sekvenci manja od 1 na 100000 baza.

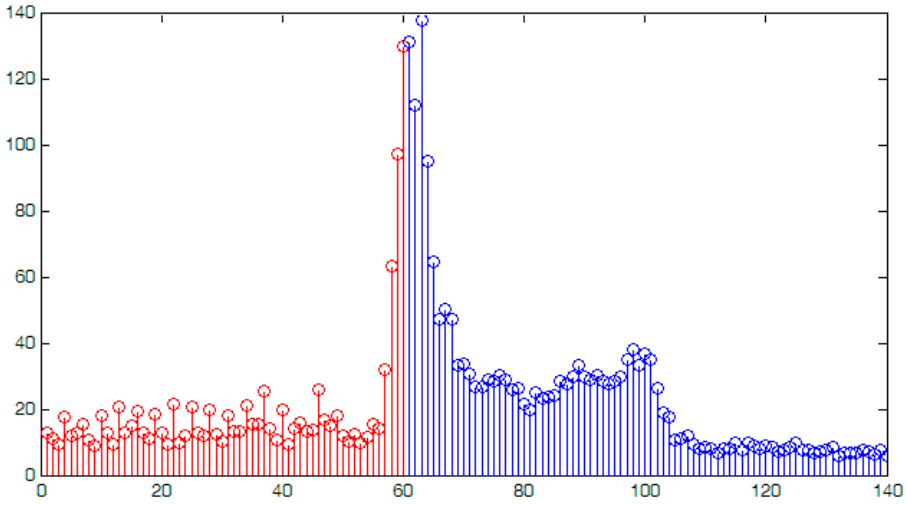
Genom C.elegansa sastoji se od više od 100 milijuna parova baza (~100.2Mb) i približno 20000 gena, raspodijeljenih u 6 kromosoma podjednakih veličina, označenih kao I, II, III, IV, V i X. U prosjeku veličina jednog gena je 3Kb i sadrži 5 introna.

Na Internetu je ovaj genom dostupan u različitim oblicima: u obliku potpune i slijedno zapisane sekvence genoma, u obliku datoteka u kojima je potpuni genom već organiziran po određenim značajkama, i u obliku datoteka u kojima su zapisani nukleotidi za pojedini gen. U genomu se naizmjenično i naizgled slučajnim redoslijedom pojavljuju nukleotidi A, C, G i T, te se nizovi ovih slova koriste kao signali u istraživanju i detekciji značajki gena.

U istraživanju su korištene datoteke s pojedinačnim genima izoliranim iz genoma, i datoteke u kojima su podaci reorganizirani prema 2 poznate značajke: da mjesta izrezivanja introna započinju s kombinacijom GT, a završavaju s kombinacijom AT. Na temelju toga genom je preoblikovan posebno za donore (početke izrezivanja introna), posebno za akceptore (krajeve izrezivanja introna), tako da su u genomu pronađena sve mjesta gdje se pojavljuje GT ili AG, i prema

njima su kreirani vektori duljine 398, kojima se na pozicijama 200 i 201 u donoru nalazi GT, a na pozicijama 198 i 199 u akceptoru AG. Svaka od ovih kombinacija u genomu je pronađena više od milijun puta, od čega je 64453 stvarnih mjesta početka izrezivanja, a 64838 stvarnih mjesta kraja izrezivanja.

Ovi podaci korišteni su i na Projektu u kojem je za detekciju mjesta izrezivanja korišten klasifikator Parallel Random Forest (PARF). Zbog nedostatka memorije za rad s ovako velikim podacima datoteka je skraćena tako da su vektori skraćeni na veličinu 140 nukleotida (60 lijevo od GT i 80 desno od GT), a broj vektora je smanjen na sve točne i 10% netočnih. PARF kreira detektor i formira nekoliko izlaznih datoteka, od kojih jedna sadrži brojeve koji predstavljaju važnost pozicije svakog od 140 nukleotida u vektoru. Crtanjem dobivenih podataka dobivena je slika.



Slika 6. Važnost pozicije nukleotida u okolini početka mjesta izrezivanja

U intronu, s desne strane mjesta izrezivanja, pokazuje se da su prvih 40 nukleotida veće važnosti, ali među njima uglavnom ne postoji veća razlika, dok je u eksonu, koji je s lijeve strane mjesta izrezivanja, vidljivo da veću važnost pokazuje svaki treći nukleotid.

2.3.1 Provjera poznatih značajki

Poznato je da donorsko mjesto izrezivanja počinje kombinacijom nukleotida GT, tj. da se u datotekama u vektorima pravih donora na pozicijama 200 i 201

nalazi GT, što je na slici mjesto između crvenih i plavih linija. Uzmu li se iz cijele datoteke donora samo oni koji su označeni kao točni, može se izračunati s kojom frekvencijom se svaki od nukleotida pojavljuje na svakoj poziciji u okolini stvarnih mjesta izrezivanja. Postavi li se prag da je veće važnosti nukleotid koji se na određenoj poziciji pojavljuje s vjerojatnošću većom od 0.5, MATLAB programski kôd ispisuje u datoteku sljedeći sadržaj:

```
Na poziciji 199 se s frekvencijom 0.556752 javlja adenin.
Na poziciji 200 se s frekvencijom 0.596678 javlja guanin.
Na poziciji 201 se s frekvencijom 1.000000 javlja guanin.
Na poziciji 202 se s frekvencijom 0.993970 javlja timin.
Na poziciji 203 se s frekvencijom 0.584372 javlja adenin.
Na poziciji 204 se s frekvencijom 0.664518 javlja adenin.
Na poziciji 205 se s frekvencijom 0.751187 javlja guanin.
Na poziciji 206 se s frekvencijom 0.618330 javlja timin.
Na poziciji 207 se s frekvencijom 0.510579 javlja timin.
```

Ispis potvrđuje poznatu pretpostavku da svaki pravi intron počinje s GT/GC te da postoje još neki značajni nukleotidi, koji se pojavljuju s visokim frekvencijama, i to samo u najbližoj okolini mjesta izrezivanja, s obje strane mjesta izrezivanja. Program je uz manje preinake (zbog pozicija mjesta izrezivanja, na kojoj se uvijek javlja AG) primijenjen i za akceptore, za koje se u datoteci dobiva ispis koji potvrđuje pretpostavku da svaki pravi intron završava s AG, te da u njegovoj najbližoj okolini sa strane introna postoji još nekoliko nukleotida koji se javljaju s višim frekvencijama. Ispis je sljedeći:

```
Na poziciji 183 se s frekvencijom 0.505383 javlja adenin.
Na poziciji 193 se s frekvencijom 0.569990 javlja timin.
Na poziciji 194 se s frekvencijom 0.888784 javlja timin.
Na poziciji 195 se s frekvencijom 0.974290 javlja timin.
Na poziciji 196 se s frekvencijom 0.673463 javlja timin.
Na poziciji 197 se s frekvencijom 0.834341 javlja citozin.
Na poziciji 198 se s frekvencijom 1.000000 javlja adenin.
Na poziciji 199 se s frekvencijom 1.000000 javlja guanin.
```

Slika 7. Poznati podaci o učestalosti najčešćih nukleotida u akceptorima

Ovi rezultati mogu poslužiti kao značajka u traženju mjesta izrezivanja na temelju svojstava introna, što je tema mnogih znanstvenih radova (slika 7, na

kojoj veličina slova koje predstavlja nukleotid proporcionalna njegovoj frekvenciji na toj poziciji). Jednako tako postoji vjerojatnost da se mjesta izrezivanja generiraju iz svojstava eksona, kao što je periodičnost s periodom 3.

2.3.2 Periodičnost s periodom 3

Kako se aminokiseline generiraju na temelju kodona od tri uzastopna nukleotida u eksonima, ideja kojoj se priklanjaju neki autori jest periodičnost eksona sa periodom 3. Prethodna slika i znanstveni radovi koji analiziraju ovo svojstvo navode na zaključak da bi se detekcijom pojave periodičnosti u genu mogle dobiti točne pozicije eksona i introna, odnosno mjesta izrezivanja. Za genom organizma C.elegans objavljeno je nekoliko radova s ciljem potvrđivanja ove pretpostavke, a s obzirom da postoji sličnost između genoma crva i čovjeka, na temelju ovih rezultata vjerojatno bi se moglo doći i do značajki za pravilnosti u genomu čovjeka.

2.4 Prilagodba genoma za digitalnu obradu

Geni su dostupni kao nizovi slova A, C G i T koji predstavljaju odgovarajuće nukleotide. Slova je potrebno pretvoriti u jednostavan numerički oblik prikladan za digitalnu obradu, što se može izvesti na više načina.

U ovom radu primijenjena je metoda kojom se iz promatranog gena kreiraju četiri binarna signala, na temelju postojeća četiri nukleotida. Prisutnost određenog nukleotida u novoj sekvenci označava se postavljanjem jedinice na istu poziciju, a neprisutnost postavljanjem nule. Na primjer, iz sekvence AATGGGTCCAGCTCCAGTTT nastati će niz 11000000010000010000 za adenin, 00000011001011000 za citozin, te analogno tome i za guanin i timin.

Preostale dvije metode baziraju se na EIIP vrijednostima nukleotida (eng. electron-ion interaction potential). Izborom prve metode nastaju nova četiri niza množenjem navedenih binarnih nizova s EIIP vrijednošću promatranog nukleotida. Izborom druge metode originalna sekvenca se ne rastavlja na četiri nove, nego se

na svakoj lokaciji duž sekvence upisuje EIIP vrijednost nukleotida koji se na toj lokaciji nalazi.

Tablica 2. EIIP vrijednosti nukleotida

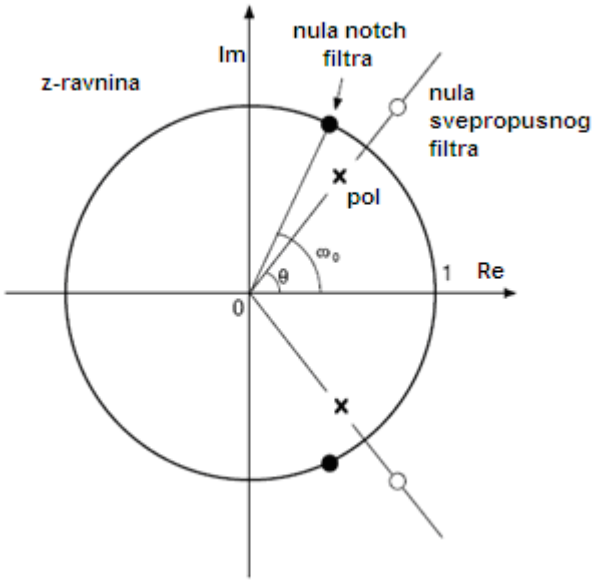
nukleotid	A	C	G	T
EIIP	0.1260	0.1335	0.1340	0.0806

3. Filtri temeljeni na svojstvima eksona

3.1 Filtri temeljeni na svepropusnom filtru

Svepropusni (eng. allpass) filtar drugog reda, s polovima u točkama $\mathrm{Re}^{\pm j\theta}$, ima prijenosnu funkciju:

$$A(z) = \frac{R^2 - 2R\cos\theta z^{-1} + z^{-2}}{1 - 2R\cos\theta z^{-1} + R^2 z^{-2}} \tag{1}$$



Slika 8. Polovi i nule svepropusnog filtra

R i θ su proizvoljno odabrani parametri, pri čemu se radi stabilnosti uzima R<1. Iz svepropusnog filtra izvode se dva filtra, notch i antinotch filtar, na sljedeći način:

$$\begin{bmatrix} G(z) \\ H(z) \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 \\ A(z) \end{bmatrix} \tag{2}$$

Notch je takav filtar koji propušta sve frekvencije signala osim frekvencije θ. Njegova prijenosna funkcija opisana je izrazom (3).

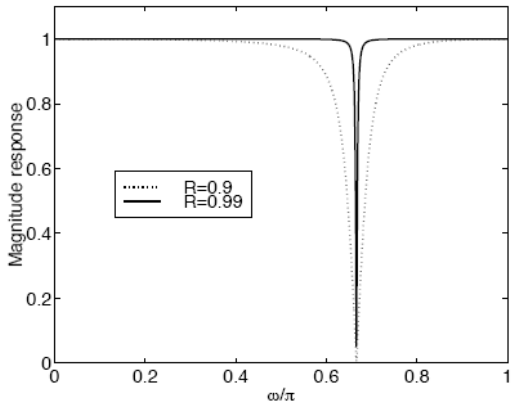
$$G(z) = \frac{1}{2}\frac{1 + R^2 - 4R\cos\theta z^{-1} + (1 + R^2)z^{-2}}{1 - 2R\cos\theta z^{-1} + R^2 z^{-2}} \tag{3}$$

Notch i antinotch su komplementarni filtri, pa antinotch propušta samo frekvenciju θ, a sve ostale prigušuje. Prijenosna funkcija antinotch filtra je:

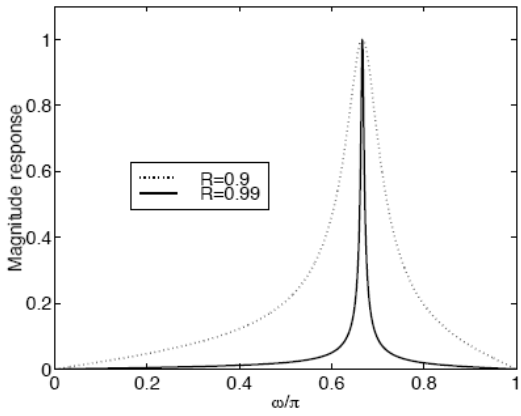
$$H(z) = \frac{1}{2}\frac{(1-R^2)+(R^2-1)z^{-2}}{1-2R\cos\theta z^{-1}+R^2 z^{-2}} \tag{4}$$

Za različite R mijenja se selektivnost filtra: što je R bliže jedinici, frekvencije koje je potrebno prigušiti jače se potiskuju, tj. selektivnost filtra se povećava. Međutim, impulsni odziv ovakvih filtara je beskonačan, što smanjuje preciznost u vremenskoj domeni.

Za *R=0.99* i $\theta = \frac{2\pi}{3}$, notch filtar guši frekvenciju $\frac{2\pi}{3}$, a povratkom u vremensku domenu prigušeni su dijelovi signala periodični sa periodom 3. Za antinotch filtar s istim parametrima, biti će propušteni dijelovi signala koji se ponavljaju s periodom 3, a svi ostali će biti prigušeni. Slike prikazuju amplitudni odziv, odnosno amplitudno-frekvencijske karakteristike notch i antinotch filtra za navedeni θ, te za *R=0.9* i *R=0.99*.



Slika 9. A-F karakteristika notch filtra



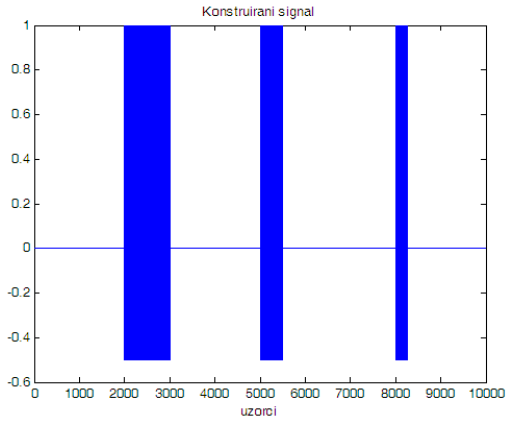
Slika 10. A-F karakteristika antinotch filtra

3.2 Konstrukcija detektora

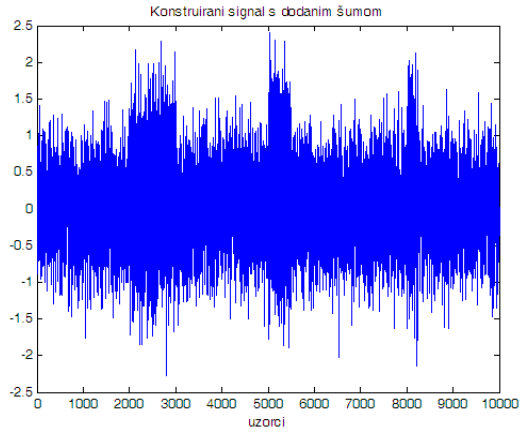
Za analizu detektora koji će se koristiti u istraživanju formira se idealni signal duljine 10000 nukleotida, koji je periodičan s periodom 3 samo na nekim svojim dijelovima. Konstruirani signal $x(t)$ opisan je formulom (5)

$$x(t) = \begin{cases} \cos(\frac{2\pi}{3}t), & t \in [2000,3000] \quad [5000,5500] \quad [8000,8250] \\ 0, & inace \end{cases} \tag{5}$$

Signali koji će se koristiti u istraživanju su stvarni biološki signali koji predstavljaju genom, sastavljeni od niza naizmjenično složenih A, C, G i T nukleotida. Ukoliko jedan ili više njih pokazuje periodičnost s periodom 3, tada preostali nukleotidi za njega predstavljaju šum, zbog čega se konstruiranom idealnom signalu (slika 11) superponira bijeli šum (slika 12).



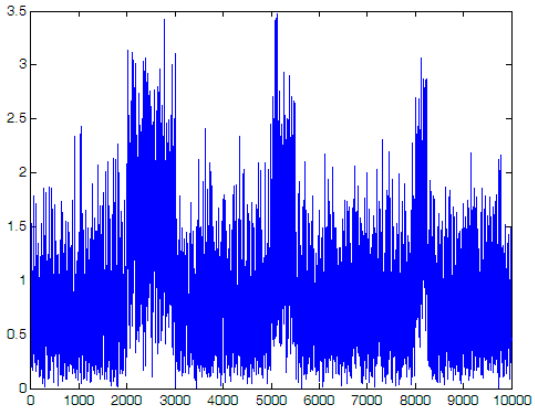
Slika 11. Konstruirani signal bez šuma



Slika 12. Konstruirani signal s dodanim šumom

Takav signal propušta se kroz FIR filtar (eng. Finite Impulse Response) impulsnog odziva *h* čime nastaje signal prikazan slikom 12. Ovaj signal koristi se kao ulazni signal za selektivne filtre antinotch i notch, koji će izdvojiti frekvenciju $\frac{2\pi}{3}$.

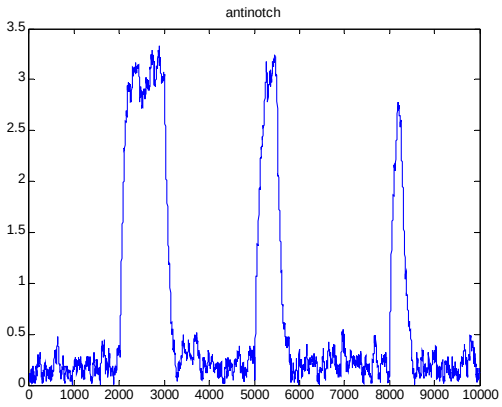
$$h = \left\{ \frac{-1-\sqrt{-3}}{2}, 1, \frac{-1+\sqrt{-3}}{2} \right\} \tag{6}$$



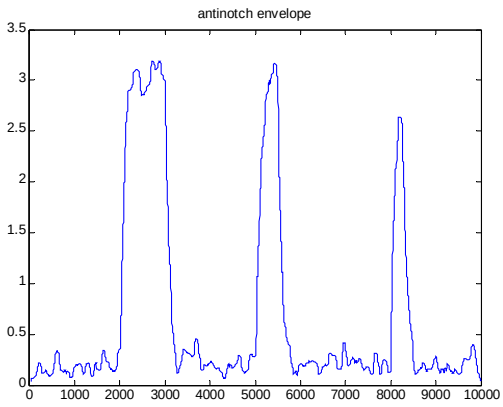
Slika 13. Signal propušten kroz FIR filtar impulsnog odziva (5)

Notch i antinotch filtri detektiraju periodične dijelove konstruiranog signala dajući na svojim izlazima signale na slikama 14 i 16, no prolaskom kroz njihovu kaskadu dobiven je signal iz kojeg je moguće preciznije očitati željene pozicije (slika 18) jer se na početku periodičnih segmenata pojavljuju strmiji nagibi.

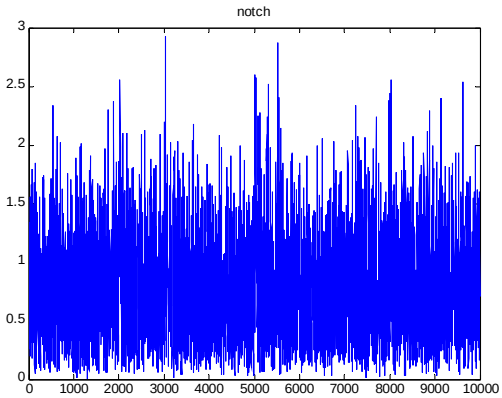
Umjesto signala na izlazu iz filtara, zbog velikog broja uzoraka signala, radi se s njihovim ovojnicama dobivenim usrednjavanjem apsolutnih vrijednosti istih signala na neparnom broju uzoraka (99). Usrednjavanje se postiže funkcijom medijan kojom se minimizira odstupanje originalnih vrijednosti od aproksimativne. Broj uzoraka je parametar koji se može mijenjati.



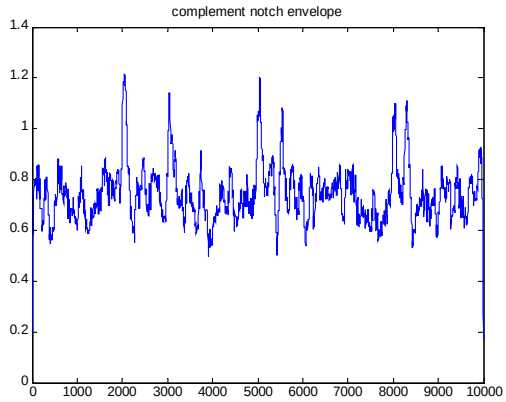
Slika 14. Signal na izlazu antinotch filtra



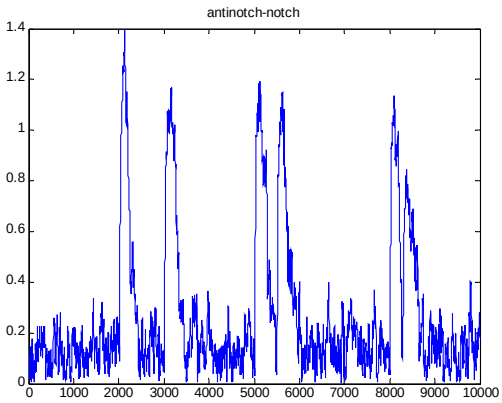
Slika 15. Njegova ovojnica



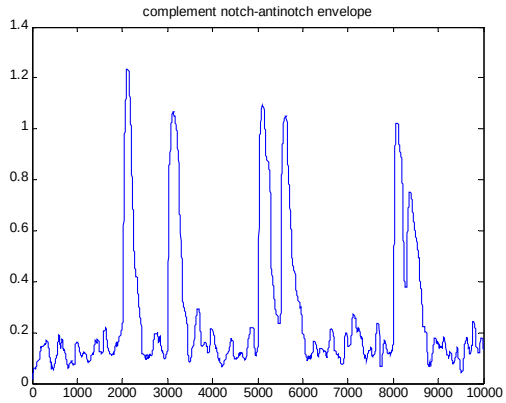
Slika 16 . Signal na izlazu notch filtra



Slika 17. Njegova ovojnica



Slika 18. Signal na izlazu antinotch-notch

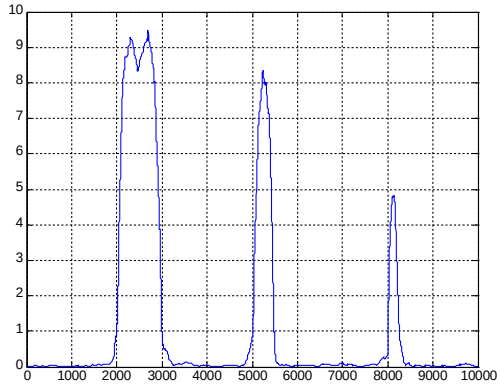


Slika 19. Njegova ovojnica

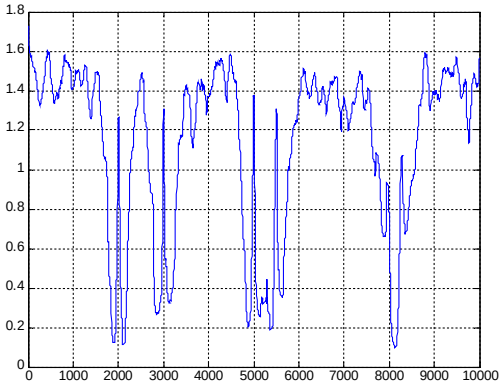
Da bi se dobili strmi rubovi i na kraju periodičnog segmenta, odnosno povećala preciznost na kraju istog na onakvu kakva je dobivena i za njegov početak, potrebno je „preokrenuti vrijeme“, što je moguće učiniti preokretanjem signala u vremenu, jer je vrijednost signala u svakom trenutku poznata unaprijed. Na ovaj način izveden je antikauzalni filtar. Ovojnice signala koje se dobiju na njihovim izlazima zrcaljene su slike ovojnica dobivenih na izlazima kauzalnih filtara.

Antinotch filtar i kaskada antinotch-notch bolje izdvajaju željene dijelove signala. Selektivnost se može povećati komplementiranjem izlaza kaskade i dijeljenjem izlaza antinotcha s izlazom kaskade. Izlaz kaskade se komplementira oduzimanjem od njegovog maksimuma, a ne od nule, kako bi se dobio broj u intervalu (-1,1). Dijeljenjem nekog broja s brojem iz tog intervala dobiva se još

veća amplituda u toj točki. Komplement se dodatno pomiče za 0.1, kako bi se pri kasnijem dijeljenju izbjeglo dijeljenje s nulom.

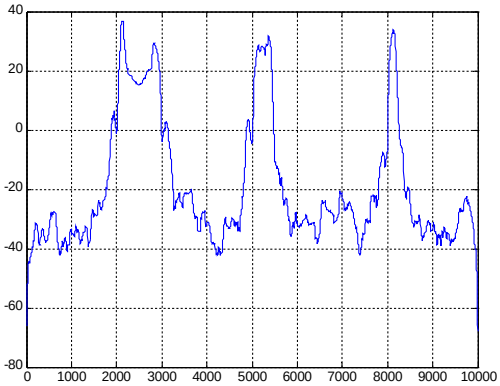


Slika 20. Produkt kauzalnog i antikauzalnog antinotcha

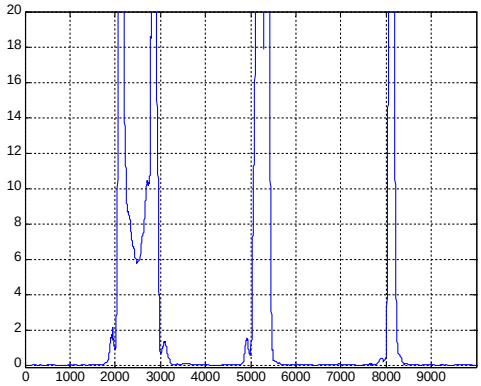


Slika 21. Komplement kauzalne i antikauzalne kaskade a-n

Konačno, željeni detektor je dobiven kao omjer produkta izlaznog signala kauzalnog i antikauzalnog antinotch filtra i produkta signala kauzalne i antikauzalne kaskade antinotch-notch. Brojnik i nazivnik iz ovog omjera prikazani su gornjim slikama.



Slika 22. Izlaz filtra u logaritamskom mjerilu



Slika 23. Izlaz filtra u linearnom mjerilu

Prolaskom konstruiranog signala kroz ovakav sustav, na izlazu se formira signal prikazan slikom 23, u linearnom mjerilu, gdje su prisutni izrazito strmi bridovi (skokovi) na samoj poziciji početka periodičnog dijela u signalu, ili vrlo blizu njega. Na ovaj način kreiran je filtarski detektor, koji izdvaja dijelove signala željenog perioda. Za idealni signal bez šuma ovakav detektor daje izrazito precizne

rezultate, a uz superponirani šum preciznost se u manjoj mjeri smanjuje. U nastavku se detektor primjenjuje na realne, biološke signale (gene) iz organizma C.elegans, u kojima je potrebno detektirati njihove periodične dijelove, što su s biološkog stajališta eksoni.

3.3 Provjera periodičnosti DFT transformacijom

Pretpostavka o periodičnosti nukleotida s periodom 3 u eksonima, navedena u nekim znanstvenim radovima, ne eliminira mogućnost da u eksonima postoje signali periodični s nekim drugim periodom. Kako bi se istovremeno ispitala periodičnost s periodom 3, ali i sa svim ostalim periodima, gen je analiziran i računanjem DFT (eng. Discrete Fourier Transformation) na segmentima gena koji odgovaraju eksonima. DFT koeficijenti računaju se prema izrazu (7).

$$X(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}nk} \tag{7}$$

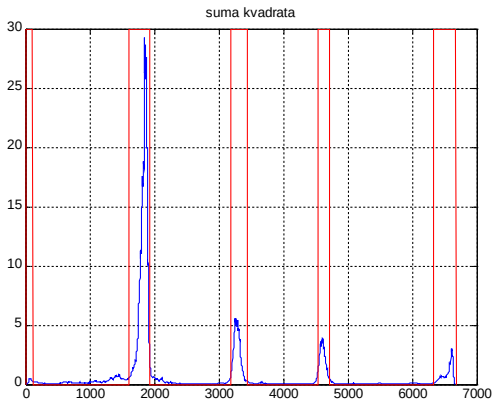
Programski kod napisan u MATLAB-u izolira eksone iz gena, te za svaki ekson podijeljen u nizove po nukleotidima računa DFT i ispisuje koji od njih ima najveću apsolutnu vrijednost Fourierovog koeficijenta za frekvenciju $2\pi/3$, crta DFT tog niza te za njega računa apsolutnu vrijednost maksimalnog koeficijenta. Ako se ove dvije vrijednosti podudaraju, tada taj nukleotid pokazuje periodičnost isključivo s periodom 3. Izračun DFT programski se prilagođava duljini sekvence eksona u genu.

4. Primjena detektora i DFT-a na C.elegans

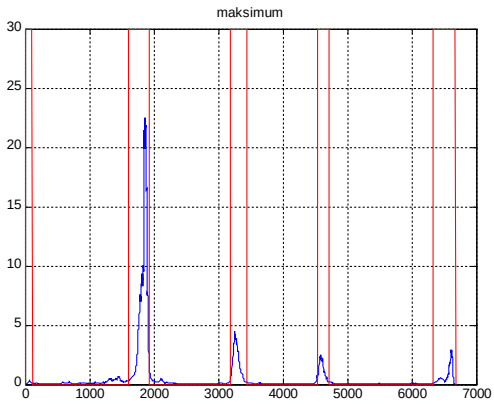
4.1 Primjena detektora i DFT-a na pojedinačne gene

4.1.1 Gen F56F114a

Gen F56F114a je gen koji se analizira u znanstvenim radovima temeljenim na detekciji kodirajućih područja u genima detekcijom pojave periodičnosti s periodom 3. Duljine je 6677 nukleotida, raspodijeljenih u 5 eksona i 4 introna različitih duljina. Signali nastali iz ovog gena propuštaju se kroz prethodno opisani detektor. S obzirom da se sada ne radi s jednim signalom nego sa četiri, konačni rezultat se računa kao suma kvadrata svakog od njih. Međutim, znamo li da se iz sekvence gena aminokiseline generiraju na temelju kodona od 3 nukleotida, logičan zaključak je da se neće svi nukleotidi istovremeno periodički ponavljati duž cijelog eksona, jer bi se tada iste kombinacije četiri nukleotida u eksonu ponavljale uzastopno jedna za drugom. Zbog toga je odabran još jedan način računanja konačnog rezultata, tako da se uzima maksimum između četiri sekvence za svaku poziciju duž gena. Dobiveni rezultati prikazani su sljedećim slikama.



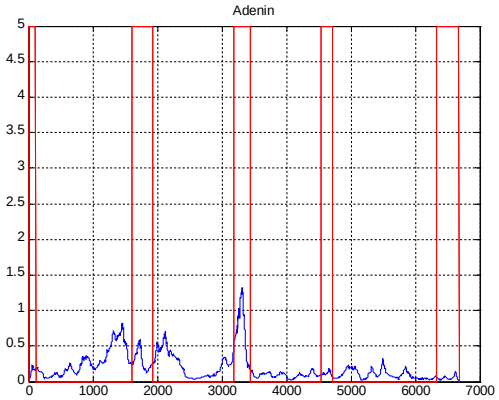
Slika 24. Rezultat detektora kao suma kvadrata



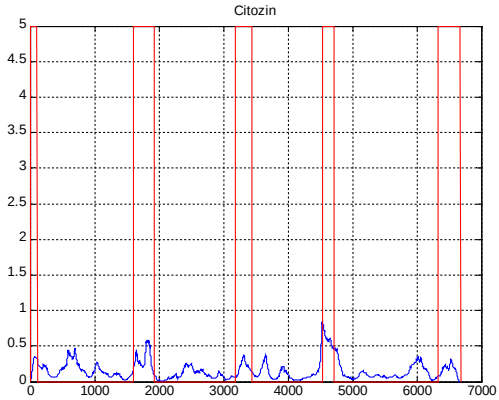
Slika 25. Rezultat detektora kao maksimum

Crvenim linijama na slikama označene su točne lokacije eksona i introna u promatranoj sekvenci, a plavim linijama lokacije dobivene konstruiranim detektorom. Uz pretpostavku da introni ne pokazuju, a eksoni pokazuju svojstvo periodičnosti s periodom 3, detekcija njihovih rubova je dosta precizna, u ovom slučaju i podjednaka za oba načina računanja konačnog rezultata. No promatra li

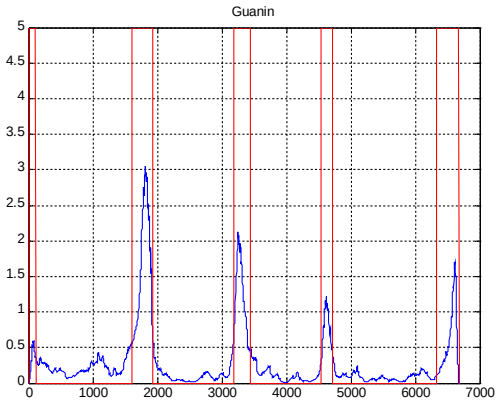
se svaki nukleotid posebno na izlasku iz detektora, može se uočiti da se zaista samo neki nukleotidi u eksonu periodički pojavljuju češće od ostalih. U ovom genu u eksonima je periodičnost s periodom 3 izraženija kod guanina te nešto manje kod timina, dok adenin i citozin zapravo i ne pokazuju periodičnost.



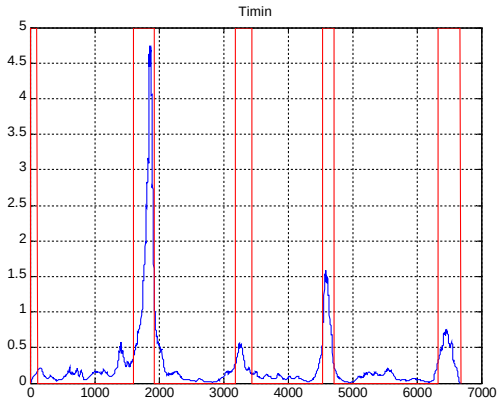
Slika 26. Periodičnost adenina u F56F114a



Slika 27. Periodičnost citozina u F56F114a



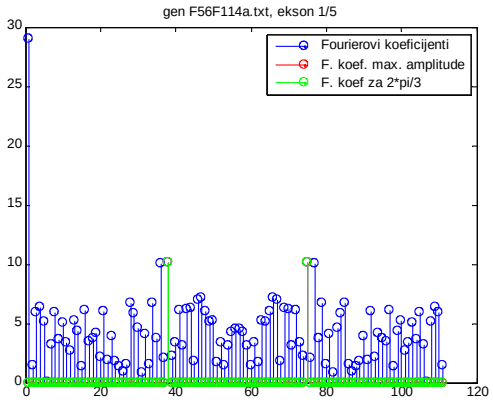
Slika 28. Periodičnost guanina u F56F114a



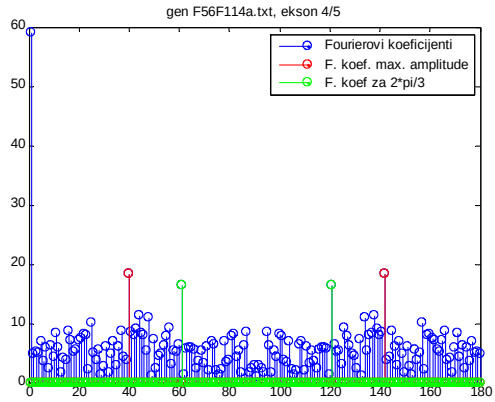
Slika 29. Periodičnost timina u F56F114a

Periodičnost s periodom 3 pretpostavka je koja se pokazala točnom za ovaj gen. To ipak ne znači da je period 3 jedini period koji se pojavljuje u njegovim eksonima, zbog čega se dodatno računa DFT kako je objašnjeno u poglavlju 2.3.

Program koji računa Fourierove koeficijente izračunava DFT. U promatranom genu za postojećih 5 eksona najveću periodičnost s 3 redom pokazuju G, T, G, T i T, ali dok je u prva 3 eksona to ujedno i maksimalni koeficijent za taj nukleotid, za 4. i 5. ekson to nije slučaj.



Slika 30. DFT za prvi ekson za guanin



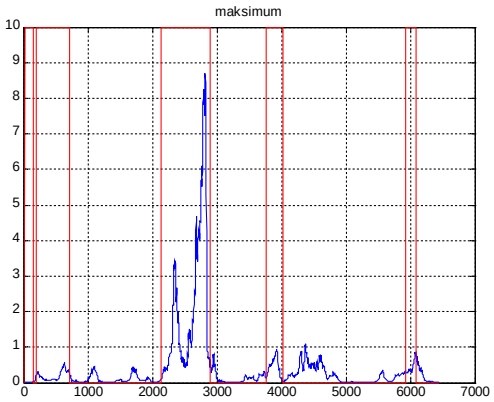
Slika 31 DFT za četvrti ekson za timin

Primjerice, za prvi ekson izračunato je da na poziciji koja odgovara frekvenciji $2\pi/3$ najveći DFT koeficijent pokazuje guanin. Na slici je prikazana DFT guanina. Crveno je obojen koeficijent najveće amplitude za guanin, a zelene boje koeficijent amplitude za frekvenciju $2\pi/3$ za guanin. Za prvi, drugi i treći ekson u genu F56F114a ova dva koeficijenta se preklapaju, čime se dobiva slika kvalitativno jednaka slici 30. Za četvrti i peti ekson dobiva se slika kvalitativno jednaka slici 31., na kojoj se vidi da nukleotid koji u eksonu pokazuje najveću periodičnost s periodom 3 istovremeno pokazuje još izraženiju periodičnost s nekim drugim periodom.

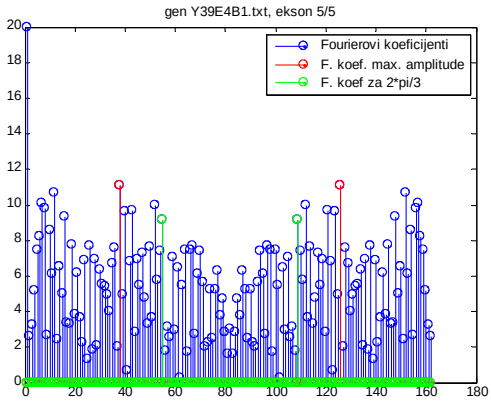
Jednako dobri rezultati dobivaju se za još nekoliko gena, kao što su K12C11.1 (duljine 3475 nukleotida i 3 eksona), YML056C (duljine 1983 nukleotida i 2 eksona), B033662 (duljine 2512 nukleotida i 5 eksona), R06F69 (duljine 1360 nukleotida i 3 eksona).

4.1.2 Gen Y39E4B1

Gen Y39E4B1 je gen duljine 6441 nukleotid, sastavljen od 5 eksona i 5 introna, dakle podjednake duljine i jednakog broja eksona i introna kao i gen F56F114a. Međutim, detektor pronalazi signal periodičnosti sa 3 izražen samo u trećem eksonu, i to s nešto manjom preciznošću u odnosu na F56F114a, dok se za ostatak eksona u genu zbog niskih amplituda izlaznog signala može reći da periodičnosti nema, jer se podjednake amplitude pojavljuju i u intronima. Prema tome, detektiran je samo jedan od pet postojećih eksona.



Slika 32. Rezultat detektora za Y39E4B1

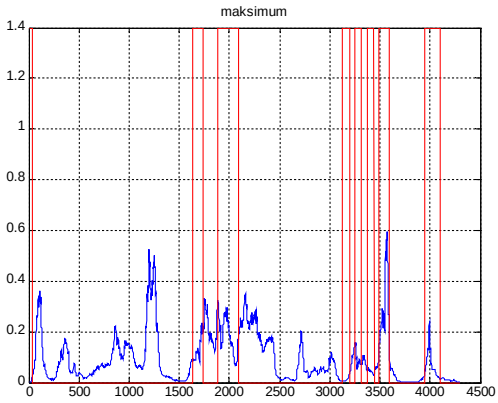


Slika 33. DFT za peti ekson za adenin

Ovo potvrđuje i analiza DFT-om, koja izračunava da po eksonima najveću periodičnost s 3 pokazuju redom T, A, G, G i A, ali promatrajući ih pojedinačno, npr. promatrajući adenin koji u zadnjem eksonu pokazuje najizraženiju periodičnost s periodom 3, koeficijent na mjestu koje odgovara ovoj frekvenciji se ne ističe, odnosno period s 3 uglavnom se ne pokazuje kao značajniji od ostalih. Dakle, u ovom slučaju introni se mogu smatrati eliminiranima, ali i neki eksoni, pa se svojstvo periodičnosti sa 3 ipak ne može smatrati dovoljnim kriterijem za pronalaženje svih eksona i za ovaj gen ta se područja ne mogu odrediti ovakvim detektorom.

4.1.3 Gen AH96

Gen AH96 je gen duljine 4297 nukleotida i 8 eksona. Kao što se vidi iz crvenih rubova na slici, detektirana periodična područja većim dijelom nalaze se u području introna, a manjim dijelom u području eksona.



Slika 34. Rezultat detektora za AH96

U ovom slučaju ne samo da eksoni ne pokazuju željeno svojstvo, nego su amplitude periodičnih segmenata veće upravo kod introna koje smo na temelju nepostojanja tog svojstva htjeli eliminirati. Ipak, ove amplitude su i dalje dovoljno niske da bi se uočile razlike između eksona i introna, i korišteni detektor ne omogućava procjenu položaja eksona u genu.

4.2 Primjena DFT na potpuni genom

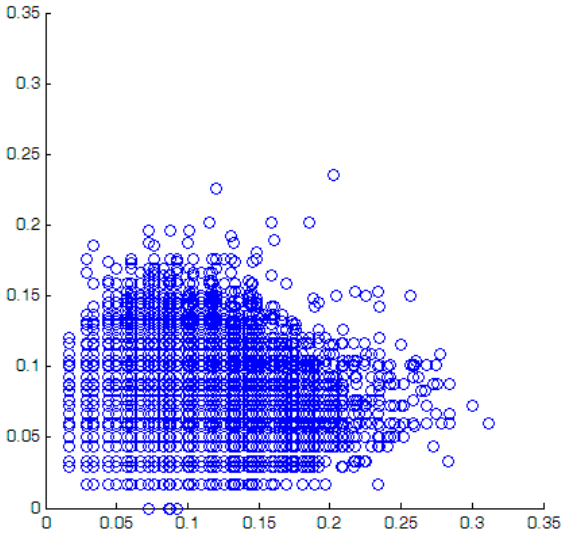
Potpuni genom organizma C.elegans dostupan je na Internetu u obliku slijednog niza nukleotida po kromosomima, ali s nepoznatim lokacijama mjesta izrezivanja, zbog čega se ovakav niz ne može koristiti za detektor pomoću filtara, jer se ne može usporediti s točnim podacima. Zbog toga je korišten drugi oblik podataka, u kojima su podaci reorganizirani u 2 datoteke, za donore i akceptore. Iz ovih datoteka uzeti su samo vektori u kojima su prisutna točna mjesta izrezivanja.

Program koji računa DFT na ovom obliku podataka izmijenjen je tako da računa DFT koeficijente posebno s lijeve strane odgovarajuće kombinacije nukleotida (GT za donore i AG za akceptore), a posebno s desne strane.

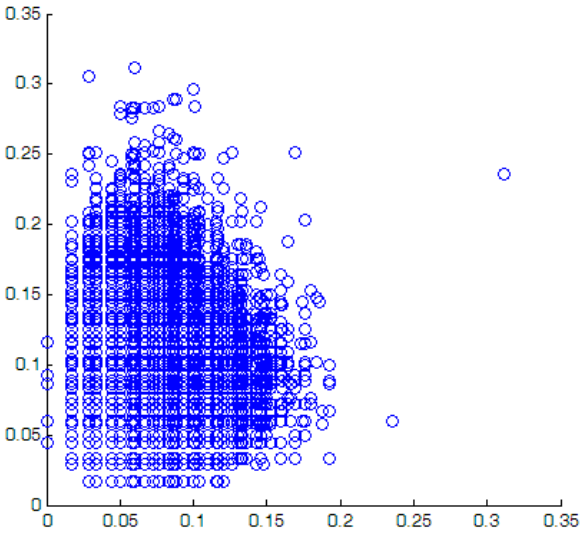
Promatraju li se donori, u sredini vektora je kombinacija nukleotida GT. S njezine lijeve strane je ekson, a s desne strane intron. Prema polaznoj pretpostavci, prema kojoj ekson pokazuje periodičnost s periodom 3, a intron to svojstvo ne pokazuje, očekuje se da će se izračunom DFT koeficijenata s lijeve strane dobiti veći koeficijent na poziciji koja odgovara frekvenciji $2\pi/3$ nego s desne strane. Crtanjem parova ova dva koeficijenta za sve vektore koji su stvoreni oko pravog mjesta izrezivanja u koordinatni sustav, točke koje predstavljaju parove koeficijenata za svaki vektor trebale bi se grupirati u donji desni rub koordinatnog sustava.

Ekvivalentno razmišljanje primijenjeno je i na akceptore. Sada je u sredini vektora kombinacija nukleotida AG, s njezine lijeve strane intron, a s desne strane ekson, i uz istu pretpostavku očekuje se da će se izračunom DFT koeficijenata s lijeve strane dobiti manji koeficijent na poziciji koja odgovara frekvenciji $2\pi/3$ nego s desne strane. Crtanjem parova ova dva koeficijenta u koordinatnom sustavu, pripadne točke trebale bi se grupirati u njegov gornji lijevi rub.

Kako su duljine eksona i introna u ovom obliku podataka nepoznate, duljine s lijeve i desne strane od središnje sekvence, kojima će se računati DFT, parametar su koji se može mijenjati. Programi su pokrenuti za donore i akceptore na način da je uzeto 60 nukleotida s lijeve i 60 nukleotida s desne strane, te su dobivene slike 35 i 36.



Slika 35. Parovi DFT za donore (60)



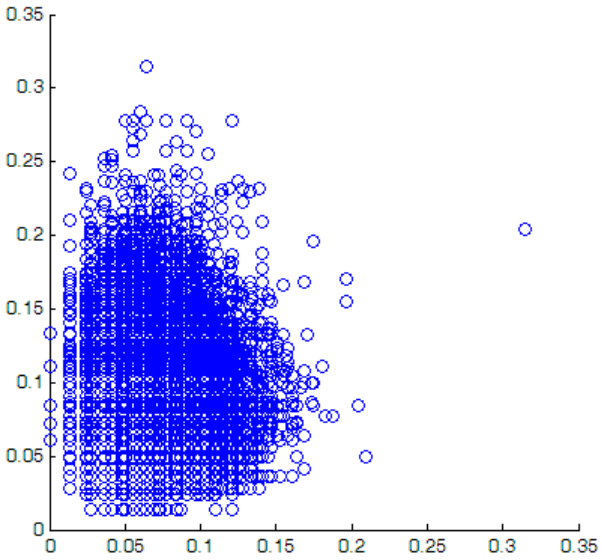
Slika 36. Parovi DFT za akceptore (60)

Očekivano grupiranje u donjem desnom kutu za donore ili gornjem lijevom kutu za akceptore značilo bi da je svojstvo periodičnosti s periodom 3 na temelju kojeg su točke dobivene realno svojstvo eksona na temelju kojeg je moguće detektirati njegov početak i kraj. Grupiranje nije ostvareno na željeni način, ali može se uočiti da ipak postoji dio gena za koje vrijedi da je periodičnost sa 3 u eksonu veća od periodičnosti sa 3 u intronu. Npr. promatraju li se samo geni čiji je DFT koeficijent u eksonima veći od 0.2, a u intronima manji od 0.1, za donore se dobiva grupa od 198 vektora koji zadovoljavaju taj kriterij, što je samo 0.0031% gena, a za akceptore se dobiva grupa od 226 vektora, što čini 0.0035% gena. Prema tome, za vrlo mali dio genoma detekcija periodičnosti sa periodom 3 može se smatrati dovoljnim detektorom za pronalaženje mjesta izrezivanja, ali za najveći dio genoma detekcija neće postići željene rezultate.

4.3 Detekcija drugih perioda na potpunom genomu

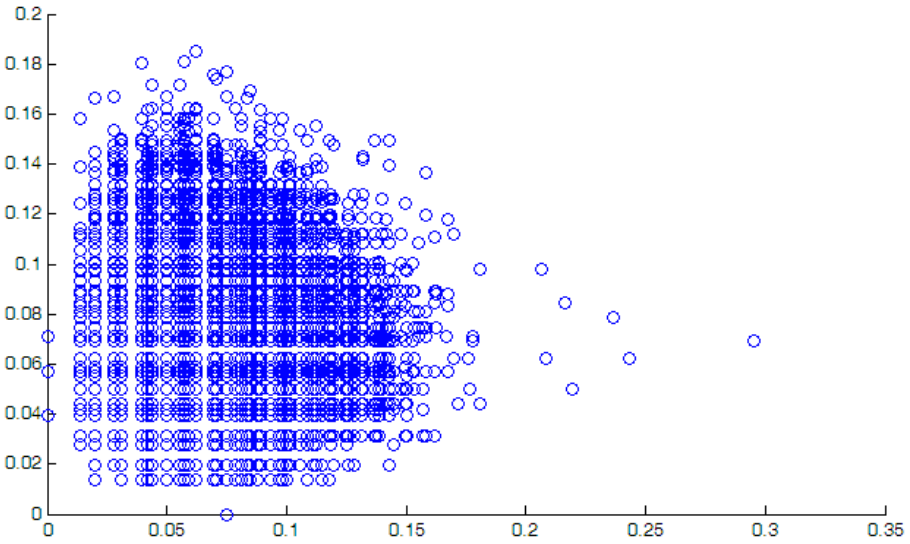
Kako je metoda detekcije eksona korištenjem filtara temeljenih na svepropusnom filtru u kombinaciji s izračunavanjem DFT koeficijenata pokazala da nukleotidi koji u eksonu pokazuju najizraženiju periodičnost s periodom 3 često istovremeno pokazuju još izraženiju periodičnost sa nekim drugim periodom, postavlja se ideja da bi se promatranjem potpunog genoma na jednak način kao u prethodnom poglavlju podaci mogli grupirati na temelju DFT koeficijenata izračunatog za neki drugi, proizvoljno odabrani period, odnosno provjeriti pojavljuje li se između eksona i introna razlika u periodičnosti s novim odabranim periodom.

Za akceptore provjereni su svi periodi između 3 i 9, na duljinama 72 s lijeve i 72 s desne strane. Za period 3 dobivena je slika slična onoj na kojoj je rezultat programa za isti period 3, ali uz vektore duljina 60. Kako je očekivano, povećanjem duljine vektora na kojima se računa DFT dobiva se lošiji rezultat, jer uzme li se dodatno kriterij da bi se točke grupirale kada bi DFT koeficijent za ekson bio veći od 0.2, a za intron manji od 0.1, tada za grupiranje preostaje samo 150 točaka.



Slika 37. Parovi DFT koeficijenata za akceptore (72) za period 3

Za period 4 dobivena je slika 38, a vrlo slične se dobivaju i za preostalih pet perioda. Iz njih je očigledno da se periodičnost s tim periodima ne ponaša različito u eksonima i intronima i nije kriterij za njihovo razlikovanje.



Slika 38. Parovi DFT koeficijenata za akceptore (72) za period 4

Prema dobivenim rezultatima, detekcijom perioda 4-9 eksoni i introni neće biti razlikovani, a detekcijom period 3 eksoni će biti prepoznati samo u manjem broju slučajeva, dok će za znatno veći broj slučajeva ipak biti potrebno upotrijebiti neke druge značajke eksona ili introna.

5. Zaključak

Dosad dobiveni rezultati znanstvenih instituta dijelom su činjenice koje se mogu dalje koristiti u istraživanjima kao dio svojstava na temelju kojih se može izvesti klasifikacija podataka, a dijelom su pretpostavke koje se temelje na manjoj ili većoj vjerojatnosti. Jedna od takvih pretpostavki je i periodičnost kodirajućih dijelova gena (eksona) s periodom 3. Periodičnost s periodom 3, ali i sa bilo kojim drugim periodom, moguće je detektirati i ispitati na više načina, i usporedbom sa poznatim podacima o lokacijama kodirajućih dijelova u genu zaključiti kolika je točnost, a zatim i preciznost ovakvih detektora.

Za detekciju eksona u radu je korišten konstruirani filtar koji izdvaja područja frekvencije 2pi/3, odnosno perioda 3. Kreirani filtar je visoke preciznosti i uspješno detektira eksone koji pokazuju navedeno svojstvo, no ono u većem broju slučajeva ipak nije prisutno i takvi eksoni neće biti detektirani. Ovo potvrđuje i DFT transformacija izračunata za ekson i intron u okolini istog mjesta izrezivanja. Kada bi pretpostavka o periodičnosti sa periodom 3 bila točna, DFT koeficijent bi za frekvenciju $2\pi/3$ uvijek bio veći za ekson i podaci za koje je izračunata DFT bi se grupirali na odgovarajućem mjestu u koordinatnom sustavu. Ekvivalentno rezultatima prvog detektora, podaci su se grupirali na način da je samo manji dio grupiran u željenoj točki, a svi ostali u području koje ne pokazuje razliku između eksona i introna. Dakle, postoji vrlo mali udio gena za koje promatrano svojstvo vrijedi, i zato se periodičnost sa 3 ne može promatrati kao svojstvo koje je samo po sebi dovoljno za detekciju eksona, no postoji li više mehanizama detekcije, što je još uvijek nepoznato, postoji mogućnost da bi se kombinacijom detektora koji prepoznaju sve mehanizme izrezivanja nekodirajućih dijelova ovi eksoni prepoznali detekcijom perioda 3, a ostali detekcijom značajki koje za njih vrijede.

6. Literatura

[1] 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

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

[3] Parađina N., *Analiza DNK metodama obrade signala u vremenskoj i frekvencijskoj domeni*, Diplomski rad br. 1143, FER, 2008.

[4] Penović M., diplomski rad, 2008.

[5] 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, 8.3.2009.

[6] Mark Blaxter, An *introduction to the genome of the nematode Caenorhabditis elegans,* http://www.nematodes.org/Caenorhabditis/caenorhabditis_genome.shtml

[7] http://en.wikipedia.org/wiki/DNA, 14.4.2009.

7. Sažetak / Abstract

Predviđanje kodirajućih regija u genomu metodama digitalne obrade signala

Činjenica da se proteini kodiraju iz aminokiselina formiranih prema kombinaciji tri uzastopna nukleotida, upućuju na mogućnost da kodirajući dijelovi gena pokazuju svojstvo periodičnosti s periodom 3. Kodirajuće eksone pokušano je na temelju ovog svojstva detektirati filtrima temeljenim na svepropusnom filtru, što je provjereno izračunom DFT. Istraživanje je pokazalo da je manji udio eksona uspješno detektiran i za njih je preciznost detektora vrlo visoka, dok je znatno veći dio eksona ostao nedetektiran. Zaključak je da svojstvo periodičnosti s 3 ne može biti samostalan kriterij detekcije eksona, ali ako postoje značajke (koje je tek potrebno otkriti) koje vrijede za druge eksone, tada bi u kombinaciji s njihovim detektorima bilo moguće istovremeno predvidjeti eksone različitih svojstava.

Ključne riječi: gen, genom, ekson, intron, period, detektor, nukleotid, kodirajući, nekodirajući, mjesto izrezivanja, filtar, DFT.

Detection of protein-coding regions in genome using digital signal processing methods

The well-known fact that proteins are generated from amino-acids based on codon-structure, leads to the possibility that the protein-coding regions of DNA exhibit a period-3 behavior. We tried to detect protein-coding regions using designed filter and DFT. The research showed that only small group of exons was successfully and very accurately detected, whereas the most of them were not detected. The conclusion is that the period-3 behavior in exons can't be used as detector itself, but if there were some other features for other exons (which are yet to be found), then the combination of those detectors could at the same time detect exons with different features.

Key words: gene, genome. exon, intron, period, detector, nucleotide, coding, non-coding, splice-site, filter, DFT.

Privitak

Provjera frekvencija nukleotida po pozicijama za točne vektore iz donora i akceptora

```
% odabir donori/ akceptori

% tocni = tocni_acc;
% br_redova = 64838;

  tocni = tocni_don;
  br_redova =64453;

lok1 = zeros(4,398);

for red = 1:br_redova
    niz = tocni(red,:);
    for stup = 1:398 % br_stupaca
        if     niz(stup)=='A'
            lok1(1,stup) = lok1(1,stup) + 1;
        elseif niz(stup)=='C'
            lok1(2,stup) = lok1(2,stup) + 1;
        elseif niz(stup)=='G'
            lok1(3,stup) = lok1(3,stup) + 1;
        elseif niz(stup)=='T'
            lok1(4,stup) = lok1(4,stup) + 1;
        end
    end
end

frekv = lok1./br_redova;

% save frekv_acc lok1 frekv
  save frekv_don lok1 frekv

% ispis najčešćih frekvencija u datoteku
fid = fopen('Frekvencije_akceptora.txt','w');  % promijeniti
                                                 donori/akceptori
fprintf(fid,'Frekvencije najcescih nukleotida za tocne vektore u
             akceptorima\n\n');
for j=1:398
    for i=1:4
        if frekv(i,j)>0.5
            if i==1 nukleotid='adenin';
            elseif i==2 nukleotid='citozin';
            elseif i==3 nukleotid='guanin';
            elseif i==4 nukleotid='timin';
            end
            fprintf(fid,'Na poziciji %3d se s frekvencijom %f javlja %s.
                        \n',j, frekv(i,j),nukleotid);
        end
    end
end
fclose(fid);
```

Detekcija filtrima i DFT-om za pojedinačne gene

```
clear all

%% odabir gena iz baze gena
% pozicionirati se u folder samoCelegans!
% definirati varijablu baza koja sadrži popis svih datoteka u folderu:
% baza ={'AfA8D5050.txt' 'AfA8D5065c.txt' 'AH96.txt' .. 'ZK11288a1.txt'};

datoteka='AH96.txt';
rb = find(ismember(baza,datoteka)==1);  % rb = redni broj datoteke u bazi
fid = fopen(baza{rb},'r');
gen = fread(fid,'*char');
len = length(gen);

nizA = zeros(1,len);
nizC = zeros(1,len);
nizG = zeros(1,len);
nizT = zeros(1,len);

for i=1:len
    nizA(i) = (gen(i)=='A' || gen(i)=='a') ;
    nizC(i) = (gen(i)=='C' || gen(i)=='c') ;
    nizG(i) = (gen(i)=='G' || gen(i)=='g') ;
    nizT(i) = (gen(i)=='T' || gen(i)=='t') ;
end

%% opis notch i antinotch filtara

Omega0 = 2*pi/3;
R = 0.99;   % Ovo je parametar s kojim se može eksperimentirati
Theta = acos( (1+R^2)/2/R * cos(Omega0) );

nazivnik  = [1   -2*R*cos(Theta) R^2];          % za sve filtre isti
A_brojnik = [R^2 -2*R*cos(Theta) 1  ];          % all-pass
G_brojnik = (1+R^2)/2*[1 -2*cos(Omega0) 1];     % notch
H_brojnik = (1-R^2)  *[1      0        -1];     % antinotch

% FFT 2*pi/3 filtar, uz N=3
h = [(-1-sqrt(-3))/2 1 (-1+sqrt(-3))/2];        % nekauzalni oblik

%% prolazak signala kroz detektor

rezA = zeros(1,length(gen));
rezC = zeros(1,length(gen));
rezG = zeros(1,length(gen));
rezT = zeros(1,length(gen));
maks_period = zeros(1,len);

for i=1:4

    if (i==1) x = nizA;
    elseif (i==2) x = nizC;
    elseif (i==3) x = nizG;
    elseif (i==4) x = nizT;
    end

yh  = filter(H_brojnik, nazivnik, x);  % antinotch
yg  = filter(G_brojnik, nazivnik, x);  % notch
```

```
yhg = filter(G_brojnik, nazivnik, yh); % antinotch-notch

% Izračunavamo ovojnice odziva usrednjavanjem apsolutnih vrijednosti
M = 99;          % Ovo je parametar s kojim se može eksperimentirati
yh_env  = medfilt2(abs(yh), [1 M]);
yg_env  = medfilt2(abs(yg), [1 M]);
yhg_env = medfilt2(abs(yhg), [1 M]);

% komplement
minimalni = 0.1; % Ovo je parametar s kojim se može eksperimentirati
yg_envc  = max(yg_env) + minimalni - yg_env;
yhg_envc = max(yhg_env)+ minimalni - yhg_env;

% "Detektor" se vrlo brzo istitrava, jedino je utitravanje problematično
% Ideja: preokrenuti vrijeme (antikauzalni filtri)
yh_anti  = filter(H_brojnik, nazivnik, x(end:-1:1));   % antinotch
yg_anti  = filter(G_brojnik, nazivnik, x(end:-1:1));   % notch
yhg_anti = filter(G_brojnik, nazivnik, yh_anti);  % antinotch-notch

yg_anti  =  yg_anti(end:-1:1);
yh_anti  =  yh_anti(end:-1:1);
yhg_anti = yhg_anti(end:-1:1);

% Izračunavamo ovojnice odziva usrednjavanjem apsolutnih vrijednosti
yh_anti_env = medfilt2(abs(yh_anti), [1 M]);
yg_anti_env = medfilt2(abs(yg_anti), [1 M]);
yhg_anti_env = medfilt2(abs(yhg_anti), [1 M]);

% komplement
yg_anti_envc  = max(yg_anti_env )+ minimalni - yg_anti_env;
yhg_anti_envc = max(yhg_anti_env)+ minimalni - yhg_anti_env;

% Iscrtavamo odziv
num = yh_env .* yh_anti_env;
den1 = yhg_envc .* yhg_anti_envc;
figure, plot(num./den1), grid

if (i==1)
    title('Adenin');
    rezA = (num./den1);
    maxA = max(rezA);
elseif (i==2)
    title('Citozin');
    rezC = (num./den1);
    maxC = max(rezC);
elseif (i==3)
    title('Guanin');
    rezG = (num./den1);
    maxG = max(rezG);
elseif (i==4)
    title('Timin');
    rezT = (num./den1);
    maxT = max(rezT);
end
end
suma2 =  rezA.^2 + rezC.^2 + rezG.^2 + rezT.^2;    % suma kvadrata
maks = ceil(max(suma2));

for i=1:len
    pom = sort([rezA(i) rezC(i) rezG(i) rezT(i)]); % od min prema max
    maks_period(i) = pom(4);
```

```
end

%% računanje stvarnih pozicija u genu

rub = zeros(1,100);
len = length(gen);
stvarnirub = zeros(1,len);
br=0;

for i=1:(len-1)
    if (i==1) && (gen(i)>='A' && gen(i)<='T')
        br=br+1;
        rub(br)=i;
    end
    if (i~=1)
       if((gen(i)>='A'&&gen(i)<='T')&&((gen(i-1)>='a')&&(gen(i-1)<='t')))
       | ((gen(i)>='A'&&gen(i)<='T')&&((gen(i+1)>='a')&&(gen(i+1)<='t')))
            br=br+1;
            rub(br)=i;
        end
    end
    if (i==(len-1))
        if ((gen(i+1)>='A') && (gen(i+1)<='T'))
            br=br+1;
            rub(br)=i+1;
        end
    end
end

rub (br+1:end)=[];      % rub = sadrži pozicije početka i kraja eksona

for i=1:2:(br-1)
    slika_ruba([rub(i):rub(i+1)]) = maks;
end
slika_ruba([len+1:len+5])=0;

%% crtanje konačnih rezultata za filtre

figure, plot(suma2), grid                    % suma kvadrata
title(datoteka)
hold on,plot(slika_ruba,'r')

figure, plot(maks_period.^2), grid           % maksimum
title('maksimum')
hold on,plot(slika_ruba,'r')

%% DFT

% ekstrakcija eksona iz gena, eksoni su različitih duljina, pretpostavimo
da su max 2000

br_eksona = length(rub)/2;
eksoni = char(zeros(br_eksona,2000));
for i=1:2:length(rub)
    eksoni(i,:) = char(zeros(1,2000));
    eksoni(i,1:rub(i+1)-rub(i)+1) = gen(rub(i):rub(i+1));
end

% analiza svih eksona u jednom genu

for k=1:2:length(rub)
```

```
% učitati u varijablu niz ekson
gen1 = eksoni(k,1:rub(k+1)-rub(k)+1);
len = length(gen1);

% traži najbliži manji broj djeljiv s 3
for i=1:10
    if mod(len,3)==0 break
    else len = len-1;
    end
end
ekson = gen1(1:len);

nizA = (ekson=='A') ;
nizC = (ekson=='C') ;
nizG = (ekson=='G') ;
nizT = (ekson=='T') ;

fA = fft(nizA,len);
fC = fft(nizC,len);
fG = fft(nizG,len);
fT = fft(nizT,len);

% traženje pozicije na kojoj je koeficijent za 2*pi/3:
t = [1:len]; funk = cos(2*pi*t/3); ff = fft(funk); % stem(abs(ff))
poz2pi3 = find(abs(ff)>1);

% oèitati Fouriera za sve 4 baze i zapamtiti najveću u varijablu f
fourier = [abs(fA(poz2pi3(1))) abs(fC(poz2pi3(1)))
           abs(fG(poz2pi3(1))) abs(fT(poz2pi3(1)))];
maks = sort(fourier);
maksf = maks(4);
nukleotid_maksf = find(fourier==maksf);

if (nukleotid_maksf==1) f = fA; disp('adenin ima max Fourierov koef.
                                     na 2*pi/3')
elseif (nukleotid_maksf==2) f = fC; disp('citozin ima max Fourierov
                                          koef. na 2*pi/3')
elseif (nukleotid_maksf==3) f = fG; disp('guanin ima max Fourierov
                                          koef. na 2*pi/3')
elseif (nukleotid_maksf==4) f = fT; disp('timin ima max Fourierov
                                          koef. na 2*pi/3')
end

% maksimum za Fourierovu transformaciju po najcescem nukleotidu
% ucrtati crveno
fsort = sort(abs(f(2:len)));
fmaks = fsort(end);
fmaks_pozicija = find(abs(f)==fmaks);
samofmaks = zeros(1,len);
samofmaks(fmaks_pozicija) = fmaks;

% frekvencija 2pi/3 za najcesci nukleotid
% ucrtati zeleno
samof2pi3 = zeros(1,len);
samof2pi3(poz2pi3) = abs(f(poz2pi3));

% crtanje
f(1)=20;
figure, stem(abs(f))
hold
```

```
    stem(samofmaks,'r')
    stem(samof2pi3,'g')
    legend('Fourierovi koeficijenti','F. koef. max. amplitude','F. koef
           za 2*pi/3');
    title(['gen ', datoteka, ', ekson ', num2str((k+1)/2), '/',
           num2str(br_eksona)])

end
```

Provjera više perioda na genomu

```
% traži Fourierove koeficijente za svaki točni vektor za periode 4-9 za
% ACGT, a maksimum od ta 4 sprema u matricu tog perioda

  br_tocnih = 64838;    % za akceptore
% br_tocnih = 64453;    % za donore
matrica_perioda3 = zeros(br_tocnih,2);
matrica_perioda4 = zeros(br_tocnih,2);
matrica_perioda5 = zeros(br_tocnih,2);
matrica_perioda6 = zeros(br_tocnih,2);
matrica_perioda7 = zeros(br_tocnih,2);
matrica_perioda8 = zeros(br_tocnih,2);
matrica_perioda9 = zeros(br_tocnih,2);

for period=3:9
    for k=1:br_tocnih

        gen(1:398) = tocni_acc(k,1:398);    % akceptori
        % gen(1:398) = tocni_don(k,1:398);  % donori

        len1=60; len2=60;    % promatrana duljina lijevo i desno

        % traži najbliži broj djeljiv s periodom koji se promatra
        for i=1:10
            if mod(len1,period)==0 break
            else len1=len1-1;
            end
        end
        gen1 = gen(198-len1:198-1);
        % gen1 = gen(200-len1:200);

        for i=1:10
            if mod(len2,period)==0 break
            else len2=len2-1;
            end
        end
        gen2 = gen(200:200+len2-1);
        % gen2 = gen(203:203+len2);

         % traži poziciju na kojoj je koeficijent za traženi period
        t=[1:len1]; f = cos(2*pi*t/period); ff = fft(f,len1)/len1;
        poz12=find(abs(ff)>0.2);
        poz1=poz12(1);

        t=[1:len2]; f = cos(2*pi*t/period); ff = fft(f,len2)/len2;
        poz12=find(abs(ff)>0.2);
        poz2=poz12(1);

        % lijeva strana vektora

        nizA1 = zeros(1,len1);
        nizC1 = zeros(1,len1);
        nizG1 = zeros(1,len1);
        nizT1 = zeros(1,len1);

        for i=1:len1
```

```
            nizA1(i) = (gen1(i)=='A') ;
            nizC1(i) = (gen1(i)=='C') ;
            nizG1(i) = (gen1(i)=='G') ;
            nizT1(i) = (gen1(i)=='T') ;
        end

        fA = fft(nizA1,len1)/len1;
        fC = fft(nizC1,len1)/len1;
        fG = fft(nizG1,len1)/len1;
        fT = fft(nizT1,len1)/len1;

        maks = sort([abs(fA(poz1)) abs(fC(poz1)) abs(fG(poz1))
                     abs(fT(poz1))]);
        M1=maks(4);

         % desna strana vektora

        nizA2 = zeros(1,len2);
        nizC2 = zeros(1,len2);
        nizG2 = zeros(1,len2);
        nizT2 = zeros(1,len2);

        for i=1:len2
            nizA2(i) = (gen2(i)=='A') ;
            nizC2(i) = (gen2(i)=='C') ;
            nizG2(i) = (gen2(i)=='G') ;
            nizT2(i) = (gen2(i)=='T') ;
        end

        fA = fft(nizA2,len2)/len2;
        fC = fft(nizC2,len2)/len2;
        fG = fft(nizG2,len2)/len2;
        fT = fft(nizT2,len2)/len2;

        maks = sort([abs(fA(poz2)) abs(fC(poz2)) abs(fG(poz2))
                     abs(fT(poz2))]);
        M2=maks(4);

        % korekcija zbog različitog broja len1 i len2, jer je dft=1/N*...
        korekcija=len1/len2;

        if period == 3       matrica_perioda3(k,1:2) = [M1,korekcija*M2];
        elseif period == 4   matrica_perioda4(k,1:2) = [M1,korekcija*M2];
        elseif period == 5   matrica_perioda5(k,1:2) = [M1,korekcija*M2];
        elseif period == 6   matrica_perioda6(k,1:2) = [M1,korekcija*M2];
        elseif period == 7   matrica_perioda7(k,1:2) = [M1,korekcija*M2];
        elseif period == 8   matrica_perioda8(k,1:2) = [M1,korekcija*M2];
        elseif period == 9   matrica_perioda9(k,1:2) = [M1,korekcija*M2];
        end

    end
end
```