SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 6679

VIZUALIZACIJA PODATAKA DOBIVENIH SEKVENCIRANJEM RNA

Jan Vrlec

Zagreb, lipanj 2020.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 6679

VIZUALIZACIJA PODATAKA DOBIVENIH SEKVENCIRANJEM RNA

Jan Vrlec

Zagreb, lipanj 2020.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 13. ožujka 2020.

ZAVRŠNI ZADATAK br. 6679

Pristupnik:	**Jan Vrlec (0036495516)**
Studij:	Računarstvo
Modul:	Programsko inženjerstvo i informacijski sustavi
Mentor:	doc. dr. sc. Krešimir Križanović

Zadatak: **Vizualizacija podataka dobivenih sekvenciranjem RNA**

Opis zadatka:

Današnji uređaji za sekvenciranje nisu u stanju sekvencirati cijele genome, već samo pročitati manje dijelove (očitanja) i pri tome unose određenu količinu pogreške. Jedan od prvih i osnovni zadataka prilikom analize podataka dobivenih sekvenciranjem jest mapiranje očitanja na referentni genom ili transkriptom. Zbog velike količine podataka, nije moguće ručno analizirati mapiranje svakog pojedinog očitanja, već je puno praktičnije koristiti prikladne alate za vizualizaciju. Potrebno je napisati aplikaciju za vizualizaciju mapiranja RNA očitanja na referentni genom. Aplikacija treba prikazati referencu, genske anotacije, pokrivenost reference mapiranim očitanjima te mapiranja pojedinih očitanja. Aplikacija treba omogućiti zumiranje prikaza do razine pojedinog nukleotida te prikaz podataka prilagoditi razini zumiranja. Treba podržati učitavanje standardnih formata datoteka kao što su FASTA, FASTQ, SAM i BAM. Rješenje mora biti napisano kao desktop aplikacija u programskom jeziku JAVA. Pri tome osigurati da radi na operacijskom sustavu Linux. Programski kod je potrebno komentirati i pri pisanju pratiti neki od standardnih stilova. Napisati iscrpne upute za instalaciju i izvođenje. Kompletno programsko rješenje postaviti na GitHub pod jednom od OSI-odobrenih licenci.

Rok za predaju rada: 12. lipnja 2020.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 6679

VIZUALIZACIJA PODATAKA DOBIVENIH SEKVENCIRANJEM RNA

Jan Vrlec

Zagreb, lipanj 2020.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 6679

VIZUALIZACIJA PODATAKA DOBIVENIH SEKVENCIRANJEM RNA

Jan Vrlec

Zagreb, lipanj 2020.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 13. ožujka 2020.

ZAVRŠNI ZADATAK br. 6679

Pristupnik:	**Jan Vrlec (0036495516)**
Studij:	Računarstvo
Modul:	Programsko inženjerstvo i informacijski sustavi
Mentor:	doc. dr. sc. Krešimir Križanović

Zadatak: **Vizualizacija podataka dobivenih sekvenciranjem RNA**

Opis zadatka:

Današnji uređaji za sekvenciranje nisu u stanju sekvencirati cijele genome, već samo pročitati manje dijelove (očitanja) i pri tome unose određenu količinu pogreške. Jedan od prvih i osnovni zadataka prilikom analize podataka dobivenih sekvenciranjem jest mapiranje očitanja na referentni genom ili transkriptom. Zbog velike količine podataka, nije moguće ručno analizirati mapiranje svakog pojedinog očitanja, već je puno praktičnije koristiti prikladne alate za vizualizaciju. Potrebno je napisati aplikaciju za vizualizaciju mapiranja RNA očitanja na referentni genom. Aplikacija treba prikazati referencu, genske anotacije, pokrivenost reference mapiranim očitanjima te mapiranja pojedinih očitanja. Aplikacija treba omogućiti zumiranje prikaza do razine pojedinog nukleotida te prikaz podataka prilagoditi razini zumiranja. Treba podržati učitavanje standardnih formata datoteka kao što su FASTA, FASTQ, SAM i BAM. Rješenje mora biti napisano kao desktop aplikacija u programskom jeziku JAVA. Pri tome osigurati da radi na operacijskom sustavu Linux. Programski kod je potrebno komentirati i pri pisanju pratiti neki od standardnih stilova. Napisati iscrpne upute za instalaciju i izvođenje. Kompletno programsko rješenje postaviti na GitHub pod jednom od OSI-odobrenih licenci.

Rok za predaju rada: 12. lipnja 2020.

Zahvala:

Htio bih se zahvaliti svom mentoru, doc.dr.sc. Krešimiru Križanoviću, koji mi je pomogao svojim savjetima i uvijek bio spreman razjasniti nejasnoće. Svakotjedni online sastanci održani u ovo vrijeme pandemije koronavirusa su mi bili velika motivacija za rad.

Također bih se htio zahvaliti svojoj obitelji, ponajviše svojim roditeljima koji su me cijelo vrijeme podržavali i omogućili mi nesmetano studiranje.

Sadržaj

Uvod

Sve genetske informacije o nekom živom biću su sadržane u lancima nukleinskih kiselina – DNA i RNA. Određivanje slijeda nekog DNA ili RNA lanca omogućava dublje razumijevanje nekog organizma i njegovih bioloških interakcija. Te informacije se primjerice mogu iskoristiti kako bi se otkrili znakovi bolesti ili potencijal razvoja neke bolesti kod osobe. Na praktičnom primjeru, podatci dobiveni sekvenciranjem se koriste kod ljudi oboljelih od raka kako bi se identificirao tip raka i u konačnici se liječenje usmjerilo u pravom smjeru.

Sam proces kojim se određuje poredak nukleinskih baza unutar nekog lanaca, a u konačnici nastoji odrediti i cjelokupne genome se naziva sekvenciranjem. U današnje vrijeme još nismo u stanju sekvencirati cijele genome, nego sekvenciramo samo manje dijelove koje nazivamo očitanjima.

Datoteke u koje se spremaju podatci dobiveni sekvenciranjem se sastoje od velike količine podataka. Tako na primjer tekstualne datoteke koje sadrže podatke o očitanjima mapiranim na neki referentni genom sežu od nekoliko stotina megabajta za jednostavnije organizme, pa do nekoliko desetaka ili stotina gigabajta za složenije eukariotske organizme.

Takve podatke je nemoguće efikasno analizirati ručnim putem, već u tu svrhu koristimo programske alate za vizualizaciju podataka. Uz pomoć vizualizacije lako je moguće vidjeti informacije kao što su npr. pozicije pojedinih očitanja s obzirom na referentni genom ili pokrivenost reference mapiranim očitanjima.

1. Bioinformatička pozadina

Nukleinske kiseline zajedno sa proteinima i ugljikohidratima su biomolekule koje nazivamo biopolimerima – esencijalnim molekulama za funkcioniranje svih poznatih živih organizama. Same nukleinske kiseline možemo podijeliti na dvije temeljne vrste – DNA i RNA. RNA molekule su nadalje podijeljene na više vrsta ovisno o svojoj ulozi. Tako razlikujemo glasničku RNA (mRNA), transportnu RNA (tRNA), ribosomsku RNA (rRNA) te veći broj regulacijskih RNA.

Molekule DNA su oblika dvostruke uzvojnice, gdje se same uzvojnice sastoje od nukleotida kao manjih gradivnih jedinica, a par nukleotida sa svake strane zavojnice se nazivaju baznim parovima. Nukleotidi u DNA se sastoje od četiri različite nukleinske baze: adenina, citozina, gvanina i timina koje označavamo početnim slovima njihovih imena A, C, G, T.

Za razliku od DNA, molekule RNA su građene od jednog lanca. Razlikuju se i u jednoj nukleinskoj bazi, pa se tako nukleotidi u RNA sastoje od: adenina, citozina, gvanina i uracila s oznakama A, C, G, U.

1.1. Geni i transkripcija

Geni su osnovne jedinice nasljeđivanja koje služe kao nacrti za proizvodnju RNA ili proteina. Sam proces koji uzima informacije iz gena kako bi se provela takva proizvodnja naziva se ekspresijom gena.

Koraci ekspresije gena su vidljivi u središnjoj dogmi molekularne biologije (podebljani dio tablice ispod), koja prikazuje sve načine protoka informacija između DNA, RNA i proteina.

Tablica 1: Središnja dogma molekularne biologije, preuzeto sa [2]

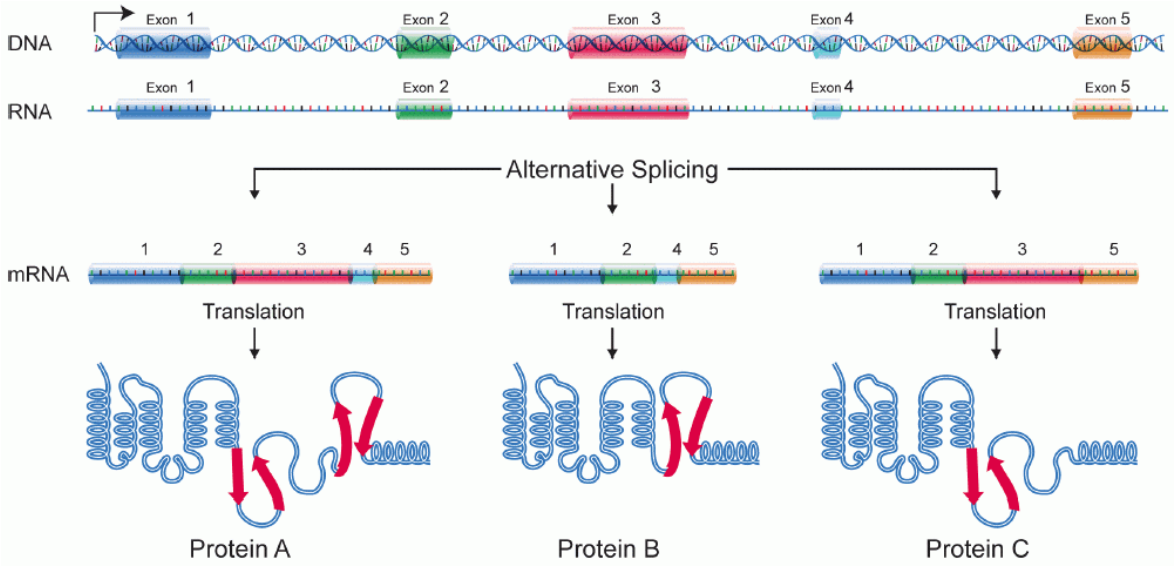
Opći prijenos	**Specijalni prijenos**	**Nepoznat prijenos**
DNA -> DNA	RNA -> DNA	protein -> DNA
DNA -> RNA	RNA -> RNA	protein -> RNA
RNA -> protein	DNA -> protein	protein -> protein

Tablica nam prikazuje svih 9 mogućih kombinacija prijenosa informacija između DNA, RNA i proteina. Opći prijenosi su prijenosi koji se odvijaju u većini stanica. Oni su redom umnažanje DNA, kopiranje informacije iz DNA u mRNA što nazivamo transkripcijom i sintetiziranje proteina iz RNA zvano translacijom. Specijalni prijenosi, kao što i samo ime sugerira, se odvijaju samo u specijalnim uvjetima, dok se za nepoznate prijenose vjeruje da se ne događaju uopće.

Dakle, ekspresija gena započinje transkripcijom koja kod prokariota direktno stvara mRNA koja je odmah spremna za daljnji proces translacije. Kod eukariota je situacija malo složenija. Geni eukariota su podijeljeni na manje dijelove – nekodirajuće regije koje nazivamo intronima i kodirajuće regije koje nazivamo eksonima. Kako bi se formirala funkcionalna mRNA iz primarnog transkripta dobivenog transkripcijom vrše se sljedeće radnje:

1. Modificiraju se krajevi transkripta – dodaje se 5' kapa na početku transkripta i 3' poli A rep na kraj transkripta
2. Vrši se proces prekrajanja introna – svi introni se izbacuju iz transkripta te se zatim preostali eksoni spajaju natrag zajedno, formirajući konačnu, zrelu mRNA

Geni koji sadrže introne i eksone imaju zanimljivo svojstvo, a to je da se u procesu prekrajanja eksoni mogu spojiti na više načina. Tako se iz jednog gena može formirati više različitih mRNA i u konačnici više različitih proteina. Taj proces nazivamo alternativnim prekrajanjem.



Slika 1.1: Alternativno prekrajanje, slika preuzeta sa [5]

1.2. Sekvenciranje

Sekvenciranje je proces kojim određujemo poredak nukleotida unutar nekog DNA ili RNA lanca. Uz pomoć današnjih metoda i uređaja za sekvenciranje možemo sekvencirati do nekoliko desetaka tisuća nukleotida. Kako ne možemo sekvencirati cijele lance odjednom, moramo ih prvo podijeliti u kraće lance.

Jedna od najkorištenijih metoda za sekvenciranje je tzv. *shotgun* sekvenciranje. Tom metodom DNA se lomi na slučajan način u manje fragmente koji se mogu sekvencirati iz čega se dobivaju očitanja. Ponavljanjem ovog postupka dobiva se više očitanja koja se mogu međusobno preklapati. Nakon što se dobiju očitanja potrebno je pronaći preklapanja u očitanjima kako bi ih spojili u kontinuiranu sekvencu.

Kvaliteta sekvenciranja se može izraziti mjerom koju nazivamo pokrivenost. Prosječna pokrivenost za promatrani segment genoma G iz kojeg je sekvenciranjem dobiveno N očitanja s prosječnom duljinom L je dana formulom:

$$C = NL/G$$

Gledajući pokrivenost za samo jednu poziciju unutar genoma, pokrivenost je jednaka broju očitanja koja postoje na toj poziciji.

Što je veća pokrivenost, to je veća vjerojatnost da neće biti praznina u dobivenoj sekvenci. Praznine odgovaraju pozicijama za koje ne postoji niti jedno očitanje.

Za razliku od DNA, RNA se uglavnom ne sekvencionira direktno. Umjesto toga RNA se prvo konvertira u takozvanu cDNA (Komplementarna DNA) procesom obrnute transkripcije. Nakon što je dobivena cDNA, nad njom se dalje vrši standardno sekvenciranje DNA.

1.3. Mapiranje i poravnanje

Nakon što se nad DNA provede postupak sekvenciranja dobije se niz sekvenci koja predstavljaju očitanja. Uz postojeću referentnu sekvencu ta očitanja je moguće mapirati ili poravnati na tu sekvencu.

Procesom mapiranja dobivaju se pozicije očitanja naspram referentnog genoma.

Mapiranje je prvi dio procesa poravnanja očitanja na referentni genom. Nakon što je očitanje mapirano svaka baza očitanja se uspoređuje s pripadnom bazom na referenci. Tim

postupkom se dobiva korespondencija baza – koje baze iz očitanja odgovaraju bazama u referenci, koje se baze brišu iz reference ili dodaju u referencu. Ta korespondencija se može zapisati u obliku CIGAR operacija koje su objašnjene kasnije u dokumentu.

Mapiranje, iako daje manje informacija od poravnanja, je i dalje izuzetno korisno. Naime informacije dobivene poravnanjem često nisu potrebne za analizu očitanja, a proces mapiranja je mnogo brži od procesa poravnanja.

2. Tipovi podataka

Kod sekvenciranja i poravnanja se koristi mnoštvo različitih tipova podataka, te je u praktičnom dijelu ovog rada obrađen samo dio tih datoteka.

Neki od tipova podataka su:

- FASTA
- FASTQ
- SAM / BAM
- BED / GTF / GFF

2.1. FASTA

FASTA je tekstualni format datoteke koji služi kako bi prikazali sljedove proteina ili nukleinskih kiselina. Datoteka se sastoji od linija zaglavlja nakon koje slijedi jedna ili više linija u kojima je zapisan definirani slijed. Dodatno, unutar datoteke mogu postojati i linije komentara koje započinju znakom ",".

Zaglavlje započinje znakom ">" nakon čega slijedi tekst koji se koristi kao identifikator slijeda.

Slijed se sastoji od niza slova gdje svako slovo određuje neki nukleotid ili aminokiselinu. U sklopu prikaza RNA slijed će se sastojati od slova A, C, G i U, ali se mogu koristiti i dodatni znakovi u slučaju da se ne može jednoznačno odrediti koja se baza nalazi na poziciji (npr. slovo R nam označava da se na poziciji nalazi adenin ili gvanin) ili ako postoji procijep u slijedu, što je označeno znakom "-".

Primjer izgleda FASTA datoteke:

```
>chrI
ccacaccacacccacacacccacacaccacaccacacaccacaccacacccacacacacacatCCTAA
>chrII
AAATAGCCCTCATGTACGTCTCCTCCAAGCCCTGTTGTCTCTTACC
```

2.2. FASTQ

FASTQ kao format se koristi za prikaz podataka dobivenih iz uređaja za sekvenciranje. Slično kao i FASTA datoteke, FASTQ sadrži sekvence i njihove identifikatore, s time da se dodatno uz svako očitanje zapisuje i njegova kvaliteta.

Primjer jednog očitanja iz FASTQ datoteke:

> @SimG1_S1_1
> ATGATTCTTCTTTTAAACGGAAGAGACATTGCAAAAGTTTGAGTGACCA
> +SimG1_S1_1
> +---.,+'+"-$)--&'...!.(&+$."&*%,#.(.*%#&%&(+../%..*.%"),%%.(,$(.(,%-!&.+-

Prva linija svakog očitanja započinje znakom "@", nakon čega slijedi identifikator slijeda i opcionalan opis.

Druga linija je slijed.

Treća linija započinje znakom "+" iza kojega se može, ali ne mora nalaziti isti identifikator slijeda kao u prvoj liniji i opis.

Četvrta linija predstavlja kvalitetu slijeda iz druge linije.

Za svaki znak iz slijeda postoji točno jedan znak koji predstavlja kvalitetu na toj poziciji. Sama vrijednost kvalitete je vjerojatnost netočnog očitanja prikazana takozvanim Sangerovim formatom kvalitete koji je dan izrazom:

$$Q_{Sanger} = -10log_{10}p$$

Taj izraz daje vrijednosti kvalitete od 0 do 93, koji se zatim prikazuju kao ASCII znakovi u rasponu od 33 do 126.

2.3. SAM i BAM

SAM i BAM datoteke služe za pohranu podataka o sekvencama očitanja mapiranim ili poravnatim na referentnu sekvencu. Oba tipa datoteka se sastoje od jednakih podataka, ali zapisanih u različitom formatu. U radu s raznim programskim alatima se uglavnom koristi BAM kao binarni komprimirani tip datoteke, dok je SAM ostvaren kao ljudima lako čitljiv tekstualni dokument. BAM datoteke često dolaze u paru sa datotekom ekstenzije ".bam.bai"

koja sadrži podatke o indeksiranju za BAM datoteku kako bi se ubrzale operacije čitanja datoteke o čemu ovisi mnogo programskih rješenja koja barataju ovim tipom podataka.

2.3.1. Format SAM datoteka

SAM i BAM datoteke se sastoje od dva dijela: opcionalnog zaglavlja i linija koje predstavljaju pojedina očitanja.

Zaglavlje, ako je prisutno, se nalazi na početku datoteke, gdje je svaka linija zaglavlja započeta znakom "@" i dva predefinirana slova koja određuju značenje zaglavlja, nakon čega slijede podatci vezani za zaglavlje odvojeni prazninama. Kao primjer zaglavlja možemo navesti takozvani *sequence header*:

```
@SQ   SN:chrI      LN:230218
@SQ   SN:chrII     LN:813184
```

On daje informaciju o imenima referentnih sekvenci na koje su očitanja mapirana, kao i duljinu za svaku od sekvenci što se može pokazati korisnim za vizualizaciju očitanja na pojedinu referentnu sekvencu.

Linije koje ne započinju znakom "@" predstavljaju očitanja, gdje se svaka linija sastoji od jedanaest obaveznih polja odvojenih prazninama, te još dodatnih opcionalnih polja.

Tablica 2: Obavezna polja za svako očitanje

Naziv polja	**Tip podatka**	**Opis**
QNAME	String	"Query template name" – očitanja sa jednakom vrijednosti polja dolaze iz istog predloška
FLAG	Int	Broj čiji svaki bit predstavlja jednu zastavicu, ukupno 12 zastavica
RNAME	String	Naziv referentne sekvence, odgovara jednoj od sekvenci navedenih u zaglavlju (ako su navedene)
POS	Int	Pozicija mapiranja očitanja na referentnu sekvencu
MAPQ	Int	Kvaliteta mapiranja očitanja
CIGAR	String	Definira CIGAR operacije

RNEXT	String	Naziv referentne sekvence sljedećeg očitanja u predlošku, ako postoji
PNEXT	Int	Pozicija mapiranja sljedećeg očitanja iz predloška na referentnu sekvencu
TLEN	Int	Duljina predloška
SEQ	String	Sekvenca danog očitanja zapisana kao niz nukleotida
QUAL	String	Kvaliteta očitanja u Sangerovom formatu (isto kao i u FASTQ datotekama)

Primjer izgleda SAM datoteke:

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

“*“ se koristi ako ne postoji informacija za neko od obaveznih polja.

Potpuna specifikacija SAM tipa datoteka može se pronaći na [3].

2.3.2. CIGAR

CIGAR je tekstualni niz koji se sastoji od baznih duljina i operacija koje su asocirane uz njih. Operacije nam indiciraju koje se baze poklapaju sa referentnim genomom, dodaju se u referencu ili se uklanjaju iz reference. Primjer CIGAR-a: 4M2D3M5N1M4I4M

CIGAR je se sastoji od parova brojki i slova, gdje svako slovo označava naredbu koja se izvršava onoliki broj puta kolika je brojka koja prethodi tom slovu.

U danom primjeru se nalaze četiri od ukupno devet različitih CIGAR naredbi:

- M(*Match*) – poravnanje odgovara referenci
- N(*Alignment gap*) – preskače se dio reference
- D(*Deletion*) – izbrisani dio iz reference
- I(*Insertion*) – dodavanje u referencu

Dvije bitne informacije koje CIGAR naredbe daju za vizualizaciju su da li se kao rezultat te naredbe pomičemo po sekvenci očitanja ili reference. Tako će duljina sekvence svakog očitanja, gdje je sama sekvenca zapisana u polju SEQ, biti jednaka zbroju duljina operacija kojima se pomičemo po očitanju.

Primjer funkcionalnosti CIGAR naredbi korak po korak (siva pozadina kod teksta predstavlja obrađen dio):

Prva naredba govori da imamo četiri nukleotida u očitanju koji odgovaraju poziciji na referenci.

CIGAR: 4M2D3M5N1M4I4M

Referenca: ATATGCAGCCTGACCTGAA

Očitanje: AAATAGCCCTCATGTA

Sljedeće imamo 2 brisanja s obzirom na referencu, ovdje označeni znakom "*" nakon čega slijede još tri nukleotida koji odgovaraju referenci.

CIGAR: 4M2D3M5N1M4I4M

Referenca: ATATGCAGCCTGACCTGAA

Očitanje: AAAT**AGCCCTCATGTA

Slijedi pet preskoka s obzirom na referencu, označenih kao prazna polja i još jedan nukleotid koji odgovara referenci.

CIGAR: 4M2D3M5N1M4I4M

Referenca:	ATATGCAGCCTGACCTGAA
Očitanje:	AAAT**AGC CCTCATGTA

Na posljetku imamo četiri umetanja u referencu, označenih znakom "*" na referenci i još četiri nukleotida koji odgovaraju referenci.

CIGAR:	4M2D3M5N1M4I4M
Referenca:	ATATGCAGCCTGACC****TGAA
Očitanje:	AAAT**AGC CCTCATGTA

2.4. Genske anotacije

Anotiranje genoma je proces kojim se identificiraju funkcionalni elementi iz sekvence genoma. Tim procesom se pridaje značenje dijelovima genoma, a same informacije su zapisane kao genske anotacije.

Neki od popularnijih tipova podataka za pohranu informacija o genskim anotacijama su BED, GTF i GFF. Fokus u ovom projektu je bio na BED tip, kao jednostavniji za obraditi u sklopu vizualizacije.

BED nam omogućava fleksibilno definiranje podataka za prikaz u anotacijskoj traci. Svaka linija BED datoteke se sastoji od tri obavezna polja i još 9 opcionalnih polja koja detaljnije opisuju anotaciju.

Tri obavezna polja su:

- chrom – naziv kromosoma
- chromStart – početna pozicija u kromosomu
- chromEnd – krajnja pozicija u kromosomu

Ostalih 9 polja:

- name – definira naziv linije
- score – vrijednost između 0 i 1000

- strand – vrijednosti “+“ ili “-“ ovisno o orijentaciji lanca, vrijednost “.“ ako nema orijentaciju
- thickStart – početna pozicija od koje se deblje iscrtava anotacija
- thickEnd – krajnja pozicija debljeg iscrtavanja
- itemRgb – RGB vrijednost podataka u liniji
- blockCount – broj blokova(eksona)
- blockSizes – zarezom razdvojene veličine blokova, broj vrijednosti treba odgovarati vrijednosti blockCount
- blockStarts – zarezom razdvojene početne pozicije blokova relativne na vrijednost chromStart, broj vrijednosti treba odgovarati vrijednosti blockCount

2.5. Alati samtools i minimap2

Minimap2 je programski alat za sekvenciranje koji vrši poravnavanje DNA ili mRNA sekvence naspram referentne baze. Nakon što se provede instalacija programa, program se može pozvati iz Linux ljuske korištenjem naredbe “minimap2“.

Uz pomoć informacija pohranjenih unutar FASTA i FASTQ datoteka moguće je generirati poravnata očitanja u obliku SAM datoteke pozivom naredbe

```
minimap2 –a ref.fa query.fq > alignment.sam
```

, gdje je ref.fa FASTA tip datoteke, a query.fq FASTQ tip datoteke.

Samtools je niz alata koji omogućuju manipulaciju poravnanjima u SAM tipu datoteka. Sortiranje podataka, sjedinjenje podataka iz više datoteka u jednu i indeksiranje su samo neke od mogućnosti alata. Sam alat se sastoji od 3 zasebna, ali međusobno koordinirana projekta:

- htslib – knjižnica napisana u programskom jeziku C, omogućava rukovanje podatcima za sekvenciranje visoke propusnosti
- samtools – alati za rukovanje SAM, BAM i CRAM tipovima datoteka
- bcftools – alati za rukovanje VCF i BCF tipovima datoteka

Jedna od korisnih opcija alata samtools je da se mogu uzeti SAM datoteke koje su generirane uz pomoć alata minimap2 i komprimirati ih u BAM datoteke.

Sve što je potrebno učiniti za to je pozvati naredbu

```
samtools view –b alignment.sam > alignment.bam
```

u Linux ljusci ako postoje @SQ linije u zaglavlju SAM datoteke (u slučaju nedostatka @SQ linija potrebna je FASTA datoteka kako bi mogli generirati BAM datoteku).

Također se lako može generirati i indeks za BAM datoteku pozivom sljedeće naredbe nad datotekom koja se želi indeksirati:

```
samtools indeks alignment.bam
```

3. Aplikacija za vizualizaciju

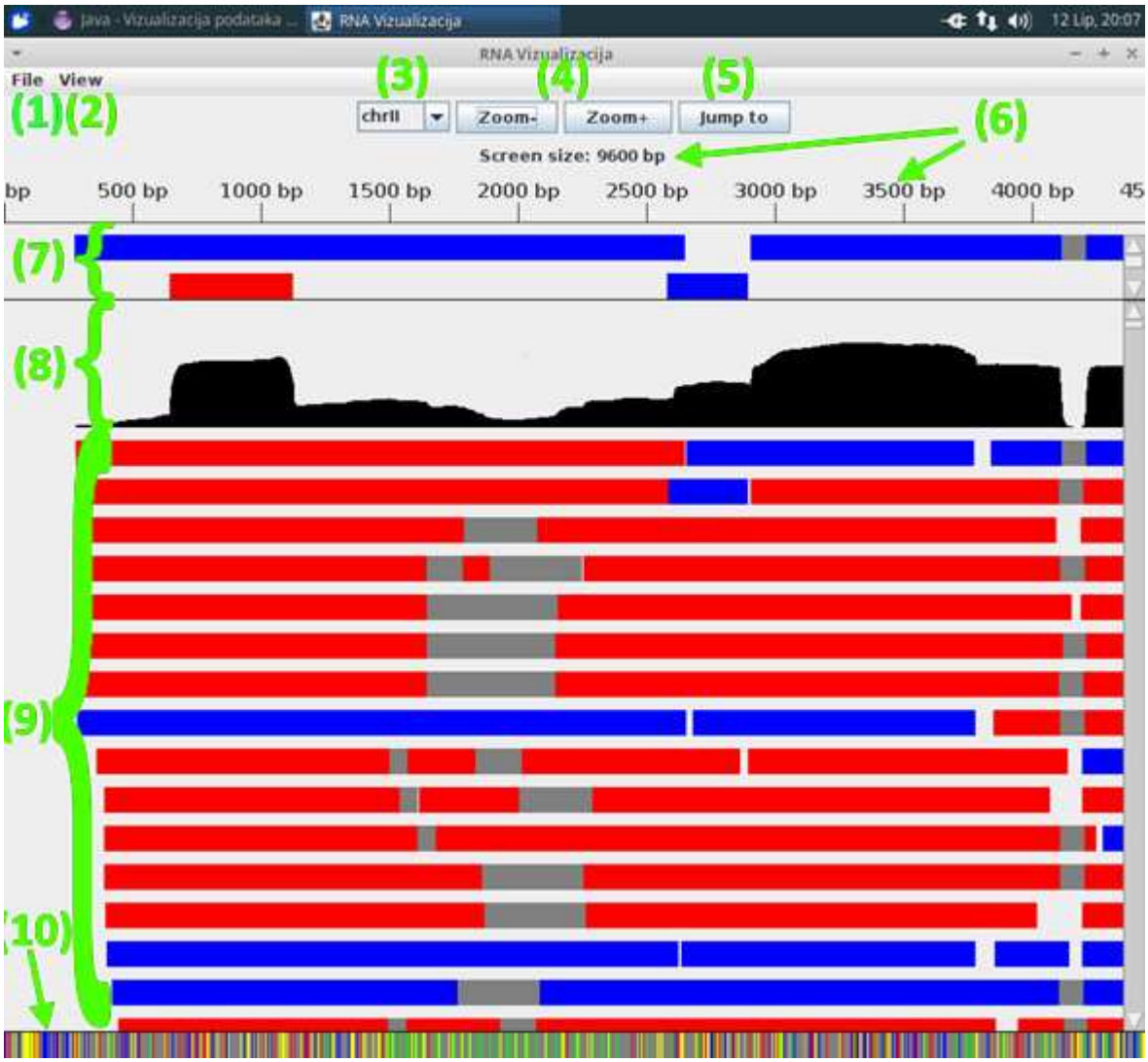
Aplikacija je napisana u programskom jeziku Java, gdje je sama vizualizacija ostvarena korištenjem paketa Java Swing i Java Awt. Kao integrirano razvojno okruženje(IDE) se koristila platforma Eclipse. Aplikacija je u cijelosti napisana unutar operacijskog sustava Xubuntu pokrenutog unutar virtualne mašine uz pomoć aplikacije Oracle VM VirtualBox.

Sam kod je napisan u objektno orijentiranom obliku. Podatci o referentnoj sekvenci, očitanjima i genskim anotacijama učitani iz različitih tipova datoteka se pohranjuju u posebne objekte. Informacije unutar objekata su filtrirane naspram informacija dobivenih iz datoteka, te sadrže isključivo podatke koji se koriste u vizualizaciji.

Unutar prozora aplikacije razlikuju se četiri glavna vizualna objekta – traka izbornika, panel koji sadrži ravnalo i dodatne objekte za prilagodbu prikaza, panel koji sadrži anotacije, panel koji sadrži očitanja i rezultirajući histogram i panel koji sadrži traku kojom je prikazana referenca.

Program sadrži i neke dodatne klase kao što su npr. klasa koja je zaslužna za učitavanje datoteka i njihovu pretvorbu u funkcionalne objekte i klasa za odabir boja pri iscrtavanju pojedinih objekata na ekranu.

3.1. Opći izgled i funkcionalnost



Slika 3.1: Aplikacija sa označenim elementima

1. Izbornik za odabir datoteka. Prvo se učitava FASTA tip datoteke koja sadrži referentnu sekvencu. Nakon što se učita FASTA datoteka, može se učitati SAM datoteka i genske anotacije u obliku BED datoteke.
2. Izbornik koji sadrži opcije pretrage očitanja i anotacija prema nazivu. Nakon upisa naziva prikaz se pomiče na poziciju pronađene stavke. Prikazuje se i ukupan broj pronađenih stavki, te je moguće pomicanje na sljedeću stavku.
3. Padajući izbornik za odabir slijeda. Moguće je odabrati jedan od sljedova definiranih unutar FASTA datoteke. Pri odabiru se cjelokupni prikaz prilagođava novo odabranom slijedu.

4. Gumbi za prilagodbu razine zumiranja. Postoji ukupno 9 razina zumiranja, gdje najdetaljniji prikaz odgovara razini zumiranja od 15x, a najširi prikaz unutar svakog horizontalnog piksela sadrži 25 podataka.
5. Gumb za skok na poziciju. Upisuje se numerička vrijednost i cjelokupni prikaz se pomiče na tu poziciju unutar slijeda.
6. Ravnalo. Prikazuje poziciju na kojoj se nalazimo unutar slijeda i skalu prikaza. Skala je prikazana mjernom jedinicom bp(bazni parovi). Iscrtava se tek pri učitavanju FASTA datoteke.
7. Anotacijska traka.
8. Histogram očitanja. Prikazuje ukupnu količinu očitanja na nekoj poziciji s obzirom na referencu. Generira se i iscrtava pri učitavanju SAM datoteke.
9. Prikaz očitanja.
10. Referentna sekvenca.

Korištenjem strelica na tipkovnici moguće je pomicanje cjelokupnog prikaza s obzorom na referentnu sekvencu(strelice lijevo i desno) i vertikalno pomicanje kroz očitanja(strelice gore i dolje). Za navedenu navigaciju, kao i za vertikalno pomicanje po anotacijskoj traci postoji i dodatna mogućnost pomicanja uz pomoć *scrollbarova*.

3.2. Referentna sekvenca

Referentna sekvenca prikazuje sadržaj odabrane FASTA datoteke. Razliku se tri načina prikaza ovisno o razini zumiranja:

1. Neutralna razina – svaki zaseban nukleotid se prikazuje na ekranu, te je kodiran bojom. Boje u ovisnosti o nukleotidu su: plava za adenin, crvena za citozin, zelena za gvanin, žuta za timin i narančasta za uracil. U slučaju da nije poznato o kojem se nukleotidu radi koristi se crna boja, a ako postoji praznina u prikazu bit će prikazana svjetlo sivom bojom.

Slika 3.2: Prikaz reference za neutralnu razinu zumiranja

2. Veća razina zumiranja – pri većoj razini zumiranja svaki nukleotid će okupirati veći broj piksela na ekranu. Uz već navedeno kodiranje bojom tada se nukleotidi prikazuju i odgovarajućim slovom za lakši pregled.

Slika 3.3: Prikaz reference za veću razinu zumiranja

3. Prikaz većeg broja nukleotida – kada se prikazuje više nukleotida nego što stane na ekran više ih se grupira unutar jednog piksela. Ako unutar grupe nukleotida neki nukleotid predstavlja većinu piksel će se prikazati prethodno navedenim bojama. U slučaju da niti jedan nukleotid nema većinski udio piksel će biti prikazan tamno sivom bojom.

Slika 3.4: Prikaz reference za veći broj nukleotida

3.3. Očitanja

Učitavanjem SAM datoteke, očitanja navedena za trenutno odabrani slijed se prikazuju na odgovarajućoj horizontalnoj poziciji. No, kao što je već navedeno u poglavlju sekvenciranje, često se dobiju očitanja koja se preklapaju. Kako bi se mogla prikazati sva očitanja, aplikacija svakom očitanju pridjeljuje i vertikalnu komponentu takvu da se izbjegnu bilo kakva preklapanja.

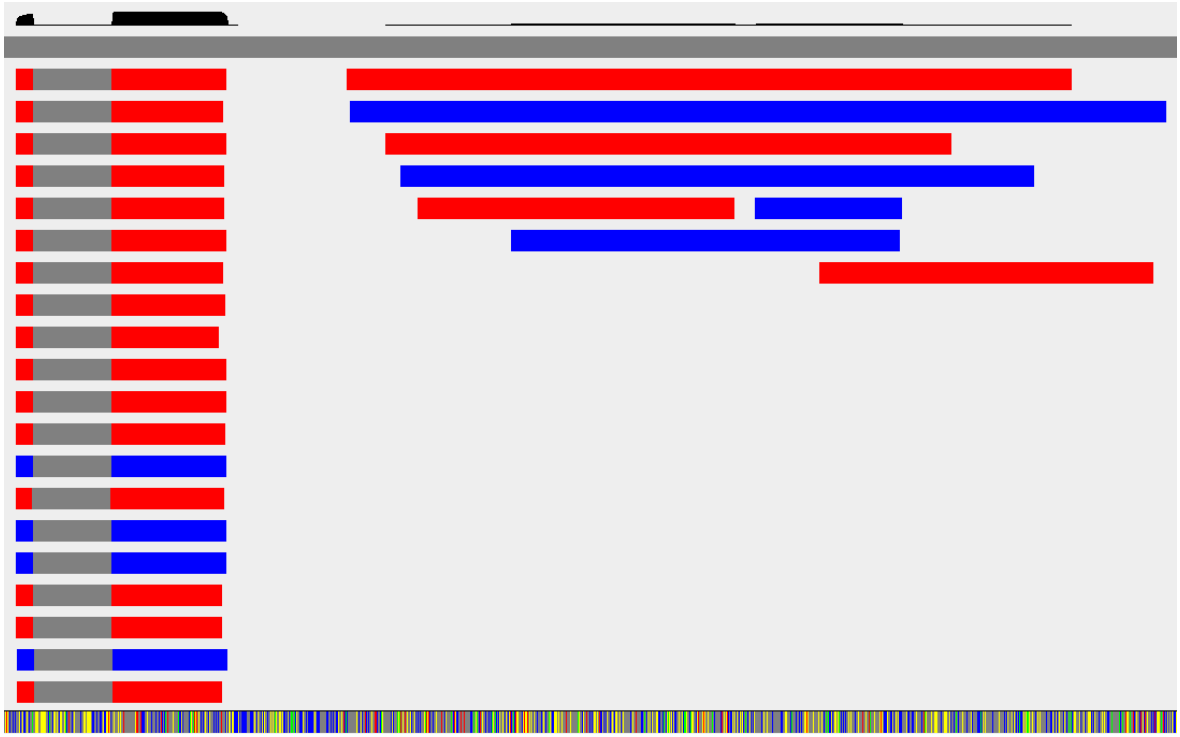
Sama duljina pojedinog očitanja na ekranu se iščitava iz pripadnog CIGAR-a. Tu se obraća pozornost na već prije spomenute operacije koje označavaju pomicanje po referenci ili očitanju. Operacije koje uzrokuju pomicanje po referenci će proširiti prikaz očitanja, a ako ne uzrokuju pomicanje po očitanju, te regije će dodatno biti različito prikazane. Operacije koje uzrokuju samo pomicanje po očitanju(npr. umetanje u referencu) se ignoriraju zbog kompleksnosti prikaza do koje bi došlo proširivanjem trake referentne sekvence za veliku količinu takvih operacija.

Kod očitanja se razlikuju tri različita prikaza bojom:

- crvenom bojom ako je postavljena zastavica *reverse complement* za očitanje
- plavom bojom ako nije postavljena zastavica *reverse complement* za očitanje

- tamno sivom bojom, neovisno o zastavici *reverse complement*, ako se na dijelu očitanja nalaze CIGAR operacije brisanja iz reference ili preskakanja reference

Na slici 3.5. je vidljiv prikaz svih navedenih tipova očitanja unutar aplikacije zajedno sa referentnom sekvencom na dnu i histogramom očitanja na vrhu.



Slika 3.5: Prikaz očitanja

3.4. Anotacijska traka

Anotacijska traka prikazuje genske anotacije dobivene iz BED datoteke poravnate s obzirom na referentnu sekvencu. U prikazu anotacija se koriste neke boje koje se koriste i kod očitanja, no samo značenje boja se razlikuje:

- Anotacija se prikazuje crvenom bojom ako je vrijednost *strand* jednaka "+"
- Anotacija se prikazuje plavom bojom ako je vrijednost *strand* jednaka "-"
- Ako su postavljene vrijednosti za blokove, praznine među blokovima su prikazane tamno sivom bojom
- U slučaju da anotacija nema orijentaciju postavljenu vrijednošću *strand*, bit će prikazana crnom bojom

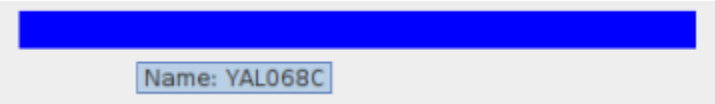
Već navedeni blokovi kod anotacija zapravo predstavljaju prije navedene eksone, dok praznine među blokovima predstavljaju introne.

Slično kao i kod očitanja, pojedine anotacije se mogu preklapati. Zbog toga je i njima pridijeljena vertikalna komponenta, pa se očitanja koja bi se preklapala prikazuju jedna ispod drugih. Sama traka zauzima manje mjesta od prikaza očitanja pošto su preklapanja kod anotacija uglavnom rjeđa.

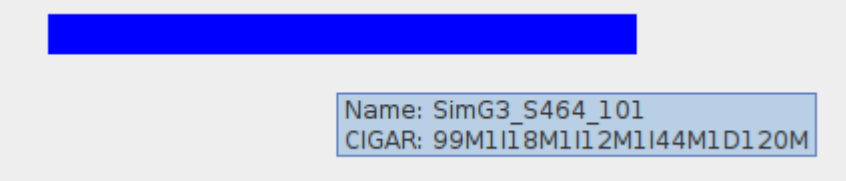
3.5. Dodatna funkcionalnost

Biranjem opcije "View" u traci izbornika daju se opcije pretrage očitanja i anotacija. Daljnjim odabirom otvara se prozor za upis naziva očitanja ili reference. Nakon potvrde upisa pretražuju se sva očitanja ili anotacije koje sadrže upisani podatak unutar svog imena. Također se otvara novi prozor koji prikazuje broj pronađenih podataka i daje korisniku mogućnost prolaska kroz sve pronađene podatke. Kod prelaska na svaki sljedeći rezultat pretrage, pregled se prebacuje na poziciju gdje je pronađeni podatak, te se podatak dodatno istakne radi lakšeg pregleda.

Prelaskom miša preko pojedinog očitanja ili anotacije se prikazuje *tooltip* na ekranu. Unutar *tooltipa* se nalaze dodatni podatci o elementu koji se ne mogu pregledno prikazati unutar samog prozora. Za pojedinu anotaciju se prikazuje njezin naziv, a za očitanje se prikazuje naziv i kompletan CIGAR.



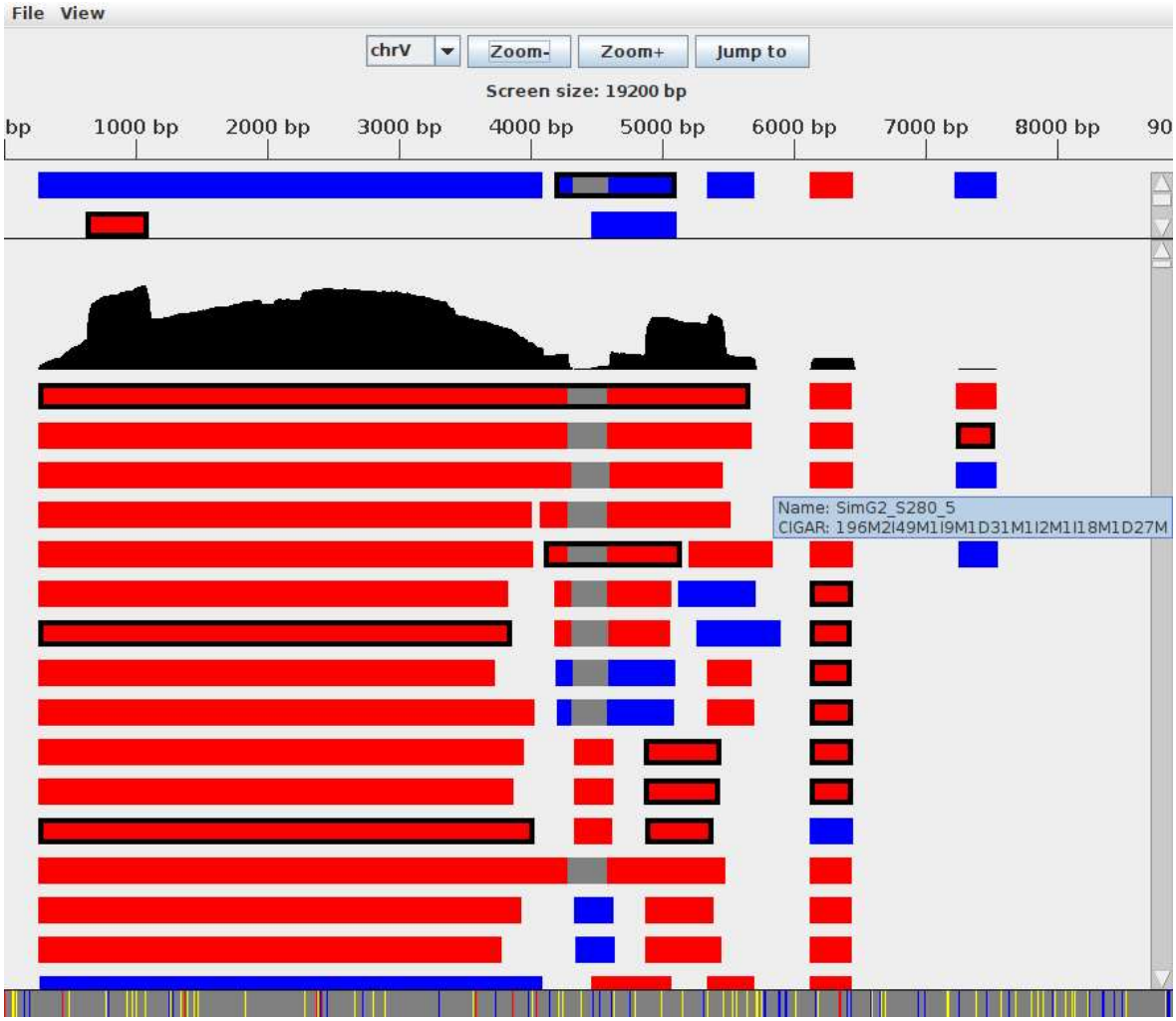
Slika 3.6: *Tooltip* uz anotaciju



Slika 3.7: *Tooltip* uz očitanje

Isto kako se pojedina očitanja i anotacije istaknu prilikom pretrage, moguće je ih je i manualno istaknuti. Pritiskom desnog klika miša na očitanje ili anotaciju izmjenjuje se je li

element istaknut ili nije. Tako je moguće primjerice proći kroz neki dio očitanja i označiti ona koja trenutno promatramo radi lakšeg pregleda.



Slika 3.8: Prikaz više istaknutih elemenata

Ako se na pojedino očitanje ili anotaciju pritisne lijevim klikom miša, kopirat će se podatci koji su inače zapisani unutar odgovarajućeg *tooltipa*. Jednom kopirani podatci se mogu dalje zalijepiti u tekstualni editor. Ova funkcionalnost je izrazito korisna ako se želi pregledati neki CIGAR. Pošto je često dugačak i ponekad niti ne stane čitav u ekran, jednostavnije ga je pregledati unutar tekstualnog editora.

Zaključak

Na osnovu podataka dobivenih iz više tipova datoteka koje sadrže informacije o referentnoj sekvenci, očitanjima i genskim anotacijama je izgrađena aplikacija koja ih vizualizira u jednostavnom i urednom obliku. Aplikacija omogućava jednostavan pregled inače velikih količina podataka. Prikaz, iako je jednostavan i prikazuje samo dio podataka sadržanih unutar korištenih datoteka, se i dalje može pokazati korisnim – već i sam prikaz pozicija pojedinih očitanja naspram referentne sekvence daje korisne informacije.

Aplikacija ima mnogo prostora za proširenje funkcionalnosti i podataka koje prikazuje. Tako bi se na primjer mogle prikazati dodatne informacije o očitanjima i anotacija prelaskom miša preko njih. Pri većoj razini zumiranja bi se mogli prikazivati pojedini nukleotidi unutar samih očitanja, kao i detaljniji prikaz CIGAR operacija.

Dakle, u jednoj rečenici, aplikacija razvijena u sklopu ovog rada daje jednostavan prikaz velike količine podataka dobivene sekvenciranjem RNA te bi ju se u budućnosti lako dalo proširiti da prikazuje i više podataka.

Literatura

1. Repozitorij predmeta *Bioinformatika*, https://www.fer.unizg.hr/predmet/bio, 11. 6. 2020.
2. Mile Šikić, Mirjana Domazet-Lošo, *Bioinformatika skripta,* https://www.fer.unizg.hr/_download/repository/bioinformatika_skripta_v1.2.pdf, 11. 6. 2020.
3. *SAM Format Specification*, https://samtools.github.io/hts-specs/SAMv1.pdf, 11. 6. 2020.
4. *Data File Formats*, https://genome.ucsc.edu/FAQ/FAQformat.html, 11. 6. 2020.
5. Khan Academy, *Eukaryotic pre-mRNA processing*, https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/eukaryotic-pre-mrna-processing, 11. 6. 2020.
6. Khan Academy, *Overview of transcription,* https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/overview-of-transcription, 11. 6. 2020.
7. Abecasis Group Wiki, *SAM*, https://genome.sph.umich.edu/wiki/SAM, 11. 6. 2020.
8. Minimap2, https://github.com/lh3/minimap2, 11.6.2020.
9. Heng Li, Bob Handsaker, John Marshall, Petr Danecek, *Samtools,* http://www.htslib.org/doc/1.2/samtools.html, 11.6.2020.

Sažetak

Vizualizacija podataka dobivenih sekvenciranjem RNA

Ovaj rad sadrži osnovnu bioinformatičku pozadinu elemenata i procesa koji su povezani sa sekvenciranjem RNA. Nadalje, govori se o samom procesu sekvenciranja i procesima mapiranja i poravnanja koji slijede nakon sekvenciranja.

Obrađuju se tipovi datoteka koji su dobiveni uz pomoć navedenih procesa. Za svaki tip se objašnjava format datoteke i što podatci unutar nje predstavljaju. Spominju se alati Mininap2 i Samtools koji služe za rad s navedenim datotekama.

Na kraju se govori o razvijenoj aplikaciji. Spominju se opće informacije vezane za razvoj aplikacije. Slikama se prikazuje izgled aplikacije, kao i njeni elementi. Svaki element je još dodatno opisan.

Ključne riječi: RNA, vizualizacija, bioinformatika , sekvenciranje, mapiranje, poravnanje

Summary

RNA Sequencing Data Visualization

This paper contains the basic bioinformatic background of elements and processes associated with RNA sequencing. Furthermore, it talks about the sequencing process itself and the processes of mapping and alignment that follow.

The information about file types obtained using the aforementioned processes are also contained in this paper. For each file type, the file format and the representation of data within it is explained. Mention is made of the Minimap2 and Samtools tools used to work with the specified files.

Lastly, the paper contains information about the developed application. General information related to the development of the application is mentioned. The appearance of the application is shown, as well as it's elements. Each element is further described.

Keywords: RNA, visualization, bioinformatics, sequencing, mapping, alignment

Skraćenice

BAM - Binary Alignment Map
BCF - BIM Collaboration Format
BED - Browser Extensible Data
CIGAR - Compact Idiosyncratic Gapped Alignment Report
DNA - Deoxyribonucleic acid
GFF – Generic Feature Format
GTF – Gene Transfer Format
RNA - Ribonucleic acid
SAM - Sequence Alignment Map
VCF – Variant Call Format