SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 3255

ALTERNATIVNO SPAJANJE EKSONA

Dorija Humski

Zagreb, lipanj 2013.

Hvala mojoj obitelji!
Hvala mom mentoru Mili Šikiću za pomoć i podršku!

1 Sadržaj

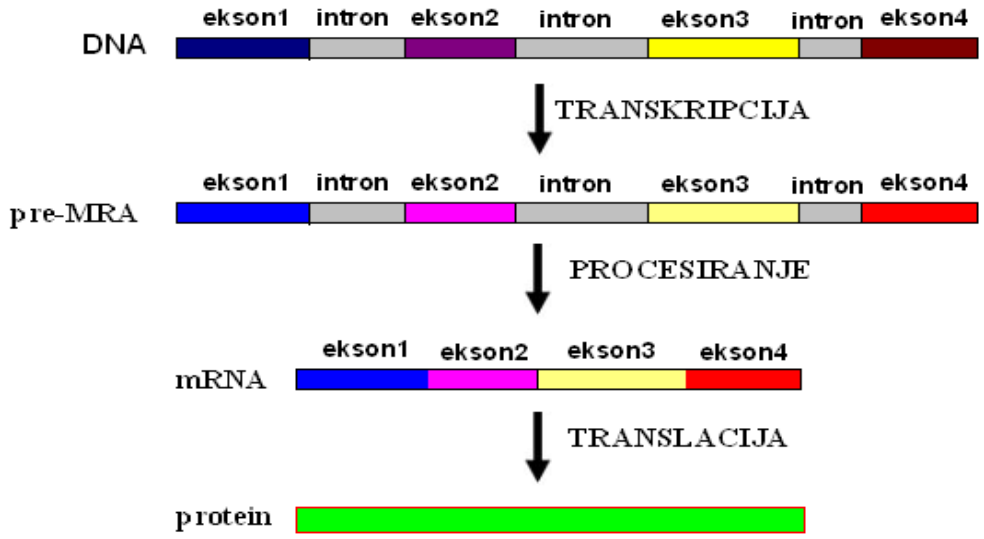
2 Uvod 4
3 Pregled područja 5
3.1 O alternativnom spajanju 5
3.2 Alati i metode 7
3.2.1 Mapiranje 7
3.2.2 Rekonstrukcija u transkripte 9
3.2.3 Analiza transkripata 13
3.3 Alati za otkrivanje alternativnog spajanja eksona 14
3.3.1 MATS 14
4 Podaci 18
4.1 Fasta 18
4.2 Fastq 18
4.3 SAM 19
4.4 BAM 21
4.5 GTF 22
5 Implementacija 22
6 Analiza rezultata 23
7 Zaključak 25
8 Literatura 26
9 Sažetak 27
10 Abstact 27
11 Dodatak A 28

2 Uvod

Izgradnja proteina u stanicama eukariota ključna je za život. Kao takva, ideja za otkrivanja procesa izgradnje proteina i potpuna kontrola nad istim predstavlja izazov današnjice.

Proces izgradnje proteina počinje od molekule DNA. Postupkom transkripcije nastaje RNA molekula iz koje se daljnjom translacijom izgrađuje protein. Kod eukariota, između transkripcije i translacije postoji međukorak – procesiranje (Slika 1.1). Rezultat transkripcije je pre-mRNA koja sadrži introne i eksone, procesiranjem se gradi mRNA molekula koja se sastoji od eksona, a početak i kraj građeni su od nekodirajućih regija. Prilikom izgradnje mRNA može doći do alternativnog spajanja eksona. Alternativno spajanje eksona tema je ovog rada.

Cilj ovog rada je razviti računalnu metodu za otkrivanje alternativnog spajanja eksona. U drugom poglavlju ukratko je opisan proces alternativnog spajanja te su opisane različite metode i alati koji nam mogu pomoći u otkrivanju istog. U trećem poglavlju opisani su podaci i formati podataka koji služe kao prikaz RNA sekvenci i genoma. U četvrtom poglavlju opisana je implementacija algoritma.



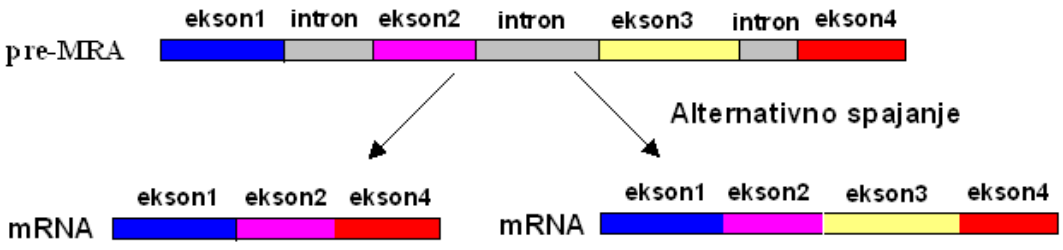
Slika 1.1 Proces izgradnje proteina u stanicama eukariota

3 Pregled područja

U ovom poglavlju opisano je alternativno spajanje, metode koje nam mogu poslužiti kao pomoć u otkrivanju i već gotovi alati.

3.1 O alternativnom spajanju

Alternativno spajanje je proces kod kojeg iz jedne pre-mRNA može nastati više različitih mRNA (Slika 2.1) . Kako iz različitih mRNA nastaju različiti proteini, alternativno spajanje eksona omogućuje nastanak više proteina iz jednog gena.



Slika 2.1 Prikaz alternativnog spajanja eksona

Do sada je otkriveno više modela alternativnog spajanja [1]:

1. Preskakanje eksona :

   Ekson može biti isključen iz rezultirajuće mRNA.

2. Međusobno isključivi eksoni:

   U rezultirajućoj mRNA nalazi se samo jedan od eksona, ni u kojem slučaju ne mogu biti uključeni oba.

3. Alternativno donorsko mjesto:

   Ekson se pridružuje eksonu koji mu prethodi i mijenja njegov 3' kraj.

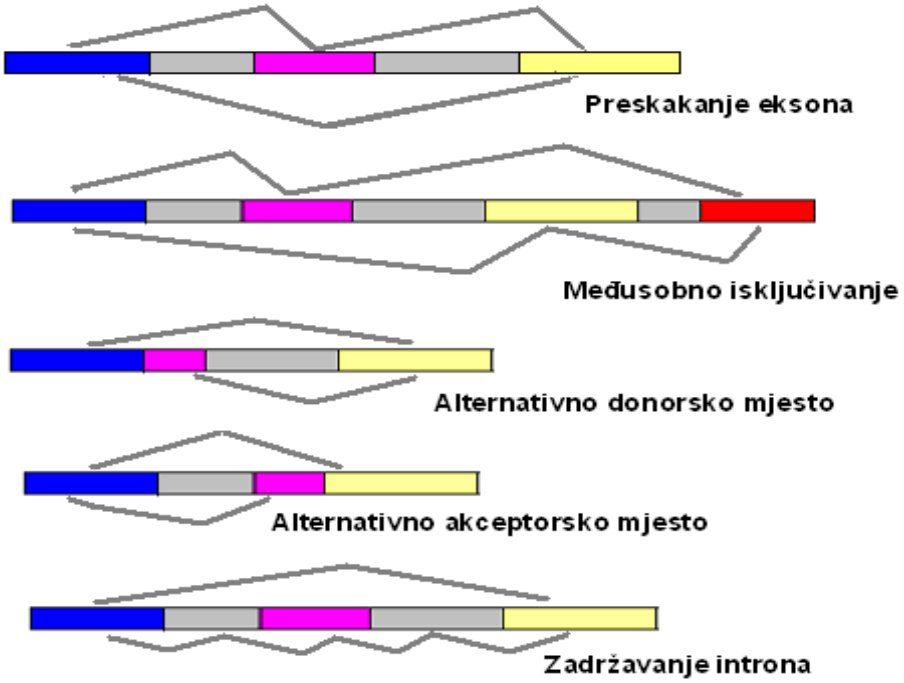
4. Alternativno akceptorsko mjesto:

   Ekson se pridružuje eksonu koji mu slijedi i mijenja njegov 5' kraj.

5. Zadržavanje introna:

   U rezultirajućoj mRNA može ostati intron.

Svaki od navedenih modela prikazan je na slici 2.2, sivom bojom predstavljeni su introni. Povezani linije povezuju eksone u mRNA.



Slika 2.2 Modeli alternativnog spajanja eksona

3.2 Alati i metode

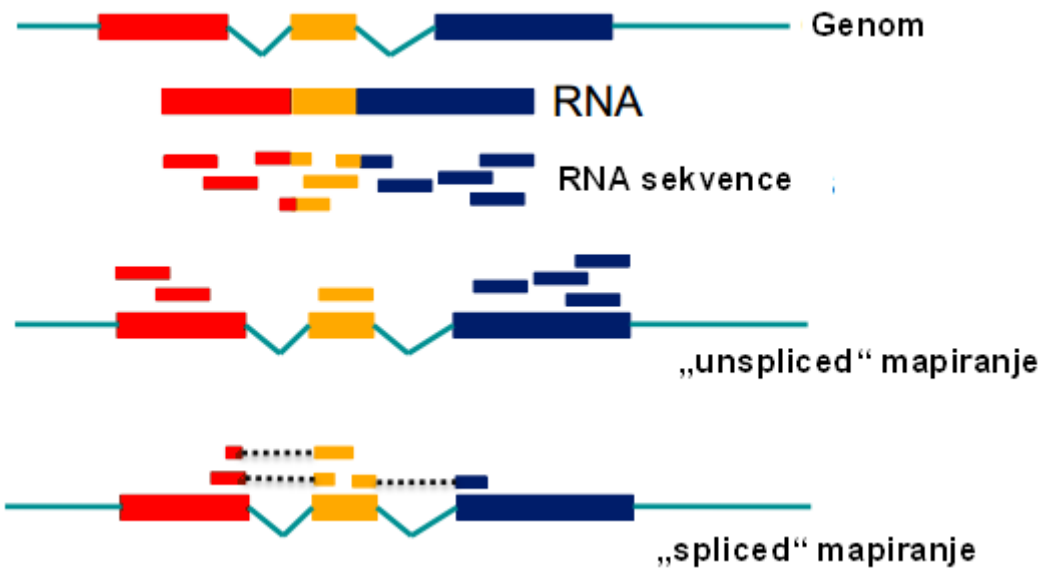
Algoritam za otkrivanje je li na zadanim skupovima RNA sekvenci došlo do alternativnog spajanja eksona je sljedeći :

1. Mapirati dobivene sekvence na genom (poglavlje 3).
2. Rekonstrukcija u transkripate.
3. Analiza transkripata.

U nastavku su opisani alati i metode za svaki korak.

3.2.1 Mapiranje

Mapiranje je proces određivanja položaja RNA sekvence u genomu. Alati za mapiranje dijele se na „unspliced“ mapiranje i „spliced“ mapiranje. „Splice“ mapiranje dopušta mapiranje RNA sekvenci preko introna. Na slici 2.3 prikazano je „unspliced“ i „spliced“ mapiranje, plavom linijom označeni su introni, dok obojani pravokutnici označavaju eksone.



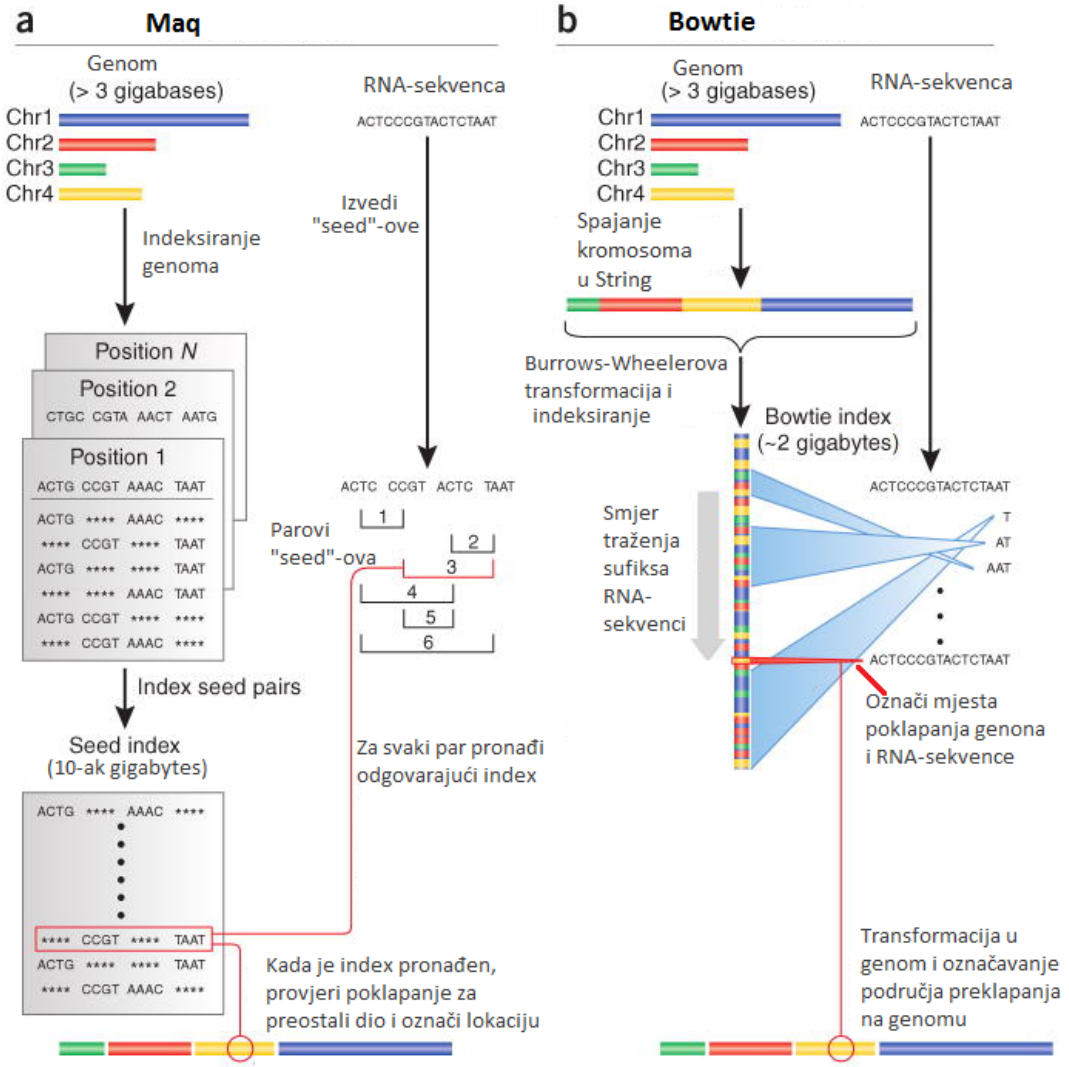
Slika 2.3 Mapiranje RNA sekvenci

Jedni od najpoznatijih alata za „unspliced“ poravnavanje su: Bowtie i Maq.

Oba koriste tehnike raspršenog adresiranja, dijelovima genoma pridružuju indeks, i na taj način povećavaju brzinu poravnavanja.

Maq (http://maq.sourceforge.net/) dijeli ulazne sekvence na četiri dijela jednake duljine („seed“), vođen idejom da se sekvenca može poravnati na genom ako se svaki dio može poravnati. Ukoliko postoji jedno neslaganje između sekvence i genoma, tada se samo jedan od četiri dijela neće moći poravnati. Ukoliko postoje dva mjesta neslaganja između sekvence i genome, najviše se dva dijela neće moći poravnati. Budući da su dopuštene dvije pogreške prilikom poravnavanja, Maq algoritam stvara parove „seed“-ova i svaki par poravnava na genom.Nakon pronalaska indeksa za odgovarajući par „seed“-ova, potrebno je provjeriti preklapanje druga dva „seed“-a. Algoritam Maq prikazan je na slici 2.4.a).

Bowtie (http://bowtie-bio.sourceforge.net/index.shtml) koristi tehniku sažimanja podataka nad ogromnim genomom i nad tako sažetim podacima provodi mapiranje. Za saživanje genoma koristi Burrows-Wheelerovu transformaciju. Mapiranje se provodi iterativno, čitanjem po jedne baze iz ulazne sekvence (sufiks). Početno stanje je lista pozicija na koje se sekvenca može mapirati i predstavlja čitavi genom. U svakoj iteraciji traženjem znaka u listi pozicija, određuje se nova lista pozicija koja može pokriti i taj znak. Ukoliko u nekom trenutku ne postoji niti jedan pozicija na koju bi se sekvenca mogla mapirati, vraća se u prethodnu iteraciju, mijenja bazu i kreće ponovno. Bowtie je prikazan na slici 2.4. b)



2.4 Maq i Bowtie

Najpoznatiji alat za „spliced“ mapiranje je TopHat (http://tophat.cbcb.umd.edu/). TopHat koristi alat Bowtie koji mapira RNA-sekvence na genome, i tako određuje eksone, a preostale RNA-sekvence poravnava između eksona [2].

3.2.2 Rekonstrukcija u transkripte

Rekonstrukcija u transkripte je postupak određivanja svih mogućih transkripata iz zadanih RNA-sekvenci te određivanja uključenosti pojedine RNA-sekvence u pojedinom transkriptu.

Rekonstrukcija u transkripte je zahtjevan proces iz više razloga. Jedna od najzahtjevnijih stvari jest odrediti uključenost RNA-sekvence u pojedinom transkriptu [3].

Postoji više metoda koje provode rekonstrukciju u transkripte, a one mogu biti ovisne ili neovisne o genomu.

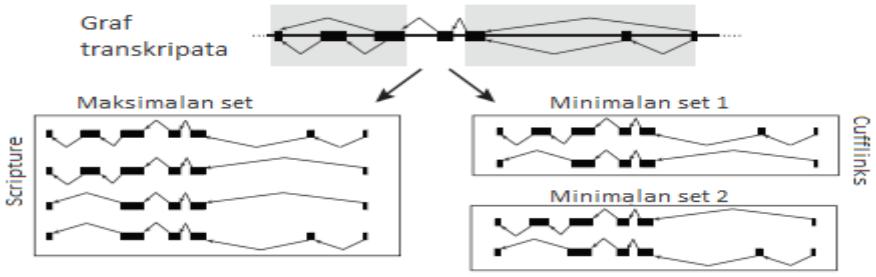
Metode ovisne o genomu, prije rekonstruckije, provode mapiranje RNA-sekvenci na genom. Kako bi dobili sve moguće transkripte koji mogu nastati iz dobivenih RNA-sekvenci, potrebno je napraviti uniju nad rezultatima mapiranja.

Dok metode ovisne o genome koriste rezultate mapiranje, metode neovisne o genomu provode uniju nad RNA-sekvencama i na taj način grade transkripte.

Najpoznatije metode koje ovise o genomu su: Cufflinks (http://cufflinks.cbcb.umd.edu/) i Scripture (http://www.broadinstitute.org/software/scripture/).

I Cufflinks i Scripture svoj rad temelje na transformaciji skupa RNA-sekvenci u graf koji sadrži sve moguće veze između eksona. Transformacijom skupa RNA-sekvenci u graf, bit problema više nije rekonstrukcija transkripta već statistika. Prilikom stvaranja veza između eksona u grafu, u obzir se uzimaju veze koje povezuju uzastopne eksone i veze koje stvaraju dijelovi razdvojene RNA-sekvence (razdvojene RNA-sekvence nastaju kao rezultat „spliced“ mapiranja, jasnije na slici 2.3).

Razlika između Cufflinka i Scripture je u njihovoj interpretaciji dobivenog grafa. Parsiranjem dobivenog grafa grade se transkripti. Prilikom gradnje transkripata, Cufflinka stavlja naglasak na maksimalnu točnost i gradi minimalan broj transkripata, dok Scripture stavlja naglasak na maksimalnu osjetljivost i gradi sve moguće transkripte. Dobiveni setovi za Cufflinks i Scripture prikazani su na slici 2.5.

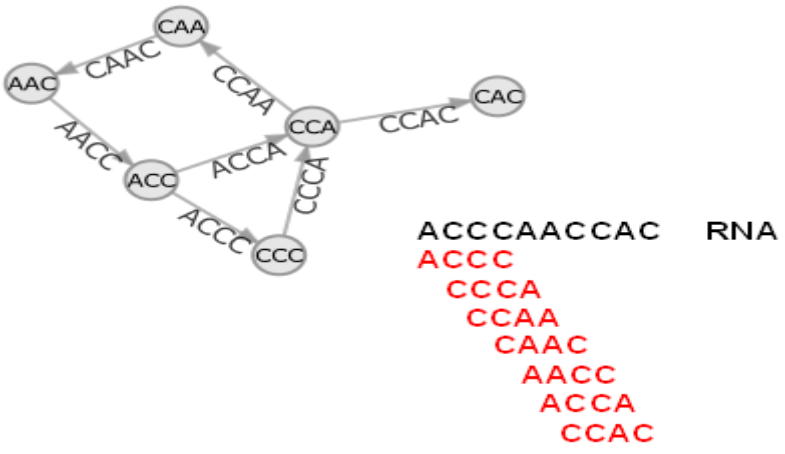


Slika 2.5 Setovi nastali iz grafa

Cufflinksa gradi skup minimalnog broj[1]. Kako postoji više skupova minimalnog broja, Cufflinksa bira samo jedan od njih koristeći se statistikom, bira onaj skup koji ima najveću vjerojatnost pojavljivanja. Za svaki transkript računa se pokrivenost RNA-sekvencama i na taj način određuje vjerojatnost pojavljivanja.

Metode koje ne ovise o genome grade transkripe preklapanjem ulaznih RNA-sekvenci. Primjer metode koja ne ovisi o genome jest transAbyss (http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss). Najpoznatija strategija za preklapanje ulaznih RNA sekvenci jest izgraditi de Bruijn graf. De Bruijn graf dijeli sekvence na na subsekvence duljine k, a preklapanjanjem k – 1 baze svih subsekvenci dobiva se graf svih mogućih sekvenci koje se mogu izgraditi [4]. Primjer de Bruijnog grafa je na slici 2.6, gdje je k = 4.

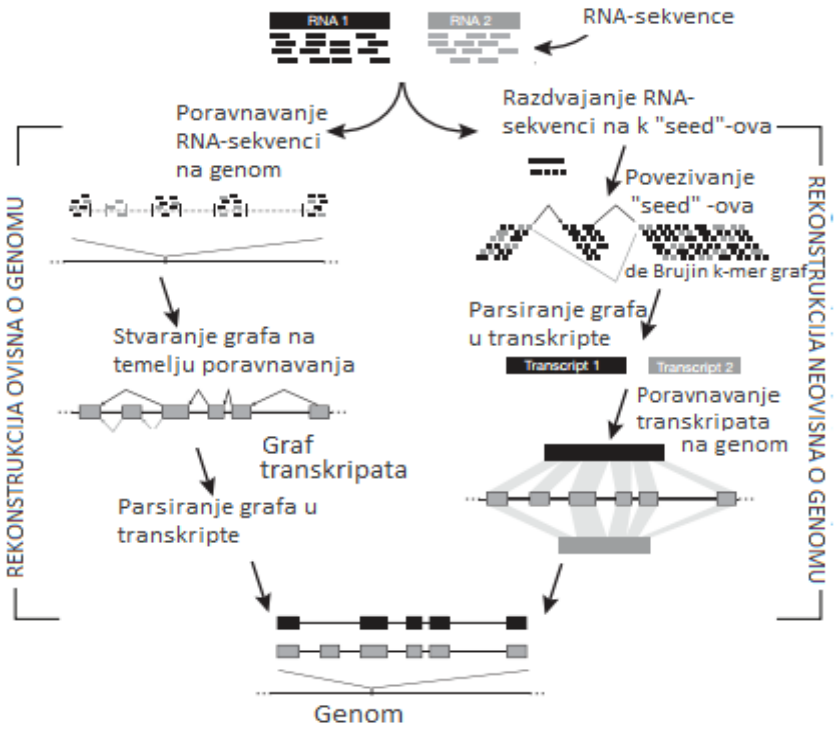
1- Skup minimalnog broja - skup koji se sastoji od minimalnog broja transkripata uz uvjet da su sve RNA- sekvence uključene u barem jedan transkript



Slika 2.6 de Bruijn graf

Kada su ulazne sekvence prevedene u de Bruijin graf, eliminiraju se oni putovi koji nisu pokriveni ulaznim RNA-sekvencama. Parsiranjem grafa dobijemo transkripte. Daljnjim poravnavanjem dobivenih transkripata na genom stvaraju se transkripti u željenom obliku (povezani eksoni).

Oba tipa metoda za rekonstrukciju transkripata prikazane su na slici 2.7.



Slika 2.7 Rekonstrukcije transkripata

3.2.3 Analiza transkripata

Prije donošenja zaključka je li za određeni ulaz došlo do alternativnog spajanja eksona, mora se provesti analiza nad dobivenim transkriptima.

Ulaz u analizator transkripata su transkripti dobiveni iz dva različita skupa RNA-sekvenci. Jedan skup predstavlja RNA sekvence u takozvanim „normalnim“ uvjetima, dok drugi skup predstavlja RNA sekvence u uvjetima u kojima bi trebalo doći do alternativnog spajanja eksona.

Metode za analiziranje transkripata služe se različitim mjerama. RPKM[2] je najčešće korištena mjera za usporedbu mRNA. RPKM računamo prema formuli 2.1, gdje g predstavlja gen, r_g označava broj RNA-sekvenci mapirani na gen, fl_g označava duljinu gena (broj nukleotida u mapiranom[3] dijelu gena), dok se R računa prema formuli 2.2 [5].

$$RPKM_g = (r_g * 10^9) / (fl_g * R) \qquad (2.1)$$

$$R = \sum_{g \in G} r_g \qquad (2.2)$$

Često korištena mjera je i FPKM[4]. FPKM definira se formulama 2.3 i 2.4, gdje je f_g broj fragmenata mapiranih na gen, fl_g označava duljinu gena, dok F označava ukupan broj fragmenata u eksperimetnu. Slično kao i RPKM, ali se umjesto RNA-sekvenci koriste se fragmenti.

$$FPKM_g = (f_g * 10^9) / (fl_g * F) \qquad (2.3)$$

$$F = \sum_{g \in G} f_g \qquad (2.4)$$

Primjeri alata za analizu transkripata su CuffCompare (http://cufflinks.cbcb.umd.edu/manual.html) i Alexa-seq (http://www.alexaplatform.org/alexa_seq/).

2 - reads per kilobase per million
3 - Mapirani dio gena- dio gena oduhvaćen RNA-sekvencama
4 - fragments per kilobase of exon per million fragments mapped

3.3 Alati za otkrivanje alternativnog spajanja eksona

Postoje brojni alati za otkrivanje alternativnog spajanja eksona, temeljeni na dva različita pristupa. Postoji :

1. Pristup temeljen na eksonima te
2. Pristup temeljen na transkriptima.

Pristup temeljen na eksonima promatra vjerojatnost pojavljivanja svakog pojedinog eksona u skupovima. Na temelju te vjerojatnosti donose se zaključci.

Pristup temeljen na transkripta promatra transkripte koji su izgrađeni od čistih RNA-sekvenci. Za svaki skup RNA-sekvenci grade se transkripti i temeljem razlika u transkriptima donose se zaključci.

U prethodnom poglavlju opisan je Cufflinks, s Cufflinksom dolaze alati CuffCompare i CuffDiff koji zajedno otkrivaju alterniranje eksona koristeći pristup temeljen na transkriptima.

U nastavku je opisan jedan od najznačajnijih alata, MATS (http://rnaseq-mats.sourceforge.net/).

3.3.1 MATS

MATS je jedan od najznačajnijih alata za otkrivanje alternativnog spajanja eksona, a pristup je temeljen na eksonima.

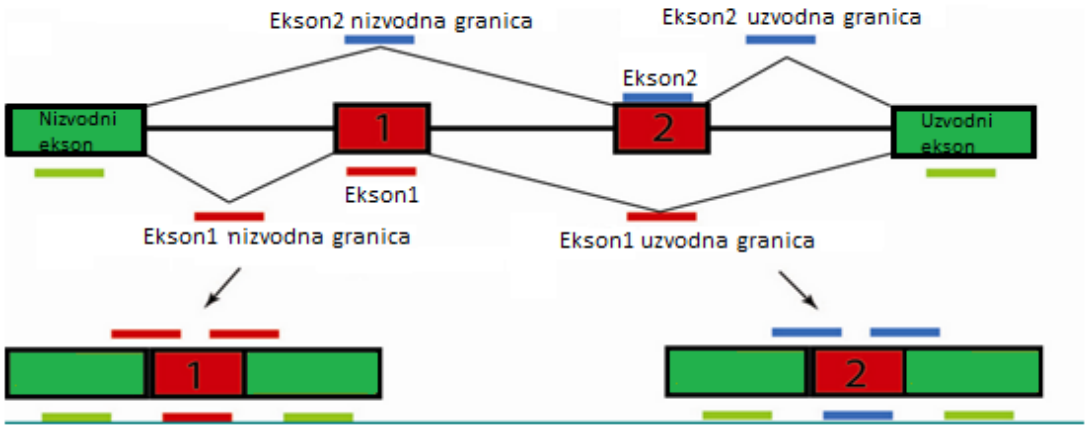
MATS je alat koji za detekciju alterniranja koristi različite izračune, u dodatku A može se naći kratak opis statističkih pojmova koje MATS koristi i koji olakšavaju razumijevanje algoritma.

MATS algoritam kao ulaz prima dva skupa RNA-sekvenci (čistih RNA-sekvenci ili rezultate mapiranja RNA-sekvenci), i njih podvrgava nul hipotezi (H_0) . Alternativna hipoteze (H_1) – došlo je do alternativnog spajanja eksona. Na ulazu prima i dodatni parametar c.

MATS algoritam je sljedeći :

1. Ulaz: 2 skupa RNA-sekvenci (.fastq ili .bam format) i genom
2. Za svaki ekson u genomu izračunaj razinu uključenosti eksona u oba skupa
3. Određuje se a priori razdioba razina uključenosti u oba skupa.
4. Računa se a posteriori vjerojatnost alternativne hipoteze.
5. Računa se p-vrijednost i FDR (False Discovery Rate) i na temelju toga određuje je li došlo do alternativnog spajanja eksona.

Uključenost eksona definira se kao ukupan broj RNA-sekvenci mapiranih na taj ekson. RNA-sekvence mogu biti mapirane na granicu tog eksona i njegovog nizvodni eksona, uzvodnog eksona ili na granicu njegovog nizvodnog i uzvodnog eksona isključujući njega (slika 2.8). Mapiranje na granici eksona naziva se „spliced“ mapiranje i objašnjeno je u poglavlju 2.2.1.



Slika 2.8 Mapiranje RNA-sekvenci

Razina uključenosti eksona (oznaka Ψ) definirana je kao postotak uključenosti eksona, a računa se prema formuli 2.3, gdje su :

- UJC = broj RNA-sekvenci mapiranih na granici eksona i njegovog uzvodnog eksona,
- DJC = broj RNA-sekvenci mapiranih na granici eksona i njegovog nizvodnog eksona te

- SJC = broj RNA-sekvenci mapiranih na granici njegovog nizvodnog i uzvodnog eksona.

$$\Psi = \frac{\frac{UJC+DJC}{2}}{\frac{UJC+DJC}{2}+SJC} \quad (2.3)$$

Uključenosti eksona podvrgava se binomnoj razdiobi sa parametrima n = (I+ S) i p = Ψ, gdje su I i S definirani formulom 2.4.

$$I = \frac{(UJC+DJC)}{2}$$
$$S = SJC \quad (2.4)$$

MATS za mapiranje koristi Tophat.

Nakon što su izračunate razine uključenosti eksona u oba skupa za sve eksone (oznake Ψ_1 i Ψ_2 za svaki pojedini ekson), Ψ_1 i Ψ_2 se podvrgavaju nul hipotezi. Nul hipoteza se prihvaća ukoliko vrijedi 2.5, gdje je c ulazni parametar kojeg definira korisnik.

$$| \Psi 1 - \Psi 2 | \leq c \quad (2.5)$$

Da bi mogli izračunati a posteriori vjerojatnost alternativne hipoteze (tj. P(| Ψ_1 - Ψ_2 | > c | Podaci)), prema Bayesovoj formuli potrebno je izračunati a priori vjerojatnost i izglednost.

MATS definira a priori vjerojatnost kao dvodimenzijsku razdiobu između Ψ_1 i Ψ_2 , kod koje su marginale razdiobe varijabli Ψ_1 i Ψ_2 uniforme radiobe na intervalu [0, 1], a njihova ovisnosti također je podvrgnuta uniformnoj razdiobi na intervalu [0,1]. MATS je jedini alat koji u obzir uzima ovisnost jednog eksona u jednom uzorku o istom eksonu u drugom uzorku.

A posteriori vjerojatnost hipoteze H_1 za ekson *i* definira se kao $P_i = P(|\ \Psi_{i1} - \Psi_{i2}\ | > c \mid I_{i1}, S_{i1}, I_{i2}, S_{i2}, I_{-i1}, S_{-i1}, I_{-i2}, S_{-i2})$, gdje je *–i* oznaka za sve ostale. MATS za izračun a posteriori vjerojatnosti koristi simulaciju Monte Carlo Markovljevih lanaca, točnije JAGS program (Just Another Gibbs Sampler). JAGS program dodatno računa i parametar *ρ,* koji određuje ukupnu ovisnost razina uključenosti eksona svih alterativno spojenih eksona.

MATS za izračun P-vrijednost koristi sljedeći algoritam:

Za svaki ekson *i*:

Odredi Ψ_{i1}^{c} i Ψ_{i2}^{c} prema :

$$(\Psi_{i1}^{c}\ ,\ \Psi_{i2}^{c}\ ) = \arg\max f()$$

$$f() = (I_1 \log \Psi_{i1} + S_1 \log (1 - \Psi_{i1}\ ) + I_2 \log \Psi_{i2} + S_2 \log (1 - \Psi_{i2}\ ))$$

Dohvati parametar *ρ.*

Za j = 1, …,M:

1) generiraj podatke (I_{i1j}, S_{i1j}, I_{i2j}, S_{i2j}):

$$I_{i1j} \sim \text{binomnaRazdioba}(n = I_{i1j} + S_{i1j,}\ p = \Psi_{i1}^{c}\ )$$

$$S_{i1j} = n - I_{i1j}$$

$$I_{i2j} \sim \text{binomnaRazdioba}(n = I_{i2j} + S_{i2j,}\ p = \Psi_{i2}^{c}\ )$$

$$S_{i2j} = n - I_{i2j}$$

2) Računaj a posteriori vjerojatnost koristeći simulaciju MonteCarlo Markovljevi lanci:

$$P_{ij}^{\ sim} = P(|\ \Psi_{i1j} - \Psi_{i2j}\ | > c \mid I_{i1j}, S_{i1j}, I_{i2j}, S_{i2j}, \rho)$$

3) Izračunaj P-vrijednost kao: $(\sum_{j = 1, M} I(P_i \leq P_{ij}^{\ sim})\ ) / M$

Parametar M određuje preciznost p-vrijednosti. Ukoliko se želi postići preciznost 0.01 za P-vrijednost, parametar M tada je jednak 100. Ukoliko P-vrijednost bude 0 ili jako blizu 0, tada je vrijednost parametra M nedovoljno velika da bi se mogla procijeniti P-vrijednost. Za sve eksone za koje je P-

vrijednost manja od trostruke preciznosti (u prethodnom primjeru to je 0.03), parametar M se mijenja za 10^{-1} puta (u prethodnom primjeru to je 0.001) i ponavlja se postupak.

Kada su izračunate P-vrijednosti za sve eksone, pomoću Benjamini-Hochberovog modela provodi se postupak računanja FDR.

4 Podaci

Ulazni podaci mogu biti u različitim formatima, prikazane s različitom točnošću. U nastavku su opisani samo neki formati za čuvanje podataka.

4.1 Fasta

Fasta format je tekstualni prikaz sekvenci gdje su nukleotidi predstavljeni slovima abecede. Fasta format prikazan je na slici 3.1. Prva linija je opisna linija, a od ostalih se razlikuje početnim znakom '>'. Svako slovo ima pridruženo značenje, tako npr. slovo M označava ili A ili C.

```
>SEQUENCE_1
MTEITAAUTGMVKELRESTGAMMDCKNALSETNGDAFDKAVQLLREKGLG
KAAAAKADRLAAEGLVSKKVSDDFFTAAGUUTTAAMRPSYEDLDMTEVEN
EYKALVAELEKENEERRRLDKDPNKPEHKIPQADNSQQLDSKLTLRRLAA
```

Slika 3.1 Fasta format

4.2 Fastq

Fastq je složeniji format od Fasta. Uz tekstualni prikaz nukleotida, sadrži kvalitetu, koja je također prikazana u tekstualnom obliku. Linija s početnim znakom '@' je opisna linija. Nakon nje slijedi sekvenca. Linija s početnim znakom '+' je dodatna linija u kojoj može stajati dodatan opis sekvence. Posljednja linija je linija kvalitete sekvence. Fastq format prikazan je na slici 3.2.

```
@SEQ_ID
GATTTAGGTAACCGATTTAGTACAGTTACAGTAAAGGGGTTACCCTACG
+
!'*(((((+++*))%%%++55CCF>>>>>>CCC56
```

Slika 3.2 Fastq format

4.3 SAM

SAM format služi za pohranu rezultata mapiranja ili poravnavanja. Primjer SAM formata prikazan je na slici 3.3 b), pridružen mapiranju RNA-sekvenci koje je prikazano na 3.3.a). Linije započete znakom '@' su opisne linije. Linije koje sadrže podatke podijeljene su na 11 obaveznih dijelova, a mogu sadržavati i više dodatnih dijelova.

Mapiranje

```
coor   12345678901234  5678901234567890123456789012345
ref    AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

r001+        TTAGATAAAGGATA*CTG
r002+       aaaAGATAA*GGATA
r003+     ~~gcct~~aAGCTAA
r004+                  ATAGCT..............TCAGC
r003-                        ~~ttagct~~TAGGC
r001-                                     CAGCGCCAT
```

SAM format

```
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTA *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

Slika 3.3 SAM format

Značenja pojedinih dijelova su slijedeća:

1. Ime sekvence.
2. Vrijednost zastavice:

   Vrijednost je prikazana u dekadskom obliku. Broj se pretvara u binarni broj duljine 11 i svakom bitu se pridružuje vrijednost (počinje od bita najmanje težine) :

- RNA-sekvenca ima pridružen par
- Pravilno mapiranje para RNA-sekvenci (ovisi o protokolu)
- RNA-sekvenca nije mapirana
- Par RNA-sekvence nije mapiran
- RNA-sekvenca je negativna.
- Par RNA-sekvence je negativan.
- RNA-sekvenca je prva sekvenca u paru
- RNA-sekvence je druga sekvenca u paru
- Mapiranje nije primarno
- RNA-sekvenca pada na provjeri kvalitete
- RNA sekvenca je kopija

3. Ime uzorka referentne RNA/DNA.
4. Pozicija na referentnoj RNA/DNA na koju se preklapa prvi lijevi nukleotid RNA-sekvence.
5. Kvaliteta mapiranja.
6. CIGAR vrijednost

   Sastoji se od niza parova oblika vrijednost:oznaka.

   Vrijednost predstavlja broj nukleotida koji su zahvačeni oznakom. Postoji više oznaka:

   - M – nukleotidi su mapirani na referentni RNA/DNA
   - I – nukleotidi su višak u odnosu na referentnu RNA/DNA
   - D – nukleotidi su manjak u odnosu na referentnu RNA/DNA
   - N – broj preskočenih nukleotida (kod „spliced“ mapiranja)
   - S – izrezani nukleotidi, ali prikazani u RNA-sekvenci (takvi nukleotidi prikazani malim slovima)
   - H – izrezani nukleotidi, nisu prikazni u RNA-sekvenci (prekriženi)
   - P – ne postojanje nukleotida u RNA-sekvenci, ali mapirani na umetnute nukleotide u RNA/DNA

7. Naziv uzoraka RNA na koji se poravnava par RNA -sekvence (* - nepoznato, = - isti).

8. Pozicija uzorka RNA na koji se poravnava par RNA-sekvence (0 ukoliko sekvenca nema para).
9. Udaljenost između pozicija mapiranja krajnjeg desnog nukeotida RNA-sekvence i krajnjeg lijevog nukleotida para RNA-sekvence (razlike između parova označene sa +/-). 0 ukoliko RNA-sekvenca nema para.
10. RNA-sekvenca. Ako sekvenca nije pohranjena, oznaka '*'.
11. Kvaliteta sekvence (isto kao i kod FASTQ formata). Oznaka '*' ako kvaliteta nije pohranjena.

Primjer :

Na slici 3.3 promatramo 2 liniju.

r00 je naziv RNA-sekvence.

163 je vrijednost zastavice. Pretvorba 163 u binarni broj je : 00010100011.

ref je ime uzorka RNA/DNA na koju radimo mapiranje.

7 je pozicija na koju se preklapa prvi lijevi nukleotid .

30 je kvaliteta mapiranja.

8M2I4M1D3M – CIGAR vrijednost (prvih 8 nukleotida mapirano na RNA/DNA, sljedeća 2 su umetnuta, sljedeća 4 su mapirana, jedan nukleotid nedostaje, posljednja 3 su mapirana).

= označava da se par od r001 mapira na istu RNA/DNA.

37 – pozicija na uzorku RNA/DNA na koju se mapira prvi lijevi nukleoid.

39 – udaljenost između r001 i para r001.

TTAGATAAAGGATACTA – promatrana RNA-sekvenca.

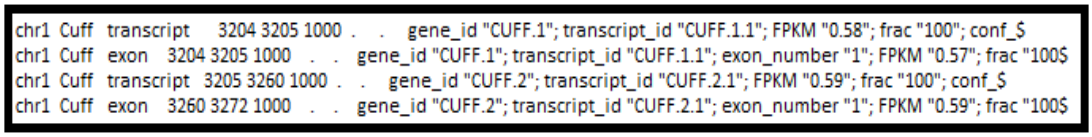
'*' – kvaliteta nije pohranjena.

4.4 BAM

BAM format sadrži iste informacije kao i SAM format, ali u drugom obliku. BAM format je binarna, sažeta verzija SAM formata. A pretvorba u SAM format provodi se korištenjem alata SAMTools.

4.5 GTF

GTF najčešći je format za pohranu transkripata. Primjer GTF formata prokazan je na slici 3.4. Svaka linija podijeljena je na 9 dijelova odvojenih TAB-om. Sadrži redom: ime kromosoma, ime alata kojim je generirano, tip podatka, početna pozicija, krajnja pozicija, kvaliteta, negativno/pozitivno, i dodatne atribute.



Slika 3.4 GTF format

5 Implementacija

U nastavku je opisana implementacija alata za otkrivanje eksona koristeći pristup temeljen na transkriptima.

Algoritam:

1. Ulaz su dva skupa čistih sekvenci.
2. Alatom TopHat/Bowtie provedeno je mapiranje za oba skupa.
3. Alatom Cufflinks provedena je rekonstukcija u transkripte.
4. Implementiran je alat za analizu rezultata.

Izlaz Cufflinksa je datoteka .GTF formata. Algoritam za analizu rezultata je slijedeći:

```
Za svaki kromosom:
    U skupu1:
        odredi sve transkripte
        za svaki transkript:
            odredi početak i kraj
            odredi sve eksone
            za svaki ekson:
                odredi početak i kraj
    U skupu2:
        Pronađi transkript
        Odredi početak i kraj
        Ako u skupu1 postoji transkript sa istim početkom i krajem:
```

```
                Pronađi ekson
                Ako ne postoji isti ekson u skupu1:
                        Zabilježi mjesto alternativih transkripata.

            Inače:
                        Zabilježi mjesto alternativih transkripata.
        Ako postoji u skupu1 transkript ili ekson koji nije bio pokriven:
                Zabilježi mjesto alternativih transkripata.

        Za zadani kromosom pronađi sve obilježene gene.

    Za sve gene:
        Ako na genu postoji zabilježeni transkript:
                Na genu je došlo do alternativnog spajanja eksona.
```

Implementacija je napravljena u programskom jeziku Python.

6 Analiza rezultata

Napravljena je usporedna analiza dobivenih rezultata, sa rezultatima koje daje alat CuffDiff.

CuffDiff je dio paketa CuffLinks, koji uz gradnju transkripata pruža mogućnost otkrivanja alternativnog spajanja eksona.

Ulaz:

Dva skupa RNA-sekvenci miša – prvi skup dobiven kada su neuralne matične stanice tretirane etanolom, drugi skup dobiven je kada su neuralne matične stanice tretirane OHT-om.

Genom miša (mm9) – preuzet iz NCBI baze podataka.

Analiza:

Rezultati analize prikazani su u tablici 6.1. Razmatrana su tri različita slučaja:

1. Pozitivno podudaranje – alternirani geni pronađeni koristeći oba alata.

2. Negativno podudaranje – alternirani geni pronađeni koristeći alata CuffDiff, a nisu pronađeni implementiranim alatom.
3. Dodatno – alternirani geni pronađeni koristeći implementirani alat, a nisu pronađeni CuffDiff-om.

	Broj gena
Pozitivno podudaranje	33447
Negativno podudaranje	4224
Dodatno	392

Tablica 6.1 Rezultati

7 Zaključak

Alternativno spajanje eksona je proces do kojeg dolazi prilikom procesiranja DNA. To je važan proces koji stvara mogućnost nastanka više različitih proteina iz jednog gena. Postoji više oblika alternativnog spajanja eksona. Veliki je izazov današnjice stvoriti kontrolu nad gradnjom života, pa tako postoje brojni alati koji bi mogli omogućiti upravo to. U sklopu ovog rada, opisani su brojni alati koji mogu pomoći u otkrivanju alternativnog spajanja eksona. Opisani su alati za mapiranje sekvenci, izgradnju i analizu transkripata. Također je opisan jedan od najznačajnijih alata za otkrivanje alternativnog spajanja eksona, MATS. Navedeni su česti formati podataka i kratak opis svakog. Implementiran je jednostavniji alat za otkrivanje alternativnog spajanja eksona na razini transkripata.

Usporednom analizom rezultata implementiranog alata sa već postojećim alatima, ustanovljeno da je implementirani alat manje precizan. Postoji više mogućnosti dorade kojim bi implementirani alat postao precizniji.
Otkrivanje alternativnog spajanja eksona samo je jedno od brojnih područja kojima se bavi bioinformatika. To je relativno mlado područje na kojem se mora još puno raditi.

8 Literatura

[1] Zahler Alan M., Alternative splicing in C.elegans

[2] Trapnell C i Salzberg S.L, How to map billions of short reads onto genomes

[3] Garber M, Grabherr M. G., Guttman M. i Trapnell C., Computational methods for transcriptome annotation and quantification using RNA-seq

[4] http://gcat.davidson.edu/phast/debruijn.html, posjećeno : 03.svibanj 2013.

[5] Wager P.G., Kin K. i Lynch V.J., Measurement of mRNA abundance using RNA-seq dana: RPKM measure is inconsistent among samples

[6] N. Elezović, Statistika i procesi, 2010.

9 Sažetak

Alternativno spajanje eksona je proces do kojeg dolazi prilikom procesiranja DNA. To je važan proces koji stvara mogućnost nastanka više različitih proteina iz jednog gena. U ovom radu opisan je proces alternativnog spajanja eksona. Slijedeći jedan od mogućih algoritama za otkrivanje alternativnog spajanja eksona, opisan je svaki pojedini korak, mapiranje, gradnja transkripata te analiza transkripata. Uz svaki korak navedeni su alati koji mogu pomoći. Opisana je implementacija alata za otkrivanje alternativnog spajanja eksona, te je napravljena usporedna analiza sa već implementiranim alatima.

10 Abstact

Alternative splicing is an important process that allows individual genes to produce multiple protein isoforms. In this paper, we describe the process of alternative splicing. Following one of the approaches for detection of the alternative splicing event, we describe different tools that can help. We also implement a tool for detection of the alternative splicing event and make comparative analysis with already implemented tools.

11 Dodatak A

Dodatak A sadrži kratak uvod u vjerojatnost i statistiku koji olakšava razumijevanje alata MATS.

Bayesova formula

Bayesovom formulom računa se vjerojatnost ostvarivanja hipoteze H, ako znamo da vrijede činjenice A (7.1).

$$p(H|A) = \frac{p(A|H)*p(H)}{p(A)} \qquad (7.1)$$

Nazivlja vezana uz Bayesovu formulu su sljedeća:

- P(H) je a priori vjerojatnost hipoteze H
- P(H|A) je a posteriori vjerojatnost hipoteze H
- P(A|H) je izglednost vjerojatnost hipoteze H

Uniforma razdioba

Kontinuirana slučajna varijabla X ima uniformu razdiobu na intervalu [a,b], ako je funkcija gustoće vjerojatnosti sljedeća:

$$f(x) = \begin{cases} \frac{1}{b-a} & za\ a \le x\ \le b, \\ 0 & za\ x < a, x > b \end{cases}$$

Binomna razdioba

Slučajna varijabla X ima binomnu razdiobu s parametrima n i p, $X \sim B(n, p)$, ako X mjeri broj ponavljanja događaja A, a p je vjerojatnost realizacije događaja A, n je broj ponavljanja pokusa.

Izračun vjerojatnosti da se realizirao događaj {X = k} prikazan je u 7.2. Realizacija događaja {X = k} predstavlja slijedeće :

- u n pokusa, događaj A se ostvario točno k puta.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad (7.2)$$

Dvodimenzionalna razdioba

Dvodimenzionalna razdioba je razdioba dvodimenzionalnog slučajnog vektora. Funkcija razdiobe definirana se formulom 7.3.

$$F(x, y) := \mathbf{P}(X < x, Y < y) \qquad (7.3)$$

Uz razdiobe vektora veže se pojam marginalne razdiobe. Marginalna razdioba varijable X opisana je formulom 7.4.

$$F_X(x) = F(X \leq x; -\infty < Y < \infty) = F(x, \infty) \qquad (7.4)$$

Nul hipoteza (H_0)

Nul hipoteza je pretpostavka da za neko svojstvo nema razlike između danih skupova. Nul hipotezu odbacujemo ukoliko se temeljem statističkih podataka utvrdi značajna razlika između skupova. Ukoliko se nul hipoteza može odbaciti, tada vrijedi alternativna hipoteza (H_1).

Za nul hipotezu veže se P-vrijednost . Na temelju P-vrijednosti, koja je statistički podatak, određuje se valjanost hul hipoteze. Ukoliko je P-vrijednost mala, vjerojatnost da podaci koje razmatramo potvrđuju nul-hipotezu tada je mala i hipoteza se može odbaciti. Razina značajnosti (α) određuje graničnu P-vrijednost, ukoliko je P-vrijednost manja od α tada se nul-hipoteza može odbaciti.

Prilikom razmatranja nul hipoteze mogu nastati četiri različita zaključka :

- Nul hipoteza vrijedi, prihvaćena je,
- Nul hipoteza ne vrijedi, prihvaćena je (greška tipa II),
- Nul hipoteza vrijedi, odbijena je (greška tipa I) te
- Nul hipoteza ne vrijedi, odbijena je.

Stopa pogreška prve vrste (FDR) definirana je sa 7.5, gdje je:

- V = ukupan broj grešaka tipa I,
- R = ukupan broj odbijenih nul hipoteza.

$$FDR = Qe = E[Q] = E[\frac{V}{R}] \qquad (7.5)$$

Ukoliko je R = 0, tada je FDR = 0.

Markovljev lanac

Lanac predstavlja niz slučajnih varijabli $\{X_1, X_2, \ldots\}$, taj lanac je Markovljev, ukoliko za sve izbore stanja $i_1, \ldots, i_n$ vrijedi 7.6, tj. buduće stanje ovisi samo o sadašnjem stanju [6].

$$\boldsymbol{P}(X_{n+1}=i_{n+1} \mid X_n=i_n, \ldots, X_0=i_0) = \boldsymbol{P}(X_{n+1}=i_{n+1} \mid X_n=i_n) \qquad (7.6)$$

Veza između slučanih varijabli $\{X_1, X_2, \ldots\}$ zadana je prijelaznim vjerojatnostima.