SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Računalne metode u istraživanju nekodirajuće RNA

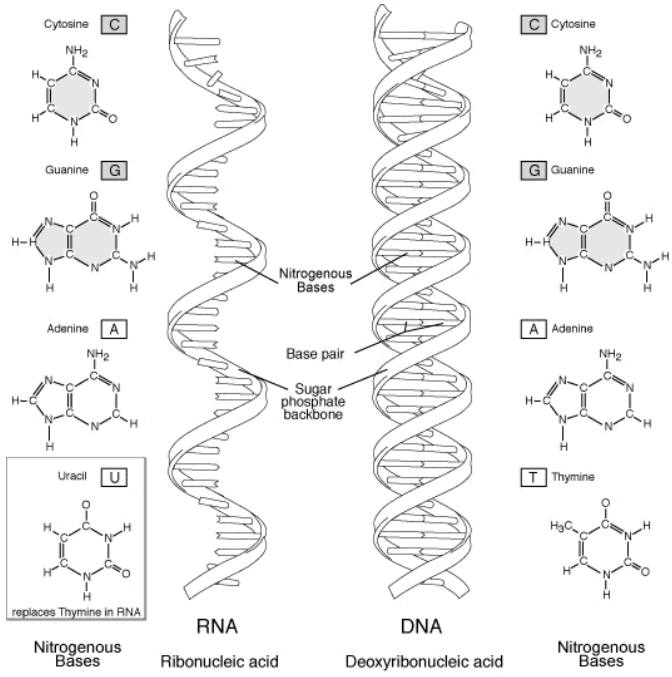
Mateja Dokleja

Voditelj: *Krešimir Šikić*

Zagreb, svibanj, 2007.

Sadržaj

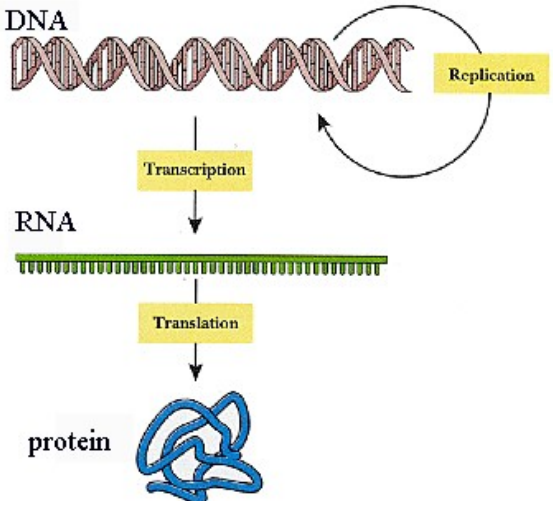
1. Uvod



Slika 1.) Struktura molekula RNA i DNA

Molekula RNA je nukleinska kiselina, polimer sastavljen od monomera koje nazivamo nukleotidima. Postoje 4 vrste nukleotida: adenin (A), guanin (G), uracil (U) i citozin (C), a bitno svojstvo je da su adenin i uracil te guanin i citozin komplementarni parovi nukleotida što znači da je između njih moguće povezivanje vodikovim vezama. Razlika između molekule RNA i DNA je u tome što se kod DNA umjesto uracila nalazi nukleotid timin, koji je sličnog kemijskog sastava kao uracil te u tome što je DNA dvolančana molekula dok RNA posjeduje samo jedan lanac. Također, RNA je sintetizirana pomoću enzima RNA polimeraza, koristeći molekulu DNA kao uzorak.

Centralna dogma molekularne biologije glasi: DNA sadrži informaciju, RNA prenosi informaciju zapisanu u DNA, a proteini su odgovorni za većinu bioloških aktivnosti i njihova sinteza je osnovna stanična funkcija. Prema ovome, molekula RNA služi samo kao poveznica između DNA i proteina, sredstvo na kojem se prenose informacije koje su zapisane na DNA. Ta tvrdnja je točna, ali samo za mRNA (engl. *messengerRNA*). Naime, podvrsta molekule RNA su nekodirajuće RNA ili ncRNA čija je osnovna karakteristika ta da su to molekule RNA koje se ne prevode u protein. No, to ni na koji način ne znači da one nemaju vrlo važne funkcije u organizmu. Upravo suprotno, one objedinjavaju čitav niz bitnih funkcionalnosti kao što su kontrola transkripcije i translacije, obrada i modifikacije RNA ili degradacija i translokacija proteina. Klasični primjer ncRNA su tRNA (engl. *transferRNA*) i rRNA (engl. *ribosomalRNA*), ali postoje i mnoge druge vrste koje se dijele na dvije osnovne kategorije, prva su male RNA molekule, npr. snoRNA, microRNA, siRNA. Druga kategorija uključuje duge RNA molekule kao što su Xist, Evf ili Air.



Slika 2.) Centralna dogma molekularne biologije

Iako je danas poznato da su ncRNA vrlo bitne molekule, njihovo proučavanje je doživjelo ekspanziju tek u posljednje vrijeme jer se dugo vremena mislilo, uglavnom zbog nerazumijevanja složenih procesa u kojima one sudjeluju, da su ncRNA neka vrsta genetičkog otpada koji nije bitan za procese u stanici. Ali, pokazalo se da je to bilo potpuno krivo uvjerenje i danas imamo zabilježen velik rast broja poznatih funkcija ncRNA. Tako je npr. molekula tmRNA koja je za sad pronađena samo u bakterijama kombinacija tRNA i mRNA koja djeluje kad iz nekog razloga dođe do problema u prevođenju mRNA. Da bi se spriječio nastanak nedovršenog proteina koji je potencijalno opasan za stanicu tmRNA generira svojevrsnu etiketu koja se pričvrsti na C-završetak nedovršenog proteina koja djeluje tako da se protein uništi.

Točan broj ncRNA kodiranih u genomu je nepoznat, ali analize predviđaju da postoji oko 30 000 dugih nekodirajućih RNA te barem toliko malih nadzornih ncRNA samo u genomu miša. Procjene broja ncRNA u ljudskog genomu sugeriraju brojku reda veličine 10^5. Također se vjeruje da postoje još mnoge ncRNA koje su još neotkrivene, ali za njihovo ispitivanje nije moguće osloniti se samo na eksperimentalne metode jer je količina genetičkih informacija danas prevelika te je nužno upotrijebiti i računalne metode da bi se došlo do željenih rezultata.

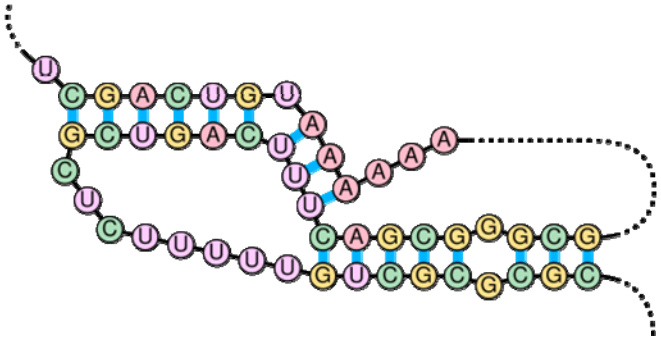
2. Seminarski rad

2.1. Sekundarna struktura molekule RNA

Kao što je prije spomenuto, u molekuli DNA se komplementarne baze vežu vodikovim vezama između 2 lanca te tako omogućuju specifičnu strukturu te molekule. Do iste pojave dolazi i kod RNA molekule, ali kako ona raspolaže samo s jednim lancem, dolazi do njegovog uvijanja. Takva dvodimnezionalna struktura koja je nastala kao rezultat uparivanja baza naziva se sekundarnom strukturom molekule RNA.

Postoji nekoliko karakterističnih načina uvijanja RNA. Kao najrasprostranjeniji se pojavljuju oblik ukosnice (uključuje ugniježđene baze) i pseudočvora (karakteriziraju ga križne interakcije) te, u nešto manjoj mjeri, izbočina i internih petlji. Struktura ukosnice je struktura do koje dolazi uglavnom kod palindromskih sekvenci kada se komplementarne baze povežu i formiraju dvostruki lanac s omčom na kraju. Ako imamo jedan par nukleotida na lokacijama *i* i *j* te jedan na lokacijama *k* i *l* (*i*<*j*, *k*<*l*) parovi su ugniježđeni ako je zadovoljen uvjet *i*<*k*<*l*<*j* ili *k*<*i*<*j*<*l*. S druge strane, pseudočvor je nešto kompliciranija struktura kod koje dolazi do spajanje dvije ukosnice i to tako da omča (engl. *stem*) prve formira dio peteljke (engl. *loop)* druge ukosnice. Ovakva struktura nije dobro ugniježđena te zadovoljava uvjet *i*<*k*<*j*<*l* ili *k*<*i*<*l*<*j* pod pretpostavkom da se parovi baza nalaze na lokacijama *i* i *j* te *k* i *l* (*i*<*j*,k<l).

Slika 3a.) Struktura ukosnice



Slika 3b.) Pseudočvor

Sekundarna struktura molekule RNA računalno se predviđa računanjem strukture s minimalnom slobodnom energijom (MFE) za različite kombinacije vodkovih veza i domena, tako se traže optimalni parovi baza. Na internetu se to može učiniti korištenjem aplikacija MFOLD ili RNAfold. Važno je napomenuti da takve aplikacije, bazirane na standardnoj metodi dinamičkog programiranja ne mogu identificirati pseudočvorove te je za njihovu detekciju potrebno pribjeći kompleksnijim metodama. Važnost sekundarne strukture se očituje u tome što je kod ncRNA struktura same molekule obično vrlo blisko povezana sa njezinom funkcijom te ju je esencijalno dobro poznavati da bi se moglo proučavati njezinu funkcionalnost.

2.2. Traženje homolognih sekvenci

Kod analize nekodirajuće RNA kao bitan faktor pojavljuje se homologija između dvije molekule. Dvije sekvence su homologne ako imaju zajedničko porijeklo i tada se mogu grupirati u homologne porodice. To svojstvo je bitno jer porodice često dijele određen broj zajedničkih karakteristika te pripadnost određene sekvence nekoj porodici može uvelike pomoći pri identifikaciji njezinih funkcija.

Kod pronalaženja homolognih sekvenci molekula DNA i proteina dobre rezultate postižu metode koje su bazirane na sličnosti sekvenci. Takve metode se baziraju na pronalaženju sličnosti poravnavanjem dviju sekvenci (BLAST, FASTA), traženju uzoraka koji se često pojavljuju (PROSITE) te vjerojatnosnom opisu cijele porodice sekvenci (profil-HMM). Iako one dobro funkcioniraju u navedenim slučajevima, te metode gube na efikasnosti u slučaju ncRNA jer je kod ncRNA sekundarna struktura katkad očuvana iako ne postoje prevelike sličnosti u primarnoj. Do te pojave dolazi zbog svojstva RNA molekula da često koriste kompenzacijske mutacije da bi očuvale svoju sekundarnu strukturu tj. kada dođe do mutacije jedne baze, njezine komplementarna baza koja je povezana s njom vodikom vezama također se promijeni u novu komplementarnu bazu da ne bi došlo do pucanja veze.To nas dovodi do zaključka da je potrebno uzeti u obzir i sekvencijalnu i strukturalnu sličnost dvije sekvence ncRNA, ali postavlja se pitanje kako kombinirati udjele tih dviju komponenti da bi se dobili korisni rezultati. U tom slučaju standardni alati bazirani na temelju sličnosti sekvenci podbacuju jer uzimaju u obzir samo primarnu strukturu.

2.3. Kontekstno neovisna gramatika

Jedan od načina na koji se može realizirati analiza primarne i sekundardne strukture je korištenjem transformativne gramatike, samo je potrebno odlučiliti koja će biti najpogodnija. Transformativna gramatika je set pravila koja su korišteni da bi se generirao niz znakova nad danom abecedom. Sastoji se od završnih znakova (koji su znakovi konačnog niza) , nezavršnih znakova (koji se koriste za definiranje produkcija) i produkcija oblika $\alpha \rightarrow \beta$ gdje su α i β nizovi završnih i nezavršnih znakova. Dakako, najbolje je izabrati najjednostavniju gramatiku koja odgovara zadanim uvjetima. Kod spajanja baza u sekundarnoj strukturi dolazi do porasta broja simetričnih regija u primarnoj sekvenci koje su analogne palindromima (simetrične sekvence koje su iste čitajući sprijeda ili odzada). Najjednostavnija gramatika prema Chomskyjevoj hijerarhiji, regularna gramatika, nije sposobna dovoljno dobro opisati palidromski jezik te ćemo koristiti nešto složeniju, kontekstno neovisnu gramatiku . Ona zadovoljava postavljene uvjete jer posjeduje produkcije koje omogućuju emitiranja jednog znaka na krajnje lijevu, a drugog na krajnje desnu poziciju desne strane produkcije te je zbog toga efikasna kod opisivanja ugniježđenih korelacija. Tako produkcije sljedeće gramatike omogućuju generiranje sekundarne strukture RNA koja se sastoji od proizvoljnog broja peteljki i omči koje su koje su također proizvoljne duljine.

$S \rightarrow SS$

$S \rightarrow \varepsilon$

S→aSu|uSa|cSg|gSu

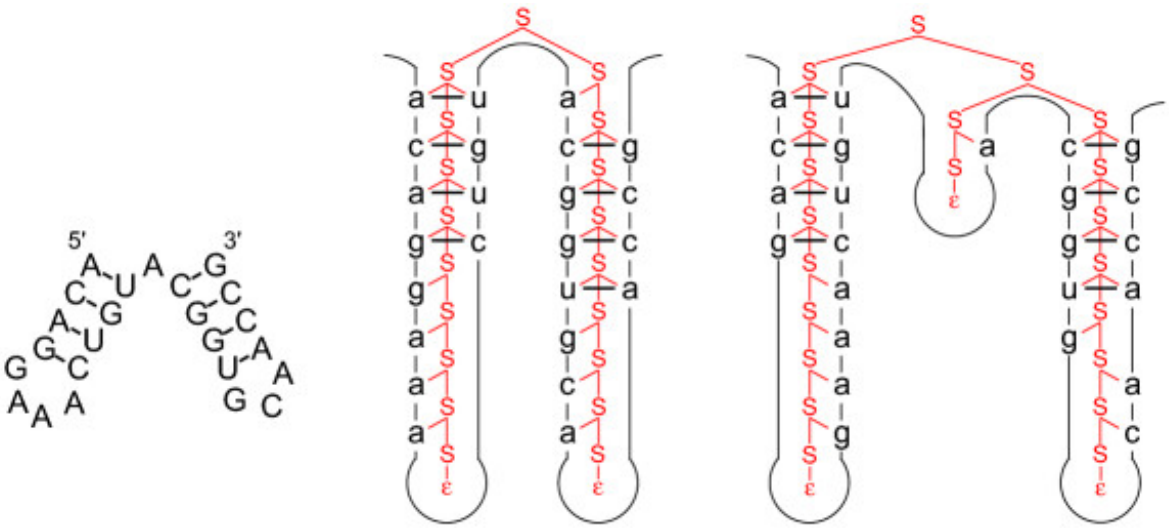
S→ aS|uS|cS|gS

S→Su|Sa|Sg|Su

Važno je napomenuti da su različite produkcije odvojene znakom | te da su nezavršni znakovi označeni velikim slovima, a završni malim.

Postupak generiranja jedne sekvence koja se sastoji od dvije peteljke sa omčama možemo prikazati postupnim korištenjem tih produkcija s time da je nezavršni znak na koji se produkcija trenutno primjenjuje podcrtan:
S→S̲S→aS̲uS→acS̲guS→ acaS̲uguS→ acagS̲cuguS→ acaggS̲cuguS→ acaggaS̲cuguS→ acagaagS̲cuguS→ acaggaaaS̲cuguS→ acaggaaacuguS̲→ acaggaaacuguaS̲→ acaggaaacuguaS̲→ acaggaaacuguacS̲g→ acaggaaacuguacgS̲cg→ acaggaaacuguacggS̲ccg→ acaggaaacuguacggS̲ccg→ acaggaaacuguacgguS̲accg→ acaggaaacuguacggugS̲accg→ acaggaaacuguacggugcS̲accg→ acaggaaacuguacggugcaS̲accg→ acaggaaacuguacggugcaaccg.

Sekvencu generiranu kontekstno neovisnom gramatikom moguće je prikazati grafom specifičnog oblika koji se naziva stablo parsiranja. Struktura tog grafa uključuje čvorove i listove (vanjske čvorove) s time da su unutarnji čvorovi označeni nezavršnim znakovima gramatike, a vanjski završnim znakovima.



Slika 4.) Dva moguća stabla parsiranja izgrađena za dani niz baza *acaggaaacuguacggugcaaccg*

Kontekstno neovisne gramatike dobro modeliraju sekundarne strukture te su korisne kod aplikacija koje pretražuju uzorke i mogu efikasno odrediti da li određena RNA sekvenca zadovoljava uvjete pretraživanja. Međutim, one nisu efikasne kod predviđanja sekundarne strukture jer je za svaku sekvencu moguće izgraditi više valjanih stabla parsiranja od kojih svako predstavlja jednu strukturu. Naš cilj je ocjeniti i rangirati sva moguća stabla za neku sekvencu kako bi pronašli optimalno. Za izvršenje tog zadatka nije dovoljna obična kontekstno neovisna gramatika već njezina proširena verzija, stohastička kontekstno neovisna gramatika.

2.4. Stohastička kontekstno neovisna gramatika

Stohastička kontekstno neovisna gramatika (engl. *stochastic context-free grammar* (SCFG)) je kontekstno neovisna gramatika u kojoj je svakoj produkciji pridodana vjerojatnost. Vjerojatnost nekog postupka generiranja međuniza jest tad umnožak vjerojatnosti svih produkcija korištenih u postupku, te su stoga neki postupci generiranja konzistentniji sa stohastičkom gramatikom od ostalih. Suma vjerojatnosti produkcija za bilo koji završni znak mora biti jedan. Oznaka Θ se koristi za potpun skup vjerojatnosti. Vjerojatnost $P($**x**, π|**H** , Θ) je produkt svih vjerojatnosti produkcija koje su korištene u stablu parsiranja π izgrađenom za sekvencu **x**. SCFG je vjerojatnosni model koji opisuje udruženu vjerojatnosnu distribuciju $P($**x**, π|**H** , Θ) nad svim RNA sekvencama **x** i svim mogućim stablima parsiranja π.

Raspolažući s parametriziranom SCFG $($**H**, Θ) i sekvencom **x** Cocke-Younger-Kasami (CYK) algoritam pronalazi optimalno stablo parsiranja π (ono najveće vjerojatnosti) za sekvencu **x**, $\pi = \text{argmax}_{\pi} P($**x**, π|**H** , Θ). Za primjer gramatike iz prijašnjeg poglavlja CYK algoritam glasi:

Inicijalizacija: $\gamma(i, i-1) = \log p(S \to \varepsilon)$

Iteracija: $$\gamma(i,j) > \max \begin{cases} \gamma(i+1, j-1) + \log p(S \to x_i S x_j) \\ \gamma(i+1, j) + \log p(S \to x_i S) \\ \gamma(i, j-1) + \log p(S \to S x_j) \\ \max_{i=k=j} \{\gamma(i,k) + \gamma(k+1, j) + \log p(S \to SS)\} \end{cases}$$

Kada algoritam završi $\gamma(1, L)$ je jednak $\log P($**x**, π|**H** , Θ), logaritamskoj vjerojatnosti najvjerojatnijeg stabla parsiranja π za sekvencu **x,** gramatiku **H** i parametre Θ.

CYK algoritam je u biti isti kao postojeći algoritam koji predviđa uvijanje RNA (engl. *RNA folding*), ali postoje određene razlike. Kod algoritama uvijanja RNA se koriste parametri za minimizaciju energije koji su većinom izvedeni iz eksperimentalnih proučavanja taljenja malih modela struktura dok su, s druge strane, SCFG logaritamske vjerojatnosti izvedene iz frekvencija promatranih u uzorcima poznatih RNA sekundarnih struktura. Tako se, umjesto bodovanja G-C para naslaganog na C-G para baza tako da se doda član koji označuje kontribuciju slobodne energije strukture GC/CG, kod SCFG dodaje logaritamska vjerojatnost koju su strukture GC/CG pokazale u promatranjima poznatih RNA struktura.

2.5. Kovarijacijski model (CM)

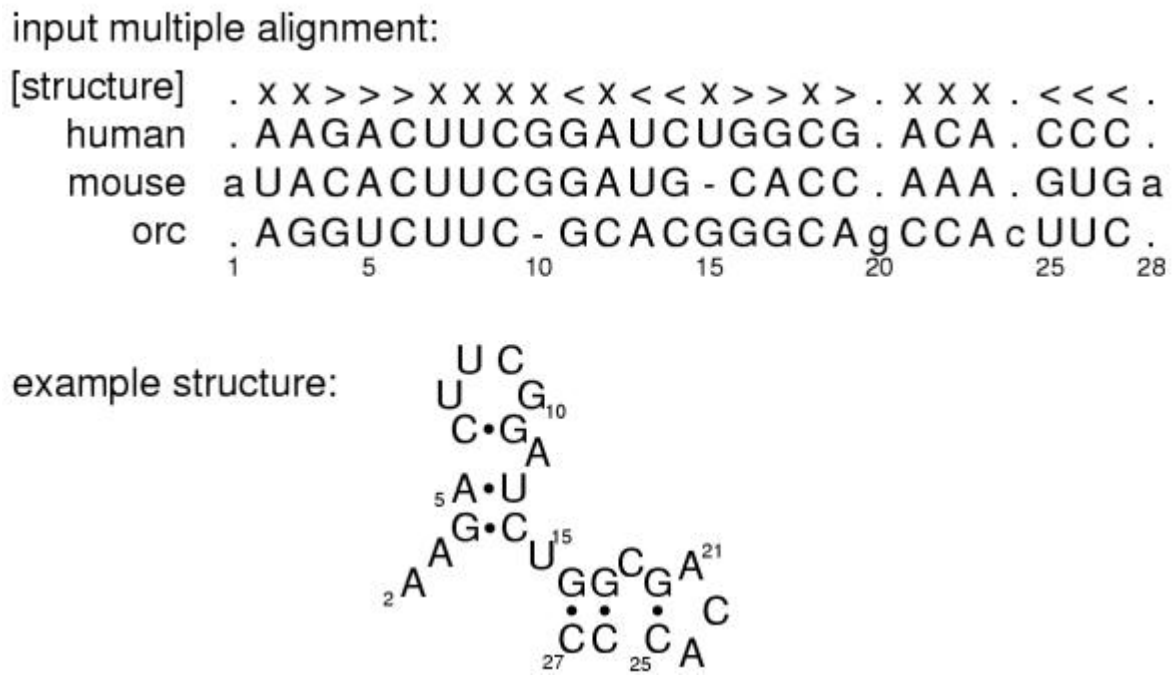
Stablo parsiranja, nastalo primjenom kontekstno neovisne gramatike, nije fleksibilno kod prikazivanja strukture molekule RNA. Da bi se prikazala porodica molekula RNA treba se omogućiti dodavanje, brisanje i neslaganje. Za tu svrhu koristi se kovarijacijski model koji je repetitivna SCFG struktura. Kovarijacijski model se sastoji od 7 tipova stanja i produkcija.

Tablica 1.) Stanja i produkcije kod kovarijacijskog modela

Tip stanja	**Opis**	**Produkcija**	**Emisija**	**Prijelaz**
P	(emisija para)	$P \rightarrow aYb$	$e_v(a,b)$	$t_v(Y)$
L	(lijeva emisija)	$L \rightarrow aY$	$e_v(a)$	$t_v(Y)$
R	(desna emisija)	$R \rightarrow Ya$	$e_v(a)$	$t_v(Y)$
B	(račvanje)	$B \rightarrow SS$	1	1
D	(obriši)	$D \rightarrow Y$	1	$t_v(Y)$
S	(kreni)	$S \rightarrow Y$	1	$t_v(Y)$
E	(kraj)	$E \rightarrow \varepsilon$	1	1

Vjerojatnost svake produkcije je umnožak vjerojatnosti emisije e_v i vjerojatnosti prijelaza t_v , a to su parametri koji ovise o poziciji, tj. o stanju *v.* Tako npr. stanje P producira dva korelirana simbola a i b (koji predstavljaju jedan od 16 mogućih parova baza) sa vjerojatnošću e_v i prelazi u jedno od od mogućih novih stanja *Y* s vjerojatnošću $t_v(Y)$.

Početni korak u stvaranju CM modela je provođenje višestrukog sravnjenja homolognih RNA te izdvajanje linije koja opisuje konsenzusnu strukturu. Taj postupak je pokazan na sljedećoj slici:



Slika 5.) Primjer porodice RNA sekvenci

Na vrhu slike su prikazane 3 sekvence, višestruko sravnjene sa ukupnim brojem od 28 stupaca od čega su 24 iskorištena da bi se prikazale konsenzusne pozicije. Linija

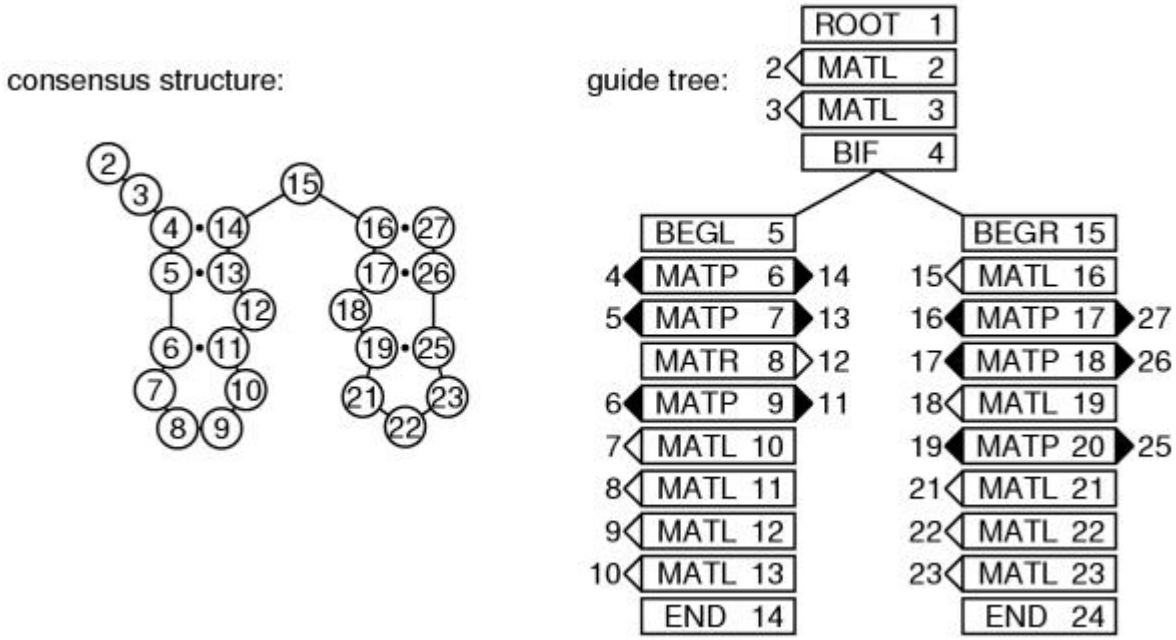
koja je označena sa *[structure]* označava konsenzusnu sekundarnu strukturu tako da simboli < i > označavaju parove baza, x označava konsenzusne nesparene pozicije baza, a . označava kolone dodavanja koje nisu uzete u obzir kod konsenzusne strukture. Na dnu slike je prikazana sekundarna struktura ljudske sekvence.

Sam kovarijacijski model je realiziran konstrukcijom usmjerenog grafa oblika stabla repetitivnim korištenjem osnovnih blokova koja zovemo CM čvorovima. Takvo usmjereno stablo je u biti stablo parsiranja za konsenzusnu strukturu, sa čvorovima kao nezavršnim znakovima i kolonama sravnjenja kao završnim. Postoji 8 vrsta čvorova koja su prikazani sljedećom tablicom:

Tablica 2.) CM čvorovi i pripadajuća stanja

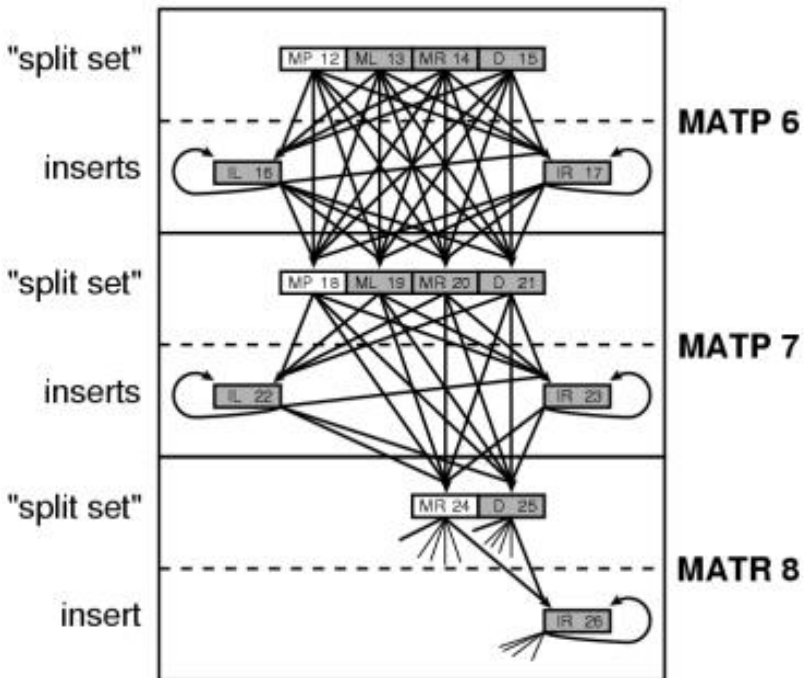
Čvor	**Opis čvora**	**Glavno stanje**	**Stanja**	**Ukupan broj stanja**	**Broj stanja podjele**	**Broj stanja umetanja**
MATP	(par)	P	[MP ML MR D] IL IR	6	4	2
MATL	(nesparena baza, lijevo)	L	[ML D] IL	3	2	1
MATR	(nesparena baza, desno)	R	[MR D] IR	3	2	1
BIF	(račvanje)	B	[B]	1	1	0
ROOT	(korijen)	S	[S] IL IR	3	1	2
BEGL	(početak, lijevo)	S	[S]	1	1	0
BEGR	(početak, desno)	S	[S] IL	2	1	1
END	(kreni)	E	[E]	1	1	0

Primjećujemo da su čvorovi MATP, MATL i MATR emitivni, tj. produciraju završne znakove (u ovom slučaju oznake baza) te su oni povezani s konsenzusnim kolonama kod višestrukog sravnjenja. Neemitivni čvorovi BIF, ROOT, BEGL, END služe da bi se formirala struktura stabla za emitivne čvorove. Pretvorba konsenzusne strukture u usmjereno stablo prikazana je slikom 6.



Slika 6.) Pretvorba konsenzusne strukture u usmjereno stablo

Lijevo na slici 6 primjećujemo konsenzusnu sekundarnu strukturu koja potječe iz podataka prikazanih slikom 5. Broj u krugu označava kolonu kod višestrukog sravnjenja. Na desnoj strani slike 6 nalazi se usmjereno stablo koje odgovara konsenzusnoj strukturi na lijevoj strani. Čvorovi su označeni s brojkama od 1 do 24 te su povezani sa kolonama koje generiraju. Parovi baza su pridruženi MATP čvorovima, a nesparene baze MATL i MATR čvorovima. Čvor ROOT je korišten za početak (korijen stabla), a BIF, BEGL i BEGR čvorovi služe za realizaciju grananja i to BIF za račvanje, a BEGL i BEGR za početak grane na lijevoj odnosno desnoj strani.

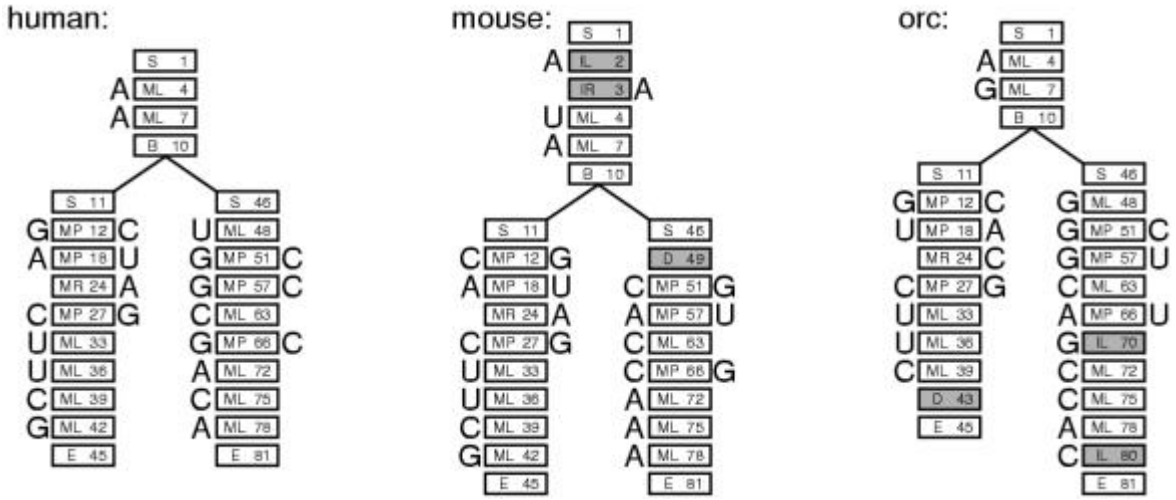


Slika 7.) Primjer interne strukture nekih čvorova

Nužno je da se omogući da sekvenca u bazi podataka ima simbole koji nisu dio konsenzusne strukture (umetanje) i da nema simbole koji su uključeni u tu strukturu (brisanje). Tu mogućnost pružaju interne strukture čvorova. Naime, svaki čvor se sastoji od skupa stanja kao što je to prikazano u tablici 2. Potrebno je primjetiti da su za interna stanja jednaka stanjima prikazanim u tablici 1. osim što su stanjima P, L i R dodatni prefiksi M

odnosno I koji označavaju slaganje odnosno dodavanje. Ta interna stanja su podijeljena u dva dijela, prvi dio čine stanja podjele slaganjem ili brisanjem simbola ili para simbola u konsenzusnoj strukturi dok drugi dio čine stanja dodavanja dodatnih simbola između simbola konsenzusne strukture i stabla ispod čvora. Tako npr. čvor MATL ima tri interna stanja: jedno za slaganje s konsenzusnim simbolom, jedno za brisanje konsenzusnog simbola i jedno za dodavanje dodatnih simbola između konsenzusnog i daljnje sekvence koja se gradi u desno.

Nakon izgradnje usmjerenog stabla koje se, kad su sve gore navedene karakteristike ispunjene, naziva CM možemo sravniti danu sekvencu prema izgrađenom CM-u i tada dobijemo stablo parsiranja za tu sekvencu.



Slika 8.) Primjeri stabala parsiranja

Na slici 8. su prikazana stabla parsiranja za naše ulazne sekvence čovjeka, miša i orke sa slike 5. Sivom bojom su prikazana dodavanja i brisanja koja se ne slažu s konsenzusnom strukturom.

Slično tome, svaku sekvencu možemo sravniti s danim CM-om i brojiti događaje emisije i prijelaza u svakom stanju CM-a da bi izračunali vjerojatnosti emisije i prijelaza. Tada jednostavno možemo koristiti te frekvencije i pokrenuti EM algoritam poznat kao unutarnje-vanjski algoritam za optimizaciju parametara modela.

2.6. Identifikacija novih nekodirajućih RNA

Izgradnja pretraživača gena opće namjene za predviđanje novih ncRNA molekula je mnogo teži zadatak od predviđanja homolognih ncRNA gena i sekvencama genoma. Do sad su mnoge metode obrade signala bile korištene za predviđanje gena koji kodiraju proteine od kojih su najvažnije diskretna Fourierova transformacija, digitalni filtri i HMM. Između njih, metode koje se baziraju na HMM-u su se pokazale osobito uspješnima te se mogu pohvaliti s uspješnošću predviđanja većom od 90% i gotovo savršenom kod jednostavnih organizama. Ali, te metode nisu pogodne za predviđanje ncRNA gena iz više razloga. Naime, za razliku od gena koji kodiraju proteine te koji se mogu podijeliti u kodone (skupove od tri nukleotida koji kodiraju jednu aminokiselinu), kod gena koji kodiraju ncRNA ne postoje takve vrste struktura. Nadalje, u ovom slučaju ne postoje tzv. ORF regije (regija koja potencijalno kodira protein, počinje s specifičnim početnim kodonom i završava s specifičnim završnim) koje se koriste kod pronalaženja kodirajućih gena jer one jasno impliciraju postojanje kodirajućeg gena. Također je i problem u tome što su mnoge ncRNA kraće nego kodirajući geni te tipična ncRNA ima manje od nekoliko stotina nukleotida. Unatoč neefikasnosti tradicionalnih alata za pronalaženje kodirajućih gena, moguće je iskoristiti izvorne karakteristike RNA za izgradnju alata za pronalazak novih ncRNA. Na primjer, mnoge ncRNA imaju dobro očuvano sekundarnu strukturu te možemo iskoristiti to svojstvo za pronalazak ncRNA gena.

Većina alata za pronalaženje ncRNA gena kao što su QRNA, ddbRNA, MSARI i RNAz koristi jednu općenitu strategiju. Prvo pronalaze regije u genomu koje su očuvane u različitim vrstama i provode višestruko poravnavanje sekvenci između tih regija. Ovisno o poravnanju oni istražuju da li postoji neka općenita sekundarna struktura koja je sačuvana u svim sekvencama. Ta informacija je iskorištena da bi se otkrilo da li te regije odgovaraju funkcionalnoj RNA. Neki od opisanih algoritama su korišteni za pretraživanje genoma nekoliko organizama i rezultate pokazuju da je ta strategija prilično efektivna. Tako npr. RNAz, koji je trenutno najsuvremeniji algoritam za predviđanje novih ncRNA, postiže prosječnu osjetljivost od 84.17% kod specifičnosti od 96.42% i 75.27% osjetljivosti kod 98.93% specifičnosti. Nedavno je RNAz korišten za izvođenje komparativnog proučavanja nekoliko genoma kralježnjaka i predvidio je više od 30 000 ncRNA gena u ljudskom genomu, a od njih je skoro tisuću gena pronađeno očuvano u sva 4 genoma kralježnjaka koji su bili proučavani što snažno sugerira njihovu biološku funkcionalnost.

Unatoč početnom uspjehu ovih alata za pronalaženje ncRNA gena, još uvijek je otvoren velik prostor za poboljšanja. Naime, prosječni postoci predivđanja za postojeće algoritme nisu toliko visoki i još uvijek ne rade dobro za određene skupine RNA. Ipak, učinkovitost tih alata se povećava se velikom brzinom i jasno je da će računalni alati za pronalaženje gena biti vrlo važni u pronalaženju novih ncRNA u budućnosti.

3. Zaključak

Istraživanje nekodirajuće RNA je relativno novo i neotkriveno područje u genetici jer, za razliku od molekule DNA, funkcionalnost ovih molekula je otkrivena relativno kasno te je samim time kasno započeo proces istraživanja nekodirajuće RNA. U usporedbi s genima koji kodiraju proteine i čija je anotacija skoro završena kod većine sekvencioniranih genoma, anotacija nekodirajuće RNA je tek započela što se vidi već iz samog podatka da je danas još nemoguće predvidjeti točan broj molekula ncRNA u genomu. Zbog toga su i metode koje se koriste pri istraživanju ncRNA još u razvoju. Mnoge metode koje se koriste su metode koje su već otprije poznate kod u području obrade signala te su prilagođene specifičnim zahtjevima koje imaju ovakva istraživanja. Primjer je SCFG koji se u originalu koristio za prepoznavanje govora.

Područje istraživanja ncRNA je danas područje koje bilježi veliku ekspanziju i karakteriziraju ga značajna otkrića novih metoda i algoritama u cilju što veće produktivnosti i preciznosti. Naime, dok su prve korištene metode mogle predvidjeti i uspoređivati samo linearne strukture, današnji modeli poput CM-a efikasno omogućuju predviđanje sekundarnih struktura koje uključuju i ugniježđenja. Još naprednija metoda koja se danas javlja, profil-csHMM i koja je svojevrsna ekstenzija klasičnog profil-HMM-a osim svega ovdje navedenog podržava i podupire i mogućnost ukrštavanja. Također, zamjetljivo je i stalno poboljšanje već postojećih metoda kako bi one postigle još veću uspješnost. Tako je već danas mogućnost predviđanja sekundarne strukture relativno jednostavnih SCGF-a vrlo bliska točnosti koju metode koje se baziraju na minimizaciji energije. Ali, na duge staze se radi na tome da se SCFG prilagodi usporedbi para sekvenci da bi se rješio problem strukturalnog sravnjenja kombinirajući strukturalne i evolucijske informacije u jednom modelu.

Unatoč ogromnom napretku na polju istraživanja nekodirajuće RNA jasno je da je ostalo još mnogo neotkrivenog, a sama mogućnost novih otkrića koja će u većoj ili manjoj mjeri utjecati na poimanje bioloških procesa te koja će definitivno biti još jedan komadić slagalice zvane ljudski genom je dovoljno intrigantna da se u bliskoj budućnosti još više energije i vremena te definitivno znanja utroši na nekodirajuću RNA.

4. Literatura

[1] Yoon B., Vaidyanthan P. P. Computational Identification and Analysis of Noncoding RNAs: Unearthing the buried treasures in the genome. IEEE Signal processing magazine. 1053-5888, 2007., str. 64-74

[2] Dowell R. D., Eddy S. R., Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction, 4.6.2004. http://www.biomedcentral.com/1471-2105/5/71, 8.5.2007.

[3] Eddy S. R., Durbin R. RNA sequence analysis using covariance models. Nucleic Acids Research. 2079-2088, broj 22 (1994.), str. 2079-2088

[4] Eddy S.R., A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure, 2.7.2002., http://www.biomedcentral.com/1471-2105/3/18, 10.5.2007

[5] Smith S. F., Acceleration of covariance models for non-coding RNA search, http://ww1.ucmss.com/books/LFS/CSREA2006/BIC4507.pdf , 11.5.2007.

[6] Non-coding RNA, http://en.wikipedia.org/wiki/Noncoding_RNA, 6.5.2007.

[7] Stohastic context-free grammar, http://en.wikipedia.org/wiki/SCFG, 7.5.2007.

[8] Pseudoknot, http://en.wikipedia.org/wiki/Pseudoknot, 7.5.2007.

[9] Stem-loop, http://en.wikipedia.org/wiki/Stem-loop, 7.5.2007.

5. Sažetak

Nekodirajuća RNA je specijalna vrsta molekule RNA koja ne kodira proteine već ima druge, vrlo važne funkcije u stanici. Kod molekule RNA bitan pojam je sekundarna struktura jer je obično ona blisko povezana sa funkcijom same molekule, a i molekula RNA pod svaku cijenu želi zadržati svoju sekundarnu strukturu, čak i na štetu primarne. Zbog toga se velika važnost pridaje predviđanje sekundarne strukture.

Traženje homolognih sekvenci je osnovni postupak pri istraživanju nekodirajuće RNA jer homologne sekvence imaju zajedničko porijeklo i određen broj zajedničkih karakteristika. Kontekstno neovisna gramatika se koristi pri generiranju nizova koji predstavljaju oznake baza RNA molekula te se na temelju tih generiranih nizova mogu izgraditi stabla parsiranja koja opisuju sekundarnu strukuturu. No, problem u tome je višeznačnost, tj. moguće je izgraditi više različitih stabala za određenu sekvencu. Stohastička kontekstno neovisna gramatika rješava taj problem tako da svakom stablu parsiranja izgrađenom za određenu sekvencu pridodaje vjerojatnost te se tada pomoću CYK algoritma pronalazi optimalno stablo, tj. optimalna struktura. Kovarijacijski model (CM) je posebna, repetitivna SCFG struktura koja omogućuje dodavanja, brisanja i neslaganja dviju ili više sekvenci. Ono se izgrađuje tako da se prvo provede višestruko sravnjenje sekvenci te pronađe konsenzusna struktura, zatim se gradi usmjereno stablo za konsenzusnu strukturu koje se naziva CM model. Naposljetku se dana sekvenca sravni s CM modelom i dobije se stablo parsiranja za tu sekvencu te konačno možemo izračunati vjerojatnosti i pronaći optimalnu strukturu.

Osim traženja homolognih sekvenci, radi se i na otkrivanju novih nekodirajućih RNA, ali to je teži zadatak zbog nedostatka svojstava koja posjeduju geni koji kodiraju proteine pa treba iskoristiti neka svojstva molekule RNA poput dobrog očuvanja sekundarne strukture.

Istraživanje nekodirajuće RNA je relativno novo područje u genetici koje je u stalnom rastu i razvoju te se u bliskoj budućnosti očekuju mnoga zanimljiva otkrića.