SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Računalne metode za detekciju gena

Vanessa Županović

Voditelj: *Mile Šikić*

Zagreb, travanj, 2011.

Sadržaj

1. Uvod

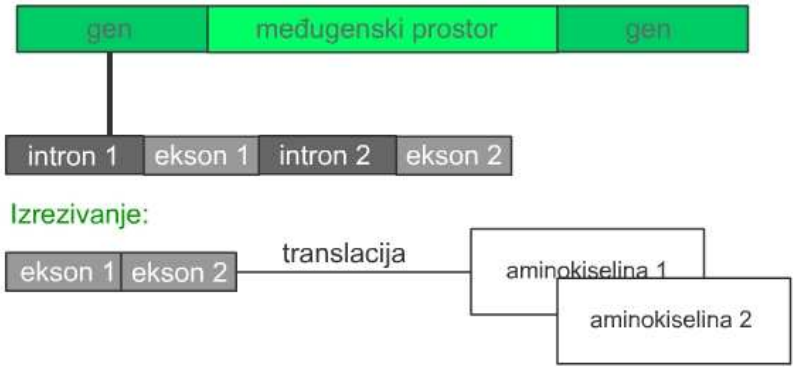
Računalne metode detekcije gena i ostalih funkcionalnih djelova genomske DNA karakterizira iznimno brz razvoj kroz proteklih dvadesetak godina. Broj cjelovito utvrđenih sekvenci genoma različitih organizama prelazi 800, no sama preciznost kojom geni bivaju detektirani još je uvijek poražavajuća. Na razini nukleotida otprilike oko 80% gena jest uspješno predviđeno (pod pojmom predviđanja gena podrazumjeva se utvrđivanje granica te položaja kodirajućih, odnosno nekodirajućih područja), na razini eksona tek 45%, ukoliko bi u obzir uzeli kompletan gen, uspješnost njegova predviđanja svela bi se na niskih 20%. Važno je naglasiti da trenutne metode pronalaska gena daju značajno bolje rezultate kod prokraiotskih organizama (nego kod eukariotskih) upravo zbog njihove jednostavne građe, tj. geni kod prokraitota ne sadrže intronske prekide što njihovu analizu čini mnogo jednostavnijom. Neki od najčešće rabljenih programa za detekciju gena jesu: FGENEH, GENMARK, GeneID, Genie, GeneParser, GENSCAN te GRAIL 2, no činjenica je da su mnogi od njih implementirani na način da za ulaznu sekvencu pretpostavljaju isključivo jedan, u potpunosti sekvencirani gen, stoga unošenje sekvence sačinjene od parcijalnih ili više gena, generalno rezultira izlazom koji nema nikakvog biloškog smisla. Kako bi uopće bilo moguće razumijeti način na koji funkcioniraju standarne metode za detekciju gena potrebno je prethodno iznijeti neke od bioloških karakteristika eksona i introna budući da se metode zasnivaju na prepoznavanju istih te okarakterizirati modele koji se pri tom rabe (Skriveni Markovljev model, neuronske mreže te SVM).

2. Biološke osnove

Sekvence dvostruke, spiralno zavijene molekule sastavljene od niza nukleotida (adenin (A), gvanin (G), citozin (C) i timin(T)) nazivamo genima. Između gena nalaze se intergenska područja koja transkripcijom ne grade mRNA pa samim time niti proteine. Svaki je gen sačinjen od niza eksona koji određuju informacije o slijedu aminokiselina u proteinu (kodirajući djelovi) te introna koji predstavljaju nekodirajući slijed nukleotida, tj. ne prevode se u proteine iako (zajedno s eksonima) sudjeluju u sintezi pre-mRNA, važno je napomenuti da prokariotski organizmi ne posjeduju introne.

DNA lanac: AATACTGCTAACCTATACGT...

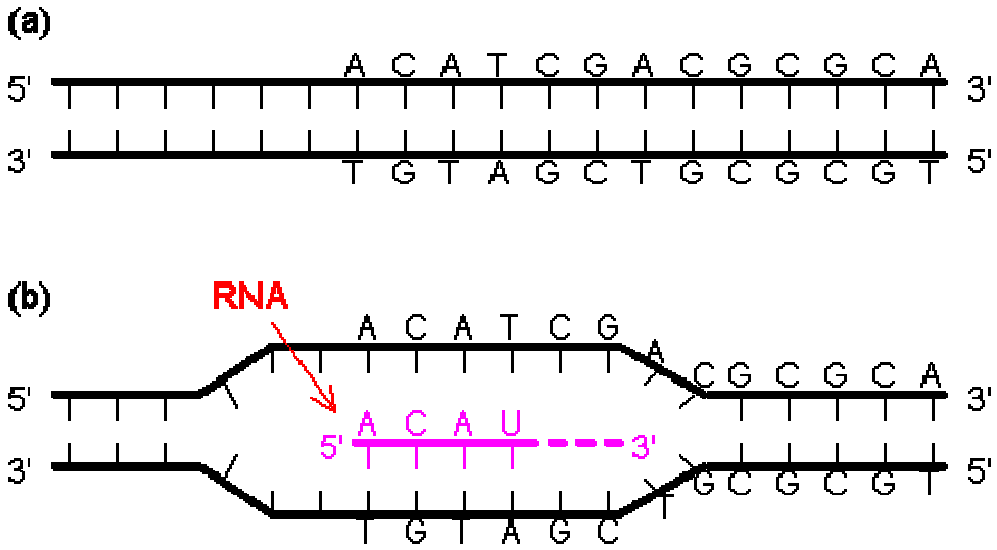
Nezrela gRNK: AAUACUGCUAACCUAUACGU...



Slika 2.1 Sinteza proteina

Čitav se poces nastanka funkcionalnog produkta (tzv. ekspresija gena) može se sagledati kroz tri faze:

1. Inicijacija: enzim RNA polimeraza se veže na točno određenu poziciju na genu koju nazivamo promotor. Promotor se u procesu transkripcije ne prepisuje te se sastoji od kratkog niza naizmjenično vezanih T i A nukleotida tzv. TATA bloka (TATA blok je prisutan u 85% gena eukariotskih organizama, ulogu promotora kod eukariota mogu imati i CAAT blok te GC blok).Transkripciju simuliraju pojačivači („enhanceri“, poboljšavaju vezu RNA polimeraze) koji se nalaze uzvodno ili nizvodno od gena na udaljenosti od nekoliko tisuća kilobaza (kb).
2. Elongacija: razdvajaju se lanci DNA od kojih jedan postaje „kalup“ RNA polimeraza se kreće duž „kalupa“ te po principu komplementarnosti veže odgovrajauće nukleotide pri čemu nastaje svojevrstan hibrid DNA i RNA.
3. Terminacija: RNA polimeraza nalijeće na niz nukleotida koji predstavljaju stop signal (stop kodon). Stop kodon se najčešće sastoji od nekoliko uzastopnih adenina povezanih palindrom sekvencom.



Slika 2.2 a) dvolančana molekula DNA b) molekula DNA u fazi transkripcije (tzv. hibrid)
Izvor: http://www.web-books.com/MoBio/Free/images

Metode otkrivanja kodirajućih područja unutar gena oslanjaju se na specifičnost građe eksona te introna. Svaki je intron započet slijedom GT ili GC. Slijed GT zastupljen je u 99% slučajeva, a GC u preostalih 1%, početak introna naziva se donorsko mjesto izrezivanja („donor splice site“) ili 5' mjesto izrezivanja .Kraj introna (3' mjesto izrezivanja, akceptorsko mjesto izrezivanja – „acceptor splice site“) sastoji se od slijeda AG. Između te dvije točke nalazi se i mjesto grananja(„branch site“) oblika CT(A ili G)A(C ili T) na koje se, netom prije početka izrezivanja, spaja snRNP koji ujedno i započinje izrezivanje čija je osnovna značajka izdvajanje introna iz lanca na donorskom i akceptorskom mjestu nakon čega se isti međusobno povežu preko mjesta grananja tvoreći zrelu glasničku RNK. Prilikom spomenutog procesa koji ubrzava i nadzire makromolekula spliceosome sastavljena od 5 malih jezgrenih ribonukleoproteina (snRNP-U1,U2,U3,U4,U5,U6) na 5' kraj pre-mRNA dodaje se RNA Cap (modificirani nukleotidi, štite mRNA od razgradnje te pomažu pri započinjanju translacije tj. sinteze proteina) dok se na 3' kraj dodaje slijed adenina tzv. polyA kraj. Važno je napomenuti da početak i kraj introna ne određuju jednoznačno položaj eksona, naime prosječan gen je sastavljen od 1500 do 6000 nukleotida, stoga je broj prihvatljivih kombinacija koje odgovarajau intronu na temelju njegove tri karakteristične točke iznimno velik , primjerice za gen duljine 6677 nukleotida dobije se 190218 mogućih kombinacija*. Pre-mRNA produkt transkripcije ne mora nužno biti isključivo jedna mRNA, već više njih koje se javljaju kao rezultat takozvanog alternativnog procesiranja („alternative splicing“- različito izrezivanje introna i povezivanje eksona), dakle ekspresija jednog gena rezulltira sintezom dvaju ili više mRNA te samim time i sintezom dvaju ili više proteina. Važno je napomenuti da su sekvence nukleotida u intronima nasumično raspoređene dok se u eksonima javlja određena periodičnost. Za uzrok te periodičnosti pretpostvljena je učestalost pojave pojednih kodona koji se najčešće javljaju u uzorku RNY (R- A/G N-C/T Y-bilo koja baza) .Važno je napomenuti da po tri baze unutar eksona čine jedan kodon

**Baze raspoređene u okolini donorskog i akceptorskog mjesta izrezivanja međusobno su zavisne (8 nukleotida lijevo i desno od G(T/C) te 26 nukleotida lijevo, zatim 8 nukleotida desno od AG. Zavisnost proizlazi iz samog procesa izrezivanja, važno je napomenuti da se baze mogu uparivati tvoreći pri tom veze različite jakosti zbog čega se nameće ideja o uporabi neuronske mreže koja bi na temelju*

„učećih“ primjera mogla ustanoviti ispravnost mjesta izrezivnja te u ovisnosti o istom vraćati 0 ili 1.

---

kojem odgovara točno jedna aminokiselina. Od 64 moguća kodona 1 otpada na strat, a 3 na stop kodon, no u prirodi postoji „samo“ 20 aminokiselina iz čega slijedi da je jednu aminokineslinu moguće odrediti s više međusobno različitih kodona. U citoplazmi stanice, točnije na ribosomu, vrši se proces translacije kodona u odgovarajuću aminokiselinu (opisani postupak vrijedi za eukariotske organizme, odgovarajući postupak za prokariote sadrži određene ključne razlike).

Tablica 1. Sve mogućnosti kombiniranja nukleotida u aminokiselinama:

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

3. Markovljevi modeli

3.1 Markovljevi lanci

Kako bi određeni prirodni proces uopće bilo moguće analizirati potrebno ga je interpretirati odgovarajućim apstraktnim modelom pri čemu osobito valja paziti na promjenjivost procesa u vremenu. Stanje procesa u nekom diskretnom vremenskom trenutku možemo promatrati kao sliku, stoga je cijeli proces zapravo sastavljen od slijeda takvih stacionarnih slika koje predstavljaju stanja procesa u diskretnim vremenskim trenutcima.

Neka je slika zadanog procesa u određenom vremenskom trenutku data s q_t gdje t predstavlja indeks vremenskog trenutka (tj. t=1,2,3,...,n,...)

Svako stanje (sliku) q_t možemo okarakterizirati skupom po volji odabranih slučajnih varijabli stanja x_i :

$q_t=(x_1,x_2,...,x_n)$ (1)

Točno određeno stanje procesa S_i opisano je skupom fiksiranih varijabli stanja:

$S_i=(x_1=x_{1i},x_2=x_{2i},...,x_n=x_{ni})$ (2)

Stoga se, konačno, promjenjivi proces može prikazati nizom Q slika procesa q_t u vremenskim trenutcima t, počev od slike procesa u početnom trenutku t_0, q_0:

$Q=\{q_0,q_1,q_2,...,q_t,..\}=q_0,q_1,...,q_t$,(3) odnosno

$S=\{S_0,S_1,S_2,...,S_t,...\}=S_0,S_1,...,S_t$ (4)

Pretpostavimo da se proces odvija po određenim zakonitostima, tada proučavanje ponašanja procesa možemo svesti na proučavanje uvjetne vjerojatnosti, tj. vjerojatnost da će proces u trenutku t+1 biti u stanju S_j ukoliko je poznat slijed predhodnih stanja (od trenutka t_0 do trenutka t) dana je izrazom :

$P(q_{t+1}=S_j \mid q_t=S_k,q_{t-1}=S_l,...)$ (5)

Međutim kod ovakvog pristupa javljaju se dva problema:

Trebalo bi definirati po jedno stanje za svaki odabir slučajnih varijabli u određenom vremenskom trenutku što dovodi do beskonačno velikog skupa stanja, nadalje svako stanje q_t ima neograničen broj prethodnih stanja. Prvi problem moguće je riješiti na način da se pretpostavi stalnost zakonitosti (u vremenu) kojima se proces podvrgava, tako dobivamo konačan skup stanja S u kojem se proces može naći u određenom vremenskom trenutku: $S=\{S_1,S_2,S_3,...,S_n\}$. Kako bi se rješio i drugi problem potrebno je uvesti tzv. Markovljevu pretpostavku, tj. prepostavku da trenutno stanje ovisi o konačnom broju prethodnih stanja. Ovako definirani procesi se nazivaju markovljevi procesi ili markovljevi lanci od kojih je najednostavniji lanac upravo lanac prvog reda gdje stanje u trenutku t+1 ovisi isključivo o stanju procesa u trenutku t, stoga izraz (5) prelazi u :

$P(q_{t+1}=S_j \mid q_t=S_k,q_{t-1}=S_l,...) =P(q_{t+1}|q_t=S_k)$(6)

Nadalje, budući da su procesi stacionarni, vjerojatnosti se ne mijenjaju tijekom vremena, stoga se može usvojiti fiksan skup parametara a_{ij} sa sljedećim svojstvom:

$a_{ij}=P(q_{t+1}=S_j|q_t=S_i)$, $1 \le i, j \le N$, za svaki t

Koeficijenti a_{ij} nazivaju se koeficijentima tranzicije te moraju zadovoljavati uvjete:

$$a_{ij} \geq 0$$

$$\sum_{j=1}^{N} a_{ij} = 1$$

Budući da se u ovako definiranom stohastičkom procesu u svakom trenutku točno zna u kojem je stanju sustav, on bi se mogao okarakterizirati kao mjerljiv.

Primjer (opis meteorološkog vremena):

Neka je skup stanja definiran na sljedeći način:

S1:padaline

S2:oblačno

S3:sunčano

Uz uvjet da se vremenske prilike tijekom dana t mogu prikazati isključivo jednim od gornjih stanja.

Neka je matrica prijelaza određena sljedećim vjerojatnostima:

$$A=a_{ij}=\begin{bmatrix} 0,4 & 0,3 & 0,3 \\ 0,2 & 0,6 & 0,2 \\ 0,1 & 0,1 & 0,8 \end{bmatrix}$$

Vjerojatnost da će vrijeme u sljedećih sedam dana biti: sunce-sunce-sunce-kiša-kiša-kiša-oblačno, tj. $O=\{S_3,S_3,S_3,S_1,S_1,S_1,S_2)$, $t=1,..,8$ uz uvjet da je prvog dana kišovito jest:

$P(O|MODEL)=P(S_1,S_3,S_3,S_3,S_1,S_1,S_1,S_2|MODEL)=$

$$P(S_1) * P(S_3|S_1) * P(S_3|S_3) * P(S_3|S_3) * P(S_1|S_3) * P(S_1|S_1) * P(S_1|S_1) * P(S_2|S_1)=$$

$\Pi_1 * a_{31} * a_{33} * a_{33} * a_{33} * a_{13} * a_{11} * a_{11} * a_{21}=1 * 0,1 * 0,8 * 0,8 * 0,8 * 0,3 * 0,4 * 0,2=0,0122=1,22\%$

3.2 Skriveni Markovljevi modeli

Kroz prethodno poglavlje razmotreni su procesi kod kojih je stanje u kojem se proces trenutno nalazi moguće izravno odrediti, međutim problemi su često mnogo složeniji te o stanju procesa možemo zaključiti samo posredno i to preko varijable koju proces u trenutnom stanju emitira.

Formalna definicija:

N: broj stanja u kojem se proces može naći u određenom vremenskom trenutku (stanja su skrivena, no postoji jasna fizička interpretacija istih)

M:broj različitih opservacijskih simbola, skup ovih simbola obično se označava s:

$V=\{v_1,v_2,\ldots,v_m\}$

A:matrica vjerojatnosti prelaska procesa iz stanja u stanje:

$$A=\lfloor a_{ij} \rfloor,\ a_{ij}=P(q_{t+1}=S_j \mid q_t=S_i),1\le i,j\le N$$

B: matrica vjerojatnosti emitiranja pojedinog simbola

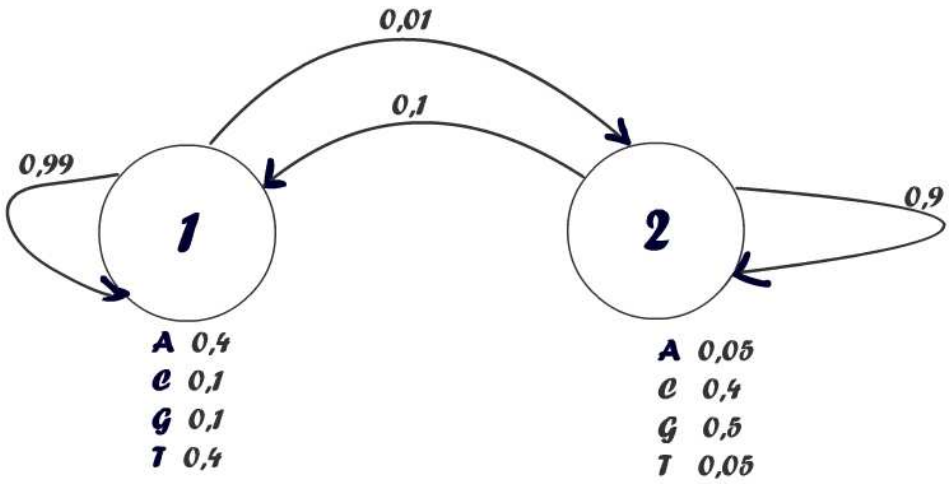
$$B=\lfloor b_j(k) \rfloor, b_j(k)=P(v_t=b_k \mid q_t=S_j),1\le j\le N,1\le k\le M$$

Π:matrica tranzicijskih vjerojatnosti

$$\Pi=[\Pi_i],\Pi_i=P(q_0=S_i),1\le i\le N$$

Potpuni opis skrivenog Markovljevog modela dan je s:

$$\lambda=(A,B,\Pi)$$



Stanja (skrivena):

...	1	1	1	1	1	2	2	2	2	1	1
Prijelazi:	?	0,99	0,99	0,99	0,99	0,01	0,9	0,9	0,9	0,1	0,99

Simboli(vidljivi):

Emisije:	A	T	C	A	A	G	G	C	G	A	T
	0,4	0,4	0,1	0,4	0,4	0,5	0,5	0,4	0,5	0,4	0,4

Primjer 3.2.1

4. Osnovni pristupi ka detektiranju gena

4.1 Ab initio pristup

Važna značajka ab initio pristupa jest da se detekcija gena ne zasniva isključivo na usporedbi s već postojećim sličnim genom (ili proteinom) koji je prethodno pohranjen u neku od bioloških baza podataka kao što su Ensembl, RefSeq,GenBank, već se on detektira na temelju određenih vlastitih bioloških karakteristika. Neki od najranijih ab intio algoritama nastoje detektirati individualne funkcionalne elemente kao što su promotori, mjesta prekida, kodirajuće regije itd.. Suvremenije metode nastoje integrirati višestruke tipove informacija uključujući signalne senzore, kompozicijska svojstva kodirajuće i ne kodirajuće DNA te samo u nekim slučajevima pretraživanje baze podataka u svrhu pronalaženja homologa kako bi se predvidjele kompletne genomske strukture. Ab initio algoritmi baziraju se ne nekoj od tehnika raspoznavanja uzoraka kao što su skriveni Markovljev model, neuronske mreže te SVM (Support vector machines).

Primjeri netom okarakteriziranih suvremenih programa jesu:

4.1.1 GRAIL(Gene Recognition and Analysis Internet Link):

(Uberbacher and Mural, 1991; Mural 1992.) Predstavlja jednu od najranijih metoda za detekciju gena te uživa široku primjenu. Javlja se u dvije verzije: GRAIL 1 – temeljen je na neuronskoj mreži kojom razaznaje potencijalne kodirajuće regije fiksne dužine (100 baza) a da pri tom ne vodi računa o start i stop kodonu te GRAIL 1a koji predstavlja proširenje netom navedene metode,tj. uzima u obzir i neposredne susjede potencijalno kodirajućih regija. GRAIL i GRAIL 1a su iznimno prikladni kod traženja jednostrukih eksona. Daljnje usavršavanje GRAIL-a rezultiralo je GRAIL-om 2 kod kojeg se u obzir uzimaju i start i stop kodon te polyA signali.

4.1.2 FGENEH/FGENES:

(Victor Solovyev 1994./1995.) Jest metoda kojom se eksoni predviđaju na način da se pogledaju strukturalne značajke kao što su donori i akceptori kod prekida, kodirajuće regije te intronske regije (obije 3' i 5'). Zasniva se na linearnoj diskriminativnoj analizi, tj. matematičkoj tehnici koja dozvoljava da podaci iz više pokusa budu kombinirani.

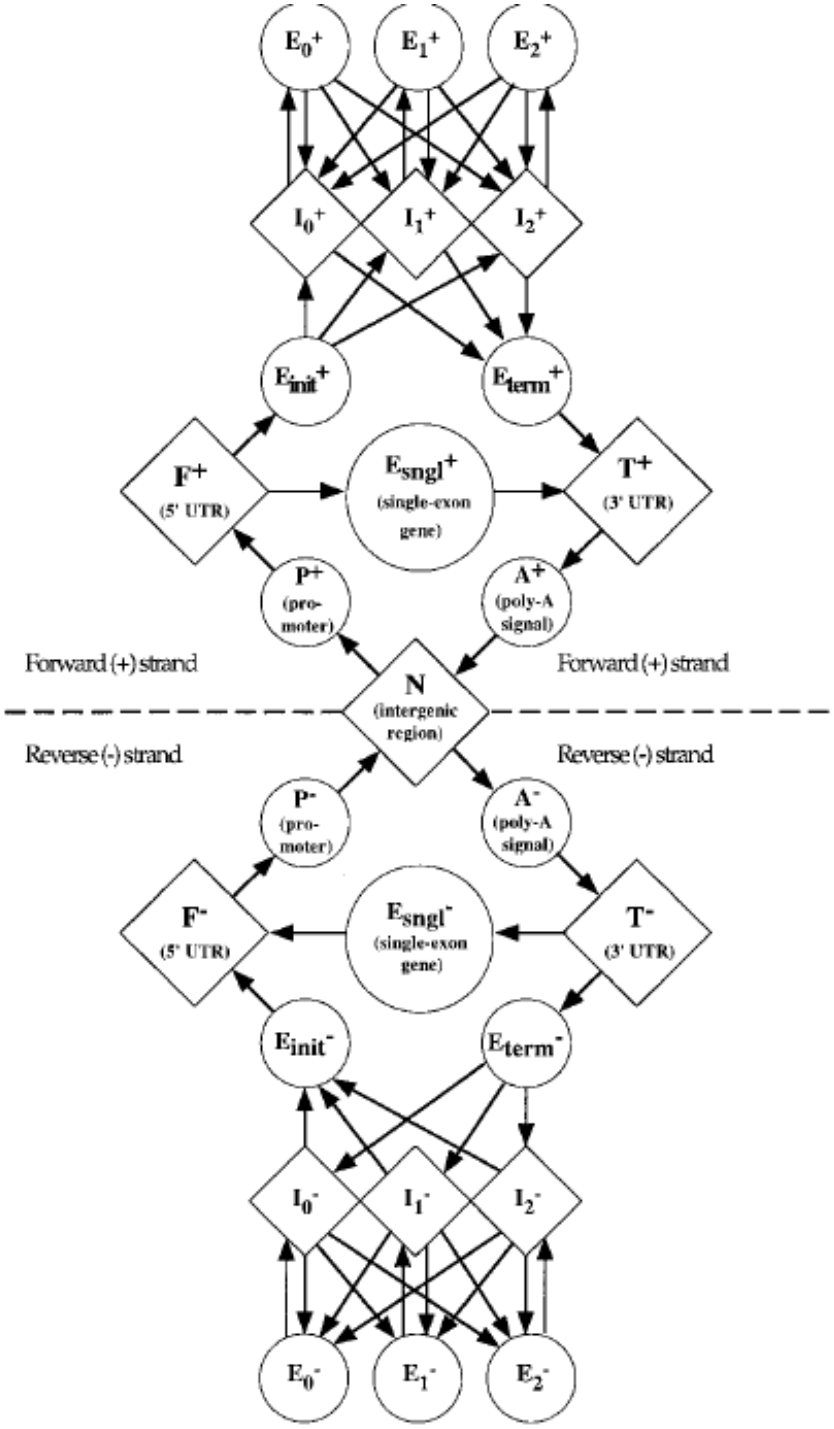
4.1.3 MZEF:

"Michael Zhang's Exon Finder"

Kod ove metode predviđanje se temelji na tehnici zvanoj kvadratna diskriminativna analiza. Pretpostavimo da je na istom jednostavnom xy grafu ucrtana učestalost mjesta prekida na eksonu te duljina istog. Ukoliko je veza između navedenih dviju

varijabli nelinearna ili multivarijantna rezultirajući graf će nalikovati na roj točaka, no tek će manji dio točaka predstavljati točna predviđanja. Kako bi se razdvojile točke iza kojih se kriju točna predviđanja od točki koje predstavljaju netočna rabi se kvadratna funkcija. Ova je metoda prevenstveno namjenjena za utvrđivanje internih kodirajućih djelova eksona, međutim ona ne daje apsolutno niti jednu drugu informaciju vezanu uz strukturu gena.

4.1.4 GENSCAN:



Slika 4.1.1 Model strukture sekvence
Izvor: Chris Burge, Samuel Karlin: Prediction of Complete Gene Structures in Human Genomic DNA str. 86. (1997.)

(Burge i Karlin 1997.,1998.) dizajniran kako bi predviđao kompletne genetske strukture. Kao takav može detektirati introne, promotore, polyA signale. Kao i FGENES, GENSCAN ne očekuje da ulazna sekvenca bude isključivo jedan gen ili jedan ekson već može vršiti relativno precizna predviđanja i na sekvencama koje predstavljaju samo djelove gena ili više gena razdvojenih intergenskom DNA. (Napomena: ulazna sekvenca prije unošenja u GENSCAN ili bilo koji drugi od ovdje navedenih programa mora biti obrađena programom CENSOR ili nekim srodnim kako bi se utvrdili repetitivni elementi) GENSCAN se zasniva na probabilističkom modelu (pojam uveli autori) kompozicije genomskih sekvenci i genetske strukture,tj. na skrivenom Markovljevom modelu petog reda, stoga ga karakterizira viša razina točnosti no kod ostalih programa rabljenih u istu svrhu, tj. 75% do 80% eksona biva točno detektirano (kod ostalih programa točnost predviđanja se kreće oko niskih 45%), no činjenica je da su rezultati dobiveni ovom metodom često komplementarni rezultatima dobivenima algoritmima čiji je rad zasnovan na homologiji sekvenci (BLAST). 1996. GENSCAN biva testiran s vrlo velikim setom sekvenci gena kralješnjaka Tražeći genske strukture koje su konzistentne s upitom (zadana sekvenca) algoritam određuje vjerojatnost da dio date sekvence predstavlja ekson, promotor i sl. "Optimalni eksoni" su upravo oni s najvećom vjerojatnošću, tj. ti djelovi sekvence imaju najveću vjerojatnost da uistinu i budu eksoni. Metoda također predviđa i "suboptimalne eksone" čija je vjerojatnost prihvatljiva, no ne tako visoka kao kod "optimalnih eksona". Autori metode preporučaju ispitivanje oba rezultata kako nestandardne genske strukture ili alternativno izdvojene regije ne bi bile previđene. GENSCAN eksoni koji imaju vjerojatnost ispod 0,5 se smatraju nepouzdanima dok vrijednost iznad 0,9 označava ispravnu detekciju eksona u 88% slučajeva.

Model strukture sekvence:

Svaki krug ili romb (prikazan na slici 4.1.1) predstavlja funkcionalno stanje gena ili genomske regije: N-intergenska regija, P-promotor, F 5' neprevodeća, E_{sngl}-jednostruki ekson (nema introne, od početka translacije- start kodona do stop kodona), E_{init} početni ekson (od početka translacije – start kodona do mjesta prekida), E_k-faza k unutarnjeg eksona (od mjesta prekida akceptora do mjesta prekida donora), E_{term}-terminirajući ekson (od mjesta prekida akceptora do stop kodona), T – 3' neprevodeća regija (od stop kodona do polyA signala), A-polyA signal, I_k-faza introna. Valja uočiti da prikazani model ima vrlo važan nedostatak, tj. model ne podržava preklapajuće transkripcijske jedinice, stoga alternativno procesiranje nije eksplicitno adresirano.

Neka je s q zadan skup stanja $q=\{q_1,q_2,...,q_n\}$, stanja su povezana sa skupom $d=\{d_1,d_2,...,d_n\}$ duljina odnosno trajanja. Tako definirana stanja generiraju DNA sekvencu S ukupne duljine: $L=\sum_{i=1}^{n} d_i$. Početno stanje q_1 odabrano je prema inicijalnoj distribuciji stanja. Matrica tranzicijskih vrijednosti jest $\Pi_i=P\{q_1=Q(i)\}$, gdje $Q_{(j)}(j=1,2,...,n)$ predstavlja točno određeno stanje procesa u trenutku j.

Generiranje raščlambe korespodentne duljini sekvence:

Duljina (trajanje stanja) d_1 odgovarajuća je stanju q_1, a generirana je uvjetno ovisno o vrijednosti $q_1=Q_{(i)}$ iz distribucije duljina $fQ_{(i)}$.Segment sekvence s_1 duljine d_1 generiran je (uvjetno) prema modelu karakterističnom za određeno stanje.

Podsekvencijsko stanje q_2 također je uvjetno generarirano ovisno o q_1 (model prvog reda) tj,.

$A_{ij}=P\{q_{k+1}=Q_{(j)}|q_k=Q_{(i)}\}$

Gornji se postupak ponavlja sve dok suma $\sum_{i=1}^{n} d_i$ ne premaši duljinu L, nakon čega se zadnje stanje odbaci. Nova se sekvenca sad jednostavno dobiva konkatenacijom segmenata tj. $S=S_1S_2...S_n$ važno je napomenuti da novodobivena sekvenca ne predstavlja nužno isključivo jedan gen, već se ona može sastojati od više gena ili samo djela gena.

Definirajmo vjerojatnosni prostor $\Omega = \Phi_L \times \ell_L$ gdje je Φ_L skup svih mogućih raščlambi duljine L te ℓ_L skup svih mogućih DNA sekvenci duljine L.

Za određenu sekvencu $S \in \ell_L$ možemo odrediti točno određenu raščlambu $\phi_i \in \phi_L$,te prema Bayesovoj formuli vrijedi:

$P\{\phi_i \mid S\}=P\{\phi_i, S\}/P\{S\}=P\{\phi_i, S\}/\sum_{\phi_j \in}$

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Init	+	4697	4801	105	1	0	64	80	103	0.651	7.58
1.02	Intr	+	5725	5838	114	0	0	48	91	116	0.993	7.62
1.03	Intr	+	10004	10081	78	1	0	61	70	78	0.809	2.13
1.04	Intr	+	10222	10317	96	0	0	94	87	117	0.999	11.49
1.05	Intr	+	11534	11640	107	1	2	118	62	31	0.953	1.59
1.06	Intr	+	13643	13712	70	2	1	88	111	32	0.950	3.77
1.07	Intr	+	15684	15776	93	2	0	45	98	59	0.782	1.84
1.08	Intr	+	16702	16797	96	0	0	70	100	26	0.709	1.29
1.09	Intr	+	17428	17658	231	0	0	69	79	233	0.911	17.55
1.10	Intr	+	19932	19986	55	2	1	90	94	29	0.805	1.33
1.11	Term	+	25128	25375	248	1	2	48	48	167	0.867	3.67
1.12	PlyA	+	25382	25387	6							1.05
2.00	Prom	+	26739	26778	40							-7.05
2.01	Init	+	27929	28093	165	1	0	77	94	65	0.948	5.68
2.02	Intr	+	28140	28223	84	2	0	69	64	142	0.901	9.00
2.03	Intr	+	29931	30071	141	2	0	126	38	55	0.262	3.93
2.04	Intr	+	52002	52164	163	2	1	99	17	149	0.194	7.53
2.05	Intr	+	53036	53243	208	0	1	48	-2	191	0.028	3.31
2.06	Intr	+	58789	58968	180	1	0	82	35	127	0.411	4.86
2.07	Intr	+	59932	60222	291	1	0	69	20	255	0.369	12.13
2.08	Intr	+	63258	63277	20	0	2	102	86	-16	0.527	-5.06
2.09	Intr	+	64481	64648	168	0	0	47	86	162	0.939	10.90
2.10	Intr	+	69012	69112	101	1	2	56	75	115	0.967	5.91
2.11	Intr	+	69496	69579	84	0	0	25	115	57	0.615	1.20
2.12	Intr	+	71019	71092	74	2	2	105	90	-21	0.950	-2.91
2.13	Term	+	73744	74779	1036	1	1	85	44	805	0.960	66.40
2.14	PlyA	+	75266	75271	6							1.05
3.11	PlyA	-	75947	75942	6							1.05
3.10	Term	-	83049	82945	105	0	0	77	38	68	0.831	-1.87
3.09	Intr	-	83245	83180	66	1	0	113	94	43	0.948	5.58
3.08	Intr	-	83509	83367	143	2	2	108	69	88	0.995	8.05
3.07	Intr	-	87709	87655	55	1	1	50	115	63	0.988	2.83
3.06	Intr	-	89754	89539	216	0	0	110	42	182	0.727	13.58
3.05	Intr	-	90488	90378	111	2	0	25	100	169	0.499	11.46
3.04	Intr	-	92145	92053	93	0	0	109	59	52	0.893	3.64
3.03	Intr	-	92307	92238	70	2	1	101	67	38	0.955	1.27
3.02	Intr	-	93882	93776	107	0	2	70	68	84	0.640	2.69
3.01	Intr	-	95364	95269	96	0	0	68	75	106	0.661	6.59

Predicted peptide sequence(s):

```
>AC002467.seq|GENSCAN_predicted_peptide_1|430_aa
MLAASPSIAVVAYAIAVSVGKVYATKYDYTIDGNQEFIAPGISNIFSGFFSCFVATTALS
RTAVQESTGGKTQVAGIISAAIVMIAILALGKLLEPLQKSVLAAVVIANLKGMFMQLCDI
PRLWRQNKIDAVIWVFTCIVSIILGLDLGLLAGLIFGLLTVVLRVQFPSWNGLGSIPSTD
<remainer of output truncated>
```

Slika 4.1.2 Primjer izlaza GENSCAN-a: Prva kolona lijevo predstavlja naziv eksona i gena, nadalje s *Type S*. označen je tip predviđanja (+/- ovisno o strani – za negativnu tj. obrnutu stranu (model sekvence)), Begin/End – početna i završna točka predviđanja, Len – duljina ORF-a, P – vjerojatnosna vrijednost
Izvor: Andreas D. Baxevanis, B. F. Francis Ouellette: *Bioinformatics – A particular Guide to the Analysis of Genes and Proteins str. 242. (2001.)*

4.1.5 GENEID:

(Guigo´ 1992.)Trenutna verzija pronalazi eksone na temelju mjere kodnog potencijala. Inicijalno, ovaj je program bio jedan od najbržih koji su koristili „rule-based“ sustav za ispitivanje eksona i označavanje potencijalnih gena. GENEID koristi metodu težinskih matrica kako bi ustanovio predstavlja li zadani dio sekvence mjesto prekida ili start, stop kodon.

4.1.6 HMMgene:

(Krogh, 1997)Predviđa kompletne gene u bilo kojoj zadanoj DNA sekvenci koristeći pri tom skriveni Markovljev model. Ova metoda osim najboljeg rješenja vraća i alternativna. Sekvence se unose u FAST-a formatu.

4.2 Komparativni pristup

Budući da je kompletan genom mnogih vrsta već sekvenciran vrlo obećavajuću metodu detekcije gena predstavlja upravo komparativna genomika koja funkcionira na način da translatirana sekvenca postaje subjekt za pretragu biloške baze podataka. Najučestalija komparativna metoda jest poravnanje sekvenci, "poravnavati" i uspoređivati se mogu samo dvije sekvence tzv. "*pairwise alignment"* te više sekvenci tzv. "*multliple alignment"*.

4.2.1 BLAST(Basic Local Alignment Search Tool)

Predstavlja heuristički algoritam za usporedbu biloških sekvenci (program koji ga interpretira se također naziva BLAST). Razvijen je 1990. godine *(Eugene Myers, Stephen Altschul, Warren Gish, David J. Lipman, Webb Miller)* s namjerom da se ubrza već postojeći FASTA algoritam. Novitet kod BLAST-a jest ideja o susjednim riječima, tj. više nije nužno da se dvije riječi podudaraju u potpunosti već je dovoljno da se riječ iz sekvence subjekta podudara sa sekvencom riječi iz baze na vrijednost veću ili jednaku od već unaprijed određenog parametra T. Takve dvije riječi koje zadovoljavaju netom navedeni uvjet nazivamo HSP („High-Scoring Segment pair"). Važno je napomenuti da ocjena poravnanja predstavlja lokalni maksimum, tj. produljenje poravnanja bez obzira na smijer smanjuje ocjenu. Sekvence se uspoređuju metodom supstitucijskih matrica. Najčešće rabljena matrica jest BLOSUM62. Kritični parametri o kojima ovisi brzina i osjetljivost algoritma su W-duljina riječi te T-zahtjevani postotak sličnosti, njihovim mjenjanjem utječemo na rezultate algoritma, npr. ukoliko povećamo T broj pogodnih riječi se smanjuje što rezultira bržim radom programa, smanjenjm parametra T možemo utvrđivati "daleke" veze među sekvencama. Osnovni BLAST algoritam se sastoji od sljedećih koraka:

1.Ulazna se sekvenca raščlanjuje na k-torke (riječi duljine k) po defaultu k=3 za aminokiseline, k=11 za nukleotide. Za određenu k-torku definiraju se susjedne riječi kao nizovi, također duljine k, koji pri poravnanju s promatranom k-torkom daju ocjenu veću od određenog praga T (riječi se uspoređuju metodom težinskih matrica, kod BLASTA se najčešće primjenjuje BLOSUM62, gdje 62 označava minimalan postotak sličnosti sekvenci unutar oređenog klastera).

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W

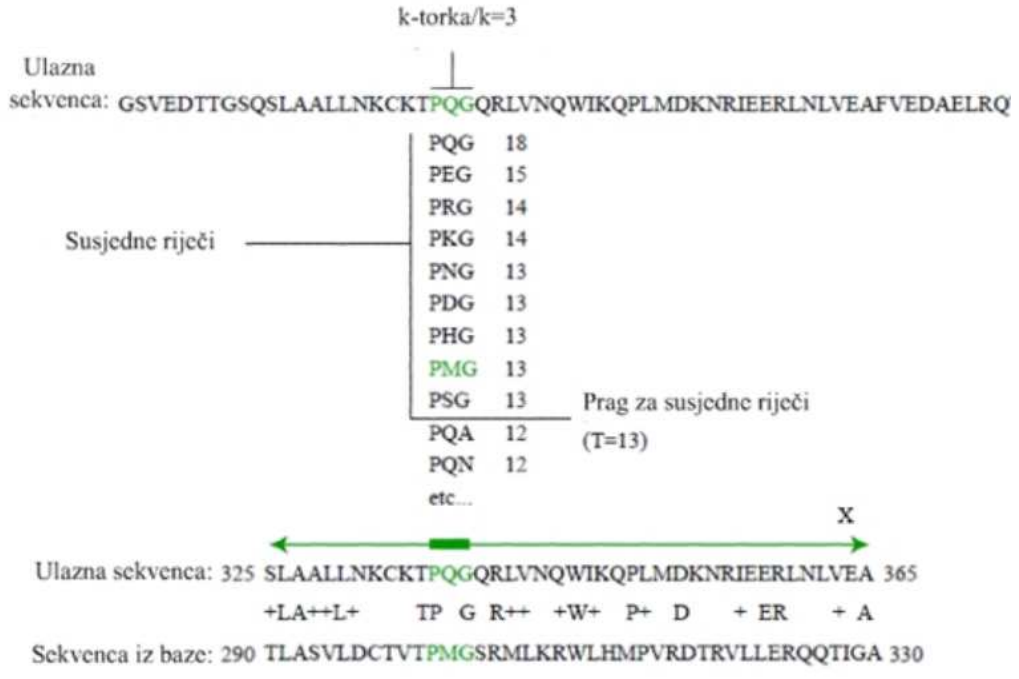
Slika 4.2.1 BLOSUM 62 supstitucijska matrica.
Izvor:http://www.users.math.umd.edu/~poorani/sampletalk/talk.html#four

---

2.Podniz sekvence iz baze koji je jednak ciljnoj riječi predstavlja hit

3.Svaki se hit proširuje i na lijevu i na desnu stranu sve dok vrijednost ukupne ocjene ne počne padati.

4.Za svaki pronađeni HSP računa se važnost dobivenog rezultata na temelju distribucije ekstremnih vrijednosti EVD mjeru statističke važnosti se označava s E te ona predstavlja očekivani broj poravnanja slučajne sekvence s ocjenom većom ili jednakom od ocjene promatranog HSP-a (E zapravo označava kolika je vjerojatnost da je određeni HSP slučajan, tj. da sekvence nisu niti funkcijski niti evolucijski povezane.



Slika 4.2.2 Grafički prikaz BLAST algoritma
Izvor: http://en.wikipedia.org/wiki/BLAST

Sekvence koje želimo analizirati unose se u genbank ili FASTA formatu. Primjer proteinske sekvence u fasta formatu:

```
>OTTMUSP00000033776 pep:all chromosome:VEGA37:16:91373028:91405834:1
Gene:OTTMUSG00000027167 Transcript:OTTMUST00000067197
MRSRCTVSAVGLLSLCLVVSASLETITPSAFDGYPDEPCTINITIRNSRLILSWELENKS
GPPANYTLWYTVMSKDENLTKVKNCSDTTKSSCDVTDKWLEGMESYVVAIVIVHRGDLTV
CRCSDYIVPANAPLEPPEFEIVGFTDHINVTMEFPPVTSKIIQEKMKTTPFVIKEQIGDS
VRKKHEPKVNNVTGNFTFVLRDLLPKTNYCVSLYFDDDPAIKSPLKCIVLQPGQESGLSE
SAIVGITTSCLVVMVFVSTIVMLKRIGYICLKDNLPNVLNFRHFLTWIIPERSPSEAIDR
LEIIPTNKKKRLWNYDYEDGSDSDEEVPTASVTGYTMHGLTGKPLQQTSDTSASPEDPLH
EEDSGAEESDEAGAGAGAEPELPTEAGAGPSEDPTGPYERRKSVLEDSFPREDNSSMDEP
GDNIIFNVNLNSVFLRVLHDEDASETLSLEEDTILLDEGPQRTESDLRIAGGDRTQPPLP
SLPSQDLWTEDGSSEKSDTSDSDADVGDGYIMR
```

Postoji nekoliko varijanti BLAST-a:

- Nukleotid BLAST: pretražuje databazu nukleotida pri čemu je sekvenca koja predstavlja upit također zadana kao niz istih. (*Algoritmi: blastn, megablast, diskontinuirani megablast*).
- Blastp: pretražuje databazu proteina rabeći pri tom proteinsku sekvencu kao upit. (*Algoritmi: blastp, psi-blast, phi-balst*).
- Tblastn: pretražuje databazu translatiranih nukleotida uz uvjet da je sekvenca koja predstavlja upit zadana kao slijed proteina.
- Blastx: pretražuje databazu proteina rabeći sekvencu translatiranih nukleotida kao upit.
- Tblastx: pretražuje databazu nukleotida rabeći pri tom sekvencu sastavljenu od slijeda istih.

5. Praktični dio

5.1 Nepostojeći geni

U većini kralješnjaka postoji gen i odgovarajući mu protein $IFNAR_2$ (interferon – alfa receptor, peptidni lanac 2), no ukoliko se pretraži jedna od databaza kompletnih genoma – Ensembl, te pregleda skup svih organizama koji navodno imaju $IFNAR_2$ može se uočiti određena nepravilnost, tj. neki od međusobno vrlo srodnih organizama, čini se, zapravo ne posjeduju prethodno navedeni protein, primjerice, isti je detektiran u mišu, no vjerujući Ensemblu ne postoji u štakoru, nadalje, nije detektiran niti u gorili iako je prema rezultatima upita prisutan kod čovjeka što je vrlo sumnjiva tvrdnja. Budući da Ensembl kao metodu za otkrivanje gena koristi upravo GENSCAN, postoji realna mogućnost da GENSCAN u ovom slučaju ne „radi“ ispravno.

Uzmimo sljedeće dvije sekvence u FASTA formatu od kojih 1. predstvalja $IFNAR_2$ kod čovjeka (Homo sapiens), a druga kod miša (Mus musculus):

```
>sp|P48551|INAR2_HUMAN Interferon alpha/beta receptor 2
OS=Homo sapiens GN=IFNAR2 PE=1 SV=1
MLLSQNAFIFRSLNLVLMVYISLVFGISYDSPDYTDESCTFKISLRNFRSILSWELKNHS
IVPTHYTLLYTIMSKPEDLKVVKNCANTTRSFCDLTDEWRSTHEAYVTVLEGFSGNTTLF
SCSHNFWLAIDMSFEPPEFEIVGFTNHINVMVKFPSIVEEELQFDLSLVIEEQSEGIVKK
HKPEIKGNMSGNFTYIIDKLIPNTNYCVSVYLEHSDEQAVIKSPLKCTLLPPGQESESAE
SAKIGGIITVFLIALVLTSTIVTLKWIGYICLRNSLPKVLNFHNFLAWPFPNLPPLEAMD
MVEVIYINRKKKVWDYNYDDESDSDTEAAPRTSGGGYTMHGLTVRPLGQASATSTESQLI
DPESEEEPDLPEVDVELPTMPKDSPQQLELLSGPCERRKSPLQDPFPEEDYSSTEGSGGR
ITFNVDLNSVFLRVLDDEDSDDLEAPLMLSSHLEEMVDPEDPDNVQSNHLLASGEGTQPT
FPSPSSEGLWSEDAPSDQSDTSESDVDLGDGYIMR
```

```
>tr|Q9D1R7|Q9D1R7_MOUSE Ifnar2 protein OS=Mus musculus
GN=Ifnar2 PE=2 SV=1

MRSRCTVSAVGLLSLCLVVSASLETITPSAFDGYPDEPCTINITIRNSRLILSWELENKSG
PPANYTLWTVMSKDENLTKVKNCSDTTKSSCDVTDKWLEGMESYVVAIVIVHRGDLTVCRC
SDYIVPANAPLEPPEFEIVGFTDHINVTMEFPPVTSKIIQEKMKTTPFVIKEQIGDSVRKK
HEPKVNNVTGNFTFVLRDLLPKTNYCVSLYFDDDPAIKSPLKCIVLQPGQESGMARFLKFA
LLF
```

Ukoliko te dvije sekvence unesemo u BLAST te izvršimo pretragu (tblastn) po kompletnom genomu spornih organizama dobit ćemo zanimljive rezultate, naime prema BLAST-u, $IFNAR_2$ je prisutan u oba sporna organizma, što dokazuje tvrdnju o nepreciznosti ab initio metoda tj. da u ovom slučaju GENSCAN zasigurno ne radi ispravno. Podršku za BLAST posjeduju i određeni programski jezici kao što su Python, Ruby i Perl. BLAST pretragu u programskom jeziku Python za dva

prethodno navedena sporna organizma, moguće je izvesti na sljedeći jednostavan način:

```
from Bio.Blast import NCBIWWW
from Bio import SeqIO
organizmi=['Gorilla','Rattus']
bliski_organizmi=['homo_sapiens','mus_musculus']
i=0;
for organizam in organizmi:
    print
'********************ORGANIZAM:'+organizam+'********************'
    record=SeqIO.read(open(bliski_organizmi[i]+'.fasta'),format="fasta")
    i=i+1;
    result_handle=NCBIWWW.qblast("tblastn","nr",record.seq)
    save_file = open(organizam+'.xml', "w")
    save_file.write(result_handle.read())
    save_file.close()
    result_handle.close()
    result_handle=open(organizam+'.xml')
    from Bio.Blast import NCBIXML
    blast_record=NCBIXML.read(result_handle)
    for alignment in blast_record.alignments:
        if alignment.title.find(organizam)!=-1:
            print 'SEKVENCA: ',alignment.title[0:50]+'...'
            print 'DULJINA: ',alignment.length
            for hsp in alignment.hsps:
                      print 'E vrijednost: ',hsp.expect
                      print hsp.query[0:65] + '...'
                      print hsp.match[0:65] + '...'
                      print hsp.sbjct[0:65] + '...'
                      print 'Max ident:
',(100.0*hsp.identities)/hsp.align_length
```

6. Zaključak

Postojeće metode za detekciju gena posjeduju određena ograničenja, stoga ih karakterizira još uvijek nedovoljna točnost pri detekciji određenih gena. Iznadprosječnu točnost zasad pokazuje isključivo GENSCAN koji se zasniva na kombinaciji svih trenutno dostupnih znanja. Činjenica je da su rezultati dobiveni ab initio metodama često oprečni rezultatima dobivenim komparativnim metodama što je ujedno i vrlo važan nedostatak postojećih modela, no treba uzeti u obzir ubrzan razvoj istih, stoga bi netom navedeni problem trebao biti riješen u skorijoj budućnosti. Neki od budućih izazova na području predikcije gena jesu:

- Bolja karakterizacija pojačivača te interpretacija alternativnog procesiranja kako bi modeli mogli predvidjeti i „alternativne“ eksone.
- Bolja identifikacija vrlo kratkih eksona te točnije predviđanje vrlo dugih eksona.
- Utvrđivanje ne – translatirajućih eksona.
- Utvrđivanje značajki mRNA koje se odnose na stabilnost iste te transport.

7. Literatura

[1] http://blast.ncbi.nlm.nih.gov/Blast.cgi

[2] http://en.wikipedia.org/wiki/BLAST

[3] Michael Q. Zhang: *Computational prediction of eukaryotic protein – coding genes (2002.)*

[4] Marcel E. Dinger, Ken C. Pang, Tim R. Mercer, John S Mattick: *Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities (2008.)*

[5] Jennifer Harrow, Alinda Nagy, Alexandre Reymond, Tyler Alioto, Laszlo Patthy, Stylianos E Antonarakis, Roderic Guigo*: Identifying protein – coding genes in genomic sequences (2009.)*

[6] Chris Burge, Samuel Karlin: *Prediction of Complete Gene Structures in Human Genomic DNA (1997.)*

[7] http://bmb.pharma.hr/predavanja/rna/sld006.htm

[8] http://en.wikipedia.org/wiki/FASTA_format

[9] http://en.wikipedia.org/wiki/Gene_prediction

[10] http://www.everythingbio.com/glos/definition.php

[11] Wiley Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics

[12] Andreas D. Baxevanis, B. F. Francis Ouellette: *Bioinformatics – A particular Guide to the Analysis of Genes and Proteins*

8. Sažetak

Računalne metode za detekciju gena te ostalih funkcionalinih djelova genomske DNA karaterizira iznimno brz razvoj kroz protekla 2 desetljeća. Postojeće se metode za detekciju gena oslanjaju na specifičnost građe introna odnosno eksona. Sama preciznost kojom geni bivaju detektirani još uvijek nije zadovoljavajuća. Ka detekciji gena moguće je pristupiti na dva načina:

- Komparativni pristup: budući da je genom mnogih vrsta već sekvenciran, gene je moguće detektirati na način da nepoznata sekvenca postane subjekt za pretragu biološke baze podataka. (GenBank, Ensembl, RefSeq...)
- Ab initio pristup: gen se detektira na temelju određenih vlastitih bioloških karakteristika. Neki od najranijih ab inito algoritama nastoje raspoznati individualne funkcijske elemente kao što su promotori, mjesta prekida, kodirajuće regije itd.. Suvremenije metode nastoje integrirati višestruke tipove informacija. Neki od suvremenih metoda i programa jesu: GENMARK, GeneID, Genie, GeneParser, GENSCAN te GRAIL 2.

Rezultati dobiveni ab initio metodama često su oprečni onima koje daje komparativni pristup.