SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 6162

PRONALAŽENJE VARIJANTI GENA IZ PODATAKA DOBIVENIH SEKVENCIRANJEM

Sanja Kosier

Zagreb, lipanj 2019.

Zahvaljujem prof. dr. sc. Mili Šikiću i doc. dr. sc. Krešimiru Križanoviću na pomoći, savjetima i vodstvu.

Zahvaljujem izv. prof. dr. sc. Deanu Konjeviću na pruženim podatcima o uzorcima jelena običnog.

Veliko hvala obitelji i prijateljima na pruženoj podršci i motivaciji tijekom studiranja.

Sadržaj

1 Uvod

Aleli predstavljaju dva alternativna gena koja određuju istu osobinu. Po jedan alel nasljeđujemo od svakog roditelja, a nalaze se na paru homolognih kromosoma. Poznavanjem alela i njihovim pronalaskom u organizmu, mogu se donijeti zaključci o obilježjima organizma, a na posljetku i populacije. Motivacija za pronalazak alela u uzorcima organizama je velika. Naime, svim granama znanosti i znanstvenih disciplina koje na neki način proučavaju obilježja organizama, informacije iz područja genetike i bioinformatike mogu biti od koristi. Medicina i farmacija napreduju brzo, genska terapija sve je češći pojam koji se može čuti. Da bi uopće bilo moguće razviti lijek koji djeluje na određeni gen, potrebno je znati koje taj gen ima varijacije – alele.

Jedan gen može biti prisutan kod više vrsta i u ovisnosti o vrsti organizma koji se proučava, zastupljenost gena varira. Mutacije, tj. izmjene u našim genetskim sekvencama također su pojave koje se događaju.

Neki od problema vezani uz dobivanje i proučavanje alela dobivenih iz skupa podataka o genu unutar nekog uzorka u bioinformatici tiču se pronalaska alela podzastupljenih u uzorku i razlikovanja alela od mutacija.

Dio genetike koji se bavi proučavanjem promjena genske učestalosti, to jest brojčanim odnosom pojedinačnih alela u svakom paru alela unutar genske zalihe određene populacije, naziva se populacijska genetika.[1]

2 Pregled područja

2.1 Algoritmi poravnanja

Poravnanje dviju sekvenci (engl. *pairwise sequence alignment*) niz je transformacija koje opisuje na koji je način moguće dobiti jednu sekvencu od druge. Često je u bioinformatici korišteno za identifikaciju sličnih regija koje mogu indicirati na strukturalnu, funkcionalnu ili evolucijsku poveznicu između dviju sekvenci proteina ili DNA i RNA lanaca.

Tri su glavna tipa algoritama za poravnanje dviju sekvenci:

- Smith-Waterman algoritam (lokalno poravnanje)
- Needleman-Wunsch algoritam (globalno poravnanje)
- Algoritam preklapanja (polu-globalno poravnanje)

Algoritmi se temelje na dinamičkom programiranju budući da se prvo rješavaju manji potproblemi čiji se rezultati progresivno koriste na putu do konačnog rezultata. Naime, poravnanje dviju sekvenci obično je prikazano 2D mrežom poravnanja čije osi predstavljaju sekvence. Točka u matrici s koordinatama (n, m) odgovara paru pozicija na sekvenci, gdje je n-ta pozicija na prvoj sekvenci, a m-ta pozicija na drugoj sekvenci. Za svaku točku mreže, moguće je kretanje u 3 smjera – horizontalno u desno, vertikalno dolje ili dijagonalno. Dijagonalnom kretnjom s točke $(n,\ m)$ prelazimo u točku $(n + 1,\ m + 1)$ te ta kretnja predstavlja podudaranje ili neslaganje dviju sekvenci. Budući da se kretanjem horizontalno u desno mijenja poziciju samo druge sekvence, takva radnja predstavlja umetanje znaka u prvu sekvencu, dok će kretnja vertikalno prema dolje rezultirati brisanjem znaka iz prve sekvence.

Poravnanje dviju sekvenci u konačnici ovisi o odabranom putu kroz mrežu, a najčešće je prikazano u formatu naziva CIGAR. Tablica 1 opisuje elemente CIGAR formata korištenih u radu.

Tablica 1 Opis elemenata CIGAR formata korištenih u radu.

Oznaka u ispisu	Značenje
M	Baze na sekvencama se podudaraju
D	Brisanje baze na očitanju
I	Umetanje baze u očitanje
X	Baze na sekvencama se ne podudaraju
S	Baze na sekvenci su uklonjene

2.2 POA graf i konsenzus

Skraćenica POA dolazi od engleskog: „*partial order alignment*“. Takav tip poravnanja rezultira izgradnjom usmjerenog acikličkog grafa koji sadrži informacije vezane uz strukturu poravnatih sekvenci.

Konsenzusna sekvenca može se očitati iz jednom izgrađenog POA grafa korištenjem dinamičkog programiranja. Vrijednosti svih čvorova u izgrađenom grafu prvobitno su postavljene na nulu, a zatim se iterira kroz graf i vrijednosti se ažuriraju. Do svakog čvora se odabire put kroz koji najviše sekvenci prolazi, a vrijednost čvora jednaka je maksimalnoj sumi težine brida i rezultata čvora iz kojeg je usmjeren. Ovakav način odabiranja maksimalnih suma pri postavljanju vrijednosti čvorova grafa, rezultira odabiranjem puta najveće vjerojatnosti, a pripadajuće baze na čvorovima slijedno čine konsenzusnu sekvencu.

U radu se za generiranje konsenzusne sekvence i izgradnju POA grafa koristio alat SPOA (*https://github.com/rvaser/spoa*) koji je detaljnije opisan u radu: *Fast and accurate de novo genome assembly from long uncorrected reads.*[2]

3 Podatci

Gen čije varijabilnosti (alele) tražimo u uzorcima je gen MHC, jedan od glavnih gena sustava tkivne podudarnosti te ima važnu ulogu u obrani od parazita. Uzorci su dobiveni sekvenciranjem metodom Ion Torrent, a prikazani su FASTQ formatom, formatom za prikaz podataka dobivenih na uređajima za sekvenciranje. Podatci korišteni u radu dobiveni su na projektu HRZZ IP-2016-06-5751 naziva: „DNA kao dokaz o distribuciji i vitalnosti ugrožene Balkanske divokoze“. Uz podatke o uzorcima divokoza imamo na raspolaganju i dva uzorka (J29_B_CE_IonXpress_005 i J30_B_CE_IonXpress_006) s podacima jelena običnog koje nam je ustupio izv. prof. dr. sc. Dean Konjević s Veterinarskog fakulteta Sveučilišta u Zagrebu. Podaci su dobiveni na projektu HRZZ IP-2018-01-8963 pod nazivom: „Interakcija nositelj-parazit: odnos tri različita tipa nositelja prema invaziji metiljem *Fascioloides magna*“. Za uzorke jelena imamo rezultate o njihovim pronađenim alelima. Rezultati su prikazani u nastavku.

Već pronađene varijante gena MHC za jelena običnog u uzorku J29_B_CE_IonXpress_005:

>jelenref05 J29B-1_M13F-pUC 249bp

CTGTATGCTAAGAGCGAGTGTCATTTCTCCAACGGGACGCAGCGGGTGGGGTTCCTGGACA
GATACTTCTATAACGGAGAAGAGTTCGTGCGCTTCGACAGCGACTGGGGCGAGTACCGGG
CGGTGACAGAGCTGGGGCGGCCGGTGGCCGAGTACCTGAACAGCCAGAAGGAGTACATG
GAGCAGACGCGGGCCGAGGTGGACACGTACTGCAGACACAACTACGGCGGCGTTGAGAG
TTTCACTGTG

>jelenref06 J29B-3_M13F-pUC 249 bp

CTGTATACTACGAGCGAGTGTCATTTCTCCAACGGGACGCAGCGGGTGGGGTTCCTGGACA
GATACTTCTATAACGGAGAAGAGTACGTGCGCTTCGACAGCGACTGGGGCGAGTACCGGG
CGGTGACAGAGCTGGGGCGGCCGTCCGCCAAGTACTGGAACAGCCAGAAGGAGTACATG
GAGCAGACGCGGGCCGAGGTGGACAGGTACTGCAGACACAACTACGGGGTTCTTGACAGT
TTCGCTGTG

>jelenref07 J29B-6_M13F-pUC 249bp

GAGCATCATAAGTGCGAGTGTCATTTCTCCAACGGGACGGAGCGGGTGCAGTTCCTGCAGA
GATACATCTATAACCGGGAAGAGTACGTGCGCTTCGACAGCGACTGGGGCGAGTACCGGG
CGGTGACAGAGCTGGGGCGGCCGTCCGCCAAGTACTATAACAGCCAGAAGGAGCTCCTGG
AGCAGAAGCGGGCCGCGGTGGACAGGTACTGCAGACACAACTACGGGGTCGTTGAGAGTT
TCACTGTG

Već pronađene varijante gena MHC za jelena običnog u uzorku J30_B_CE_IonXpress_006:

>jelenref02 J16B-1_M13F-pUC 249 bp

GAGTATGCTAAGAGCGAGTGTCATTTCTCCAACGGGACGCAGCGGGTGCGGTTCCTGGAC
AGATACTTCTATAACCGGGAAGAGTACGTGCGCTTCGACAGCGACTGGGGCGAGTTCCGG
GCGGTGACCGAGCTGGGGCGGCCGTCCGCCAAGTACTGGAACAGCCAGAAGGATTTCATG
GAGCAGAAGCGGGCCGAGGTGGACACGGTGTGCAGACACAACTACGGGGTTATTGAGAG
TTTCACTGTG

>jelenref01 J16B-4_M13F-pUC 249 bp

GAGCATCTTAAGGCCGAGTGTCATTTCTTCAACGGGACGGAGCGGATGCAGTTCCTGGCGA
GATACTTCTATAACGGAGAAGAGTACGCGCGCTTCGACAGCGACGTGGGCGAGTTCCGGG
CGGTGACCGAGCTGGGGCGGCCGGACGCCAAGTACTGGAACAGCCAGAAGGAGATCCTG
GAGCAGCACCGGGCAGAGGTGGACAGGTACTGCAGACACAACTACGGGGTCGGTGAGAG
TTTCACTGTG

>jelenref04 J20B-10_M13F-pUC 249 bp

ATGTATACTAAGAAAGAGTGTCATTTCTCCAACGGGACGCAGCGGGTGGGGCTCCTGGACA
GATACTTCTATAACGGAGAAGAGTTCGTGCGCTTCGACAGCGACTGGGGCGAGTTCCGGG
CGGTGACCGAGCTGGGGCGGCCGGCGGCCGAGGGCTGGAACAGCCAGAAGGAGCTCCTG
GAGCAGAGGCGGGCCGCGGTGGACACGTACTGCAGACACAACTACGGGGTTATTGAGAGT
TTCACTGTG

4 Početna analiza uzoraka i očekivanja

4.1 Analiza uzoraka

Da bismo mogli zaključiti daje li algoritam koji smo razvili dobre i valjane rezultate poslužilo je korištenje obrnute logike. Umjesto da pokušavamo grupirati sekvence u klastere na različite načine i računati njihov konsenzus nadajući se da ćemo dobiti podatke koje očekujemo, analizirali smo uzorke J29_B_CE_IonXpress_005 i J30_B_CE_IonXpress_006 budući da za njih imamo pouzdane rezultate o tome koje bismo alele trebali dobiti. Napravljena je i njihova međusobna usporedba.

Da bismo utvrdili koliko su dva alela međusobno udaljena ili točnije, u kojoj se mjeri oni razlikuju, korišten je algoritam globalnog poravnanja s parametrima:

- 1 – za podudaranje
- 0 – za zamjenu
- -1 – za umetanje ili brisanje

Udaljenost dvaju alela tada je predstavljena kao razlika između veličine manjeg alela i vrijednosti njihovog poravnanja. Korištenjem takvog pristupa dobivena je međusobna razlika od 16 za alel jelenref05 i alel jelenref06, 30 za jelenref05 i jelenref07, a međusobna razlika alela jelenref06 i jelenref07 iznosi 26. Rezultati međusobnih udaljenosti za testne uzorke prikazani su u tablici 2. Možemo zaključiti da su nam prva dva aleli jelenref05 i jelenref06 ipak međusobno sličnija, dok je onaj treći (jelenref07) nešto različitiji od njih.

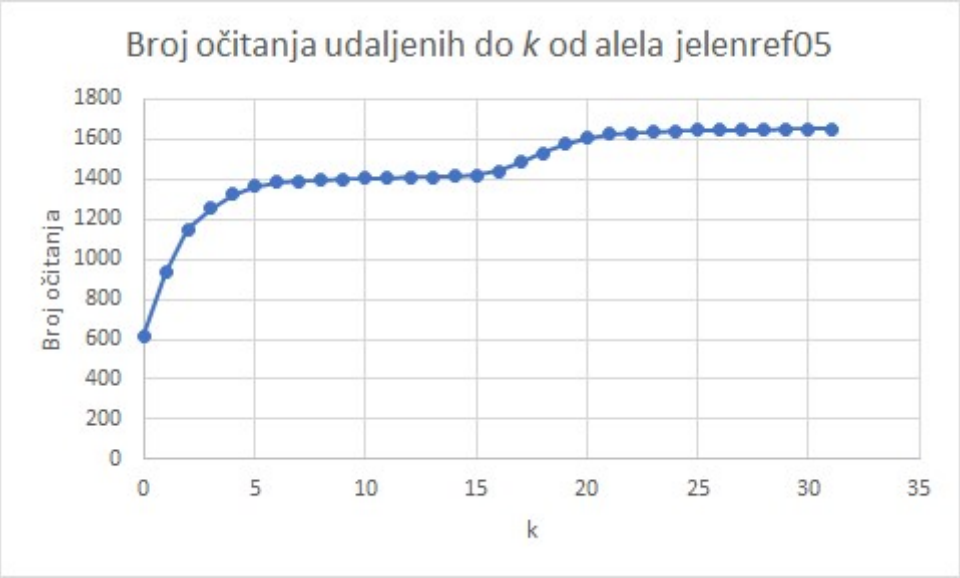
Nakon toga, provedena je analiza svih očitanja odgovarajućih duljina iz FASTQ datoteke, prvo uzorka J29_B_CE_IonXpress_005, a zatim uzorka J30_B_CE_IonXpress_006. Odgovarajuće duljine su predstavljene kao duljine najčešćih očitanja uz odstupanje od +/- 5 baza (objašnjeno kasnije u radu).

Tablica 2 Međusobne udaljenosti pronađenih varijanti gena za uzorak J29_B_CE_IonXpress_005.

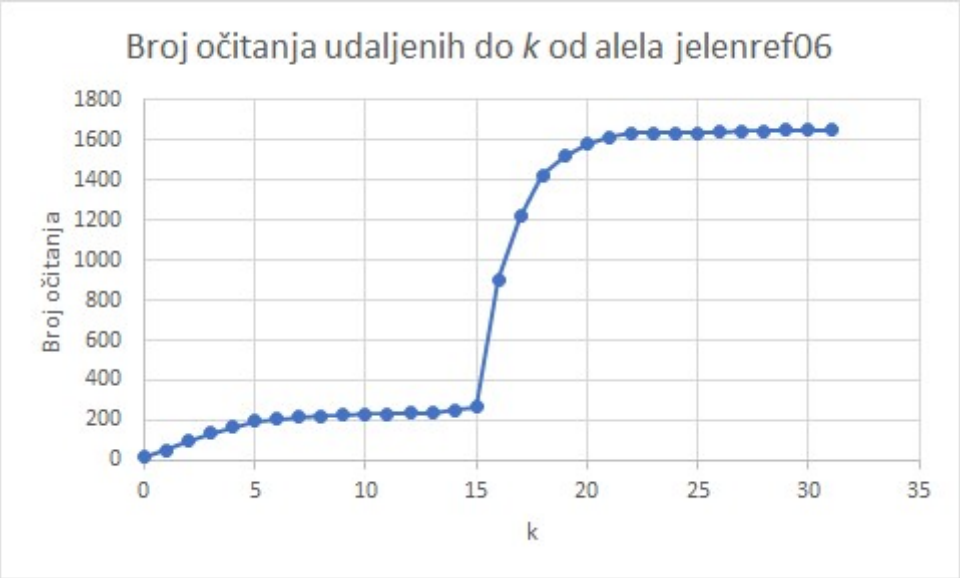
	jelenref05	jelenref06	jelenref07
jelenref05	0	16	30
jelenref06	16	0	26
jelenref07	30	26	0

Tehnika kojom sam se vodila pri analiziranju pojedinačnih alela bila je ta da sam gledala koliko je sekvenci udaljeno od dotičnog alela za neku udaljenost k, a da sam pri njihovom međusobnom računanju udaljenosti koristila već spomenuti pristup, tj. udaljenost je predstavljena kao razlika između veličine manje sekvence i vrijednosti njihovog poravnanja. Ukoliko je udaljenost manja od k, očitanje ulazi u klaster formiran oko poznatog alela.

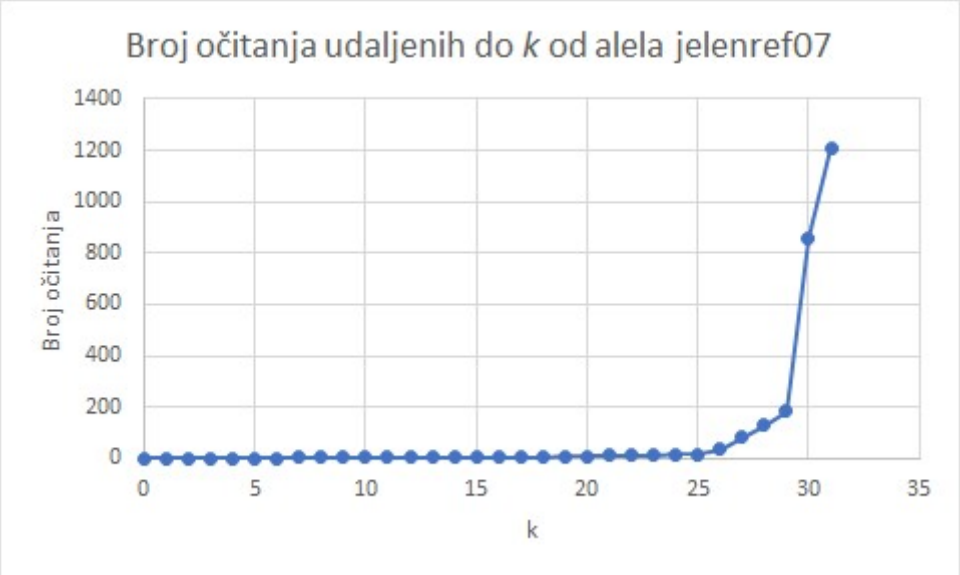
Rezultati za alele uzorka J29_B_CE_IonXpress_005 prikazani su na slikama 1, 2 i 3.



Slika 1 Prikaz ovisnosti broja očitanja o parametru udaljenosti k za alel jelenref05.



Slika 2 Prikaz ovisnosti broja očitanja o parametru udaljenosti k za alel jelenref06.



Slika 3 Prikaz ovisnosti broja očitanja o parametru udaljenosti k za alel jelenref07.

Za udaljenost $k = 5$ nema preklapanja u klasterima formiranim oko poznatih alela (niti jedno očitanje neće biti smješteno u dva takva klastera), budući da je iz tablice 2 vidljivo da su dva, već poznata alela, minimalno udaljena za 16. ($16 > 2k$). Iz grafova sa slika 1, 2 i 3 uočavamo da je za $k = 5$ broj bliskih očitanja alelu jelenref05 oko 1400, broj bliskih očitanja alelu jelenref06 oko 200, a alelu jelenref07 oko 4. Alel jelenref05 daleko je

najzastupljeniji. Rezultati govore da je čak 617 očitanja na udaljenosti 0 od njega samoga. U svakom slučaju, očitanja za detekciju tog alela ima dovoljno.

Zanimljivi su rezultati prikazani na slici 3. Samo su 4 očitanja udaljena manje od $k = 5$ od referentnog alela jelenref07. Štoviše, uzevši u obzir pogrešku sekvenciranja Ion Torrentom koja iznosi 1-3%, ne znamo jesu li ova 4 očitanja najbliža alelu jelenref07 samo posljedica greške. U svakom slučaju, tu se javljaju naznake problema, a to je da je alel jelenref07 jednostavno tako malo zastupljen da ćemo teško programskim rješenjem naći način za izdvojiti ga kao jednu od varijanti gena. Nije nepoznato u biologiji da su očitanja određene mutacije na genu ponekad češće i više zastupljene u uzorku od pojedinog alela. Ta se pojava javlja kao problem u bioinformatici – kako raspoznati mutaciju nad genom od varijante istog.

Rezultati su slični i za uzorak J30_B_CE_IonXpress_006. Bliskih očitanja oko alela jelenref02 i jelenref01 u konačnici ima dovoljno mnogo da bismo mogli detektirati prisutnost tih alela u uzorcima te svakako očekujemo njihov pronalazak u razvijenim metodama ovog rada. Za razliku od najmanje zastupljenog alela u prošlom uzorku, rezultati za alel jelenref04 nešto su optimističniji. Za udaljenosti 10 – 20 broj očitanja varira oko 20 te bismo dobrim razvojem metoda trebali izdvojiti i taj alel kao postojeći.

4.2 Očekivanje

Budući da smo analizom ustanovili da promatrani uzorci sadrže dovoljno očitanja bliskih alelima jelenref05 i jelenref06 za prvi uzorak, te alelima jelenref02 i jelenref01 drugog ispitnog uzorka, metode koje razvijamo ne bi trebale imati problema s pronalaskom upravo tih alela. Nadalje, iako je broj bliskih očitanja za treći alel u uzorku J30_B_CE_IonXpress_006 relativno malen, očekujemo da ih je dovoljno za detekciju, dok to nije slučaj za alel jelenref04 iz uzorka J29_B_CE_IonXpress_005.

5 Metode pronalaska varijanti gena

Sekvencirani podatci gena divokoza, kao i dva testna uzorka jelena običnog, analizirani su uzorak po uzorak u sljedećim razvijenim metodama:

a) Metoda velikog kažnjavanja
b) Gruba metoda
c) Pridruživanje najbližem klasteru

5.1 Metoda velikog kažnjavanja

Prva metoda od koje sam krenula (i koja je na posljetku dala prihvatljive rezultate) na početku je bila vođena idejom jednog sličnog istraživanja provedenog od strane dr. M. Bujanića pod nazivom: „Raznolikost gena glavnoga sustava tkivne podudarnosti jelena običnoga (*Cervus elaphus*) u odnosu na invaziju metiljem *Fascioloides magna*". Osim utvrđivanja postoje li razlike između različitih populacija jelena običnog i utvrđivanja možebitnu povezanost alela u odnosu na invadiranost metiljem *F. magna*, u radu navodi kako je još jedan od ciljeva istraživanja bio: „utvrditi varijabilnost MHC gena kod jelena običnoga u Hrvatskoj".[3]

Koraci analize podataka korištene u Bujanićevu radu su:

1. *Filtriranje očitanja po duljini.*
2. *Generiranje konsenzusne sekvence i višestrukog poravnanja sekvenci (engl. MSA – Multiple Sequence Alignment) za sva očitanja.*
3. *Podjela očitanja u skupine prema razlikama u odnosu na konsenzusnu sekvencu.*
4. *Ponavljanje točaka 2 i 3 za grupe očitanja dobivene u točki 3.*

5. *Određivanje reprezentativnih sekvenci za svaku grupu i odabir do četiri najbolje sekvence*.
6. *Usporedba s poznatim referencama i zbirna analiza.*

Koraci metode velikog kažnjavanja:

1. Filtriranje očitanja po duljini.
2. Generiranje konsenzusne sekvence i višestrukog poravnanja sekvenci za sva očitanja.
3. Grupiranje sličnih sekvenci u ovisnosti o parametru k u klastere.
4. Određivanje konsenzusnih sekvenci za svaku grupu i odabir do četiri najbolje sekvence.

U ovoj metodi bazirali smo se na pronalasku alela iz očitanja isključivo najčešćih duljina, a sva očitanja čija je duljina različita od tog broja, nisu sudjelovala u daljnjoj analizi. Zanimljivo je kako se razlikuje duljina najvećeg broj očitanja u uzorcima gena jelena običnog od duljine najvećeg broja očitanja u uzorcima divokoza. Naime, Bujanić navodi kako je taj broj kod jelena običnog iznosio 296 nukleotidskih baza, dok je analizom ustanovljeno da je najviše očitanja u uzorcima divokoza duljine 284 nukleotidskih baza.

Slijedila je izgradnja POA grafa koristeći alat SPOA (*https://github.com/rvaser/spoa*). Iz grafa je određeno višestruko poravnanje sekvenci te je rezultat njihovog međusobnog poravnanja zapisan u pomoću datoteku nastavka .msa. Ideja metode velikog kažnjavanja bila je uvelike kazniti ubacivanje ili brisanje baza, budući da ćemo konsenzusnu sekvencu generirati iz grupe sekvenci dobivenih upravo višestrukim poravnanjem. Iz tog razloga, bilo nam je bitno dobiti podatke tipa *string* bez oznake za te dvije kažnjavanje radnje: "-".

Shodno tome, parametri korišteni prilikom izgradnje POA grafa karakteristični su za ovu metodu.

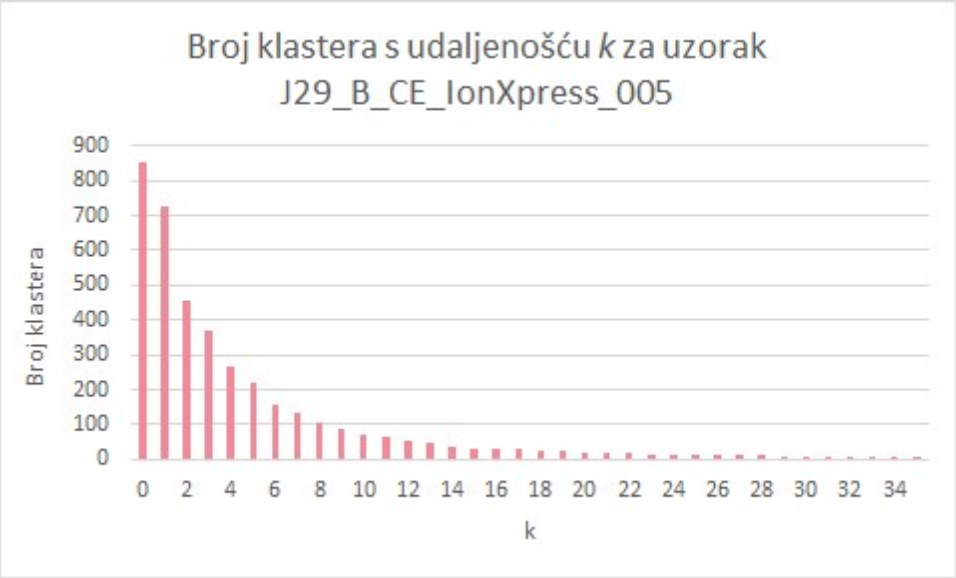
- 0 – za podudaranje
- -1 – za zamjenu
- -100 – za umetanje ili brisanje
- Globalno poravnanje

Nakon toga slijedi grupiranje sekvenci iz .msa datoteke u klastere na osnovi njihove međusobne sličnosti. U ovoj metodi sličnost i različitost dvaju očitanja određena je prema broju istih, odnosno različitih nukleotidskih baza na identičnim pozicijama pripadajućih sekvenci dobivenih višestrukim poravnanjem.

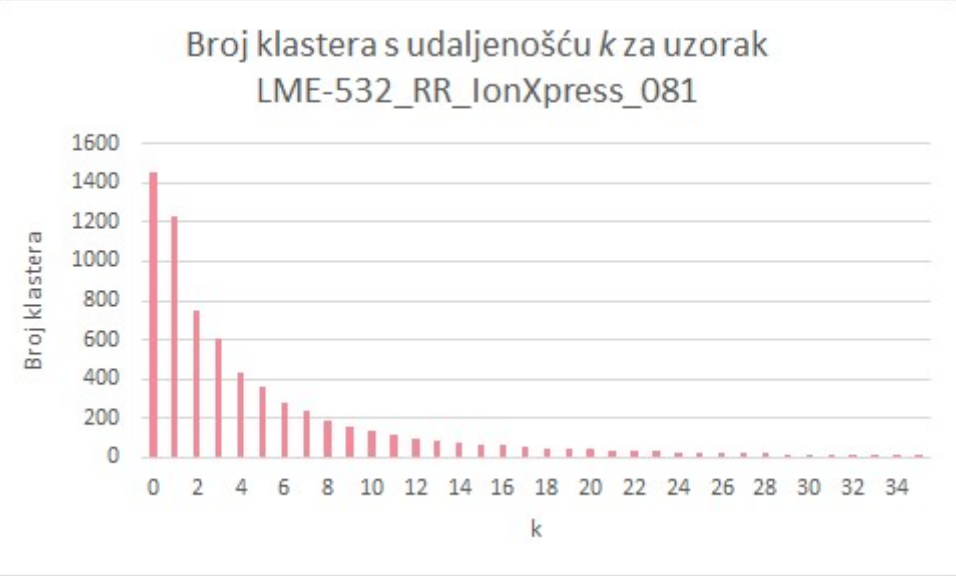
Ukoliko je određena sekvenca od sviju sekvenci iz klastera različita za manje od k, sekvenca se dodaje u klaster. U suprotnom, ukoliko sekvenca ne zadovoljava uvjet ni za jedan od već postojećih klastera, ona formira novi klaster. Korak se ponavlja za sve sekvence iz pripadajuće .msa datoteke

Po završetku klasteriranja, slijedi izdvajanje najvećih klastera, gdje je veličina klastera određena brojem očitanja koja se u njemu nalaze. Izdvajaju se najveći klasteri, uz uvjet da njihova veličina mora biti veća od parametra s.

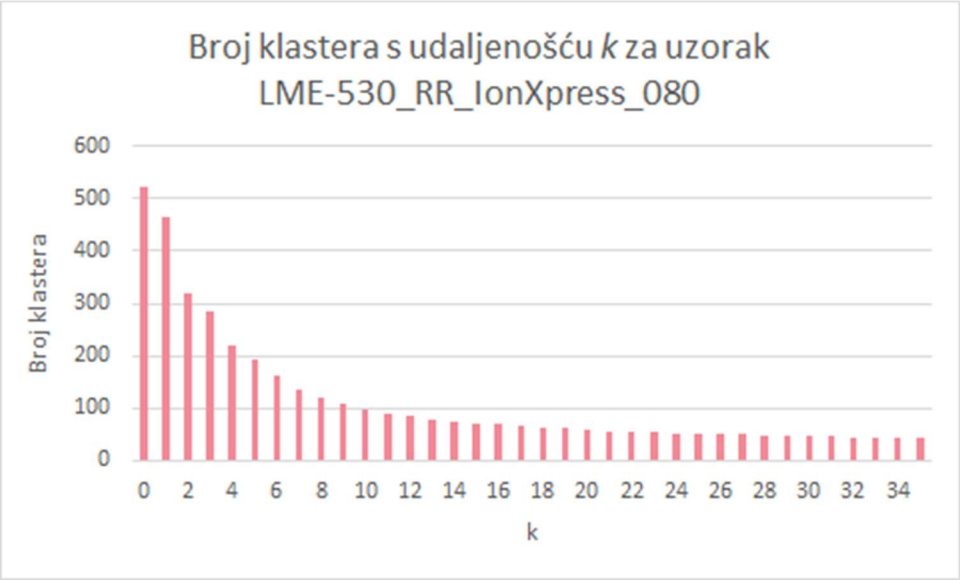
Pitanje koje se vrlo brzo počelo postavljati bilo je koja je vrijednost parametra k pogodna za određivanje hoće li određena sekvenca ući u neki klaster ili neće. Kako postavljanje defaultnog parametra k ne bi bilo samo puko pogađanje, napravljena je analiza u kojoj mjeri parametar k utječe na broj klastera. Analiza je napravljena nad uzorcima J29_B_CE_IonXpress_005, LME_530_RR_IonXpress_080 te LME_532_RR_IonXpress_081, a rezultati su prikazani na slikama 4, 5 i 6.



Slika 4 Slika prikazuje ovisnost broja klastera u ovisnosti u parametru k korištenom u metodi za uzorak J29_B_CE_IonXpress_005.



Slika 5 Slika prikazuje ovisnost broja klastera u ovisnosti u parametru k korištenom u metodi za uzorak LME-532_RR_IonXpress_081.



Slika 6 Slika prikazuje ovisnost broja klastera u ovisnosti u parametru k korištenom u metodi za uzorak LME-530_RR_IonXpress_080.

Vidljivo je kako nakon nekog k broj klastera počinje stagnirati, a da nakon $k = 12$ broj klastera padne ispod 100. Iz tog je razloga parametar k defaultno postavljen na 12, iz čega slijedi da da će očitanje ili poravnata sekvenca ući u klaster ukoliko je udaljena za manje od 12.

5.2 Gruba metoda

Koraci grube metode:

1. Filtriranje očitanja po duljini.
2. Generiranje višestrukog poravnanja.
3. Grupiranje sličnih očitanja u ovisnosti o parametru k u klastere.
4. Izdvajanje najvećih klastera.
5. Generiranje konsenzusne sekvence za izdvojene klastere, gdje konsenzusne sekvence predstavljaju pronađene alele.

U ovoj metodi proširili smo bazu očitanja na način da smo u obzir uzimali i ona očitanja čije duljine od one najčešće odstupaju za do 5 baza, a sva ostala, čija duljina proizlazi iz raspona $[duljina\_najčešćih\_očitanja - 5, duljina\_najčešćih\_očitanja + 5]$ ne sudjeluju u daljnjoj analizi.

Izgradnja .msa datoteke, provodi se na nešto drugačiji način. Naime, ovaj put se u datoteku ne zapisuju samo poravnate sekvence, nego i ime očitanja od kojeg je nastala poravnata sekvenca. Za razliku od metode velikog kažnjavanja, parametri korišteni kod izgradnje POA grafa su:

- 0 – za podudaranje
- -1 – za zamjenu
- -1 – za umetanje ili brisanje
- Globalno poravnanje

Raspodjela očitanja u klastere vrši se na isti način kao u prethodnoj metodi, ali s jednom razlikom. U klastere se ne dodaju već poravnate sekvence iz .msa datoteke, već se prema zapisanom imenu od kojeg je sekvenca došla, dohvaća očitanje iz pripadajuće FASTQ datoteke uzorka te se ono stavlja u

klaster. Pritom je međusobno računanje udaljenosti sekvenca ostalo isto – princip pregleda identičnih pozicija. Bitno je napomenuti da se u ovoj metodi konsenzusna sekvenca generira iz originalnih očitanja.

Po završetku klasteriranja, kao u metodi velikog kažnjavanja, slijedi izdvajanje najvećih klastera, gdje je veličina klastera određena brojem očitanja koja se u njemu nalaze. Izdvajaju se najveći klasteri, uz uvjet da njihova veličina mora biti veća od parametra s. Pronađene varijante gena predstavljeni su konsenzusnim sekvencama najvećih pronađenih klastera.

5.3 Metoda pridruživanja najbližem klasteru

Koraci metode pridruživanja najbližem klasteru:

1. Filtriranje očitanja po duljini.
2. Generiranje višestrukog poravnanja sekvenci.
3. Pridruživanje najbližem klasteru uz uvjet minimalne udaljenosti k.
4. Izdvajanje najvećih klastera.
5. Generiranje konsenzusne sekvence za izdvojene klastere, gdje konsenzusne sekvence predstavljaju pronađene alele.

Prva dva koraka metode u potpunosti se podudaraju s prva dva koraka grube metode.

Umjesto da očitanje dodajemo u prvi klaster za koji vrijedi uvjet da je pripadajuća poravnata sekvenca za k manja od svake sekvence u klasteru, ideja je bila pronaći najbliži klaster kojemu je očitanje najbliže. Pri tome i dalje mora vrijediti uvjet da je udaljenost manja od zadanog parametra k.

Kao što ime metode govori, sekvence se svrstavanje u najbliže klastere. Udaljenost sekvence od klastera ovdje je definirana kao srednja vrijednost svih udaljenosti sekvence koja se ispituje od svih sekvenci u klasteru. Naposljetku će sekvenca biti svrstana u onaj klaster čijim je sekvencama najbliža, odnosno u klaster s najmanjom srednjom udaljenosti.

Nakon svrstavanja očitanja u klastere, slijede koraci 4 i 5 koji se podudaraju s koracima 4 i 5 u već opisanoj, gruboj metodi.

6 Implementacija

6.1 Ulazni i izlazni formati

Program se sastoji od dva dijela. Prvi dio programa prima putanju do direktorija s uzorcima u FASTQ formatu, a rezultira direktorijem s rezultatima. Rezultati su u obliku datoteka s nastavkom .msa i sadrže rezultat pokretanja višestrukog poravnanja sekvenci za svaki od uzoraka.

Drugi dio programa ostvaren je kao konzolna aplikacija i njoj je potrebno predati dva argumenta: datoteku nastavka .msa generiranu u prvom dijelu te odgovarajući uzorak u FASTQ formatu. Program na standardni izlaz ispisuje pronađene varijante gena tog uzorka.

6.2 Programska implementacija

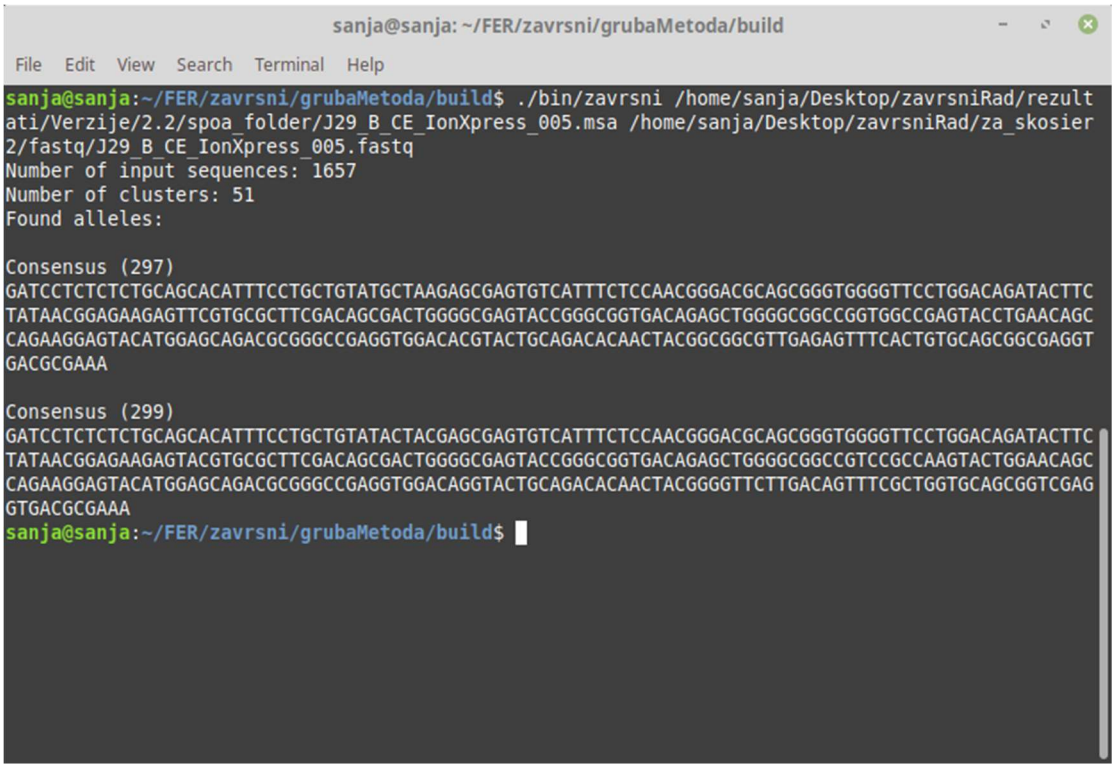
Implementacijski dio programa sastoji se od dva dijela. Prvi je dio u svim metodama odgovoran za korake broj 1 i broj 2 te je napravljen u programskom jeziku Python.

Drugi dio programa implementiran je u programskom jeziku C++ i odgovoran je za preostale korake metoda.

Oba programska ostvarenja ostvarena su kao konzolska aplikacija.

6.3 Primjer pokretanja programa

Primjer pokretanja programa prikazan je na slici 7. Za demonstraciju je odabran prikaz rezultata za testni uzorak J29_B_CE_IonXpress_005.



Slika 7 Pokretanje i rezultat programa.

7 Rezultati i diskusija

Za početak je bitno napomenuti da se u ispisu pronađenih sekvenci nalaze i početnice i završnice – sekvence koje su karakteristične za svaki gen i okružuju ga. Rezultati metoda bazirani su na testnim uzorcima MHC gena kod jelena običnog, budući da znamo koje rezultate očekujemo i razlikuju se od metode do metode. Metoda velikog kažnjavanja na uzorku J30_B_CE_IonXpress_006 kao rezultat izdvaja dvije varijante gena. Broj u zagradi predstavlja duljinu pronađene sekvence, u oba slučaja 296. Slijedi prikaz rezultata.

```
Consensus (296)

GATCCTCTCTCTGCAGCACATTTCCTGGAGCATCTTAAGGCCGAGTGTCATTTCTTC
AACGGGACGGAGCGGATGCAGTTCCTGGCGAGATACTTCTATAACGGAGAAGAGTAC
GCGCGCTTCGACAGCGACGTGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCG
GACGCCAAGTACTGGAACAGCCAGAAGGAGATCCTGGAGCAGCACGGGGCAGAGGTG
GACAGGTACTGCAGACACAACTACGGGGTCGGTGAGAGTTTCACTGTGCAGCGGCGA
GGTGACGCGAA

Consensus (296)

GATCCTCTCTCTGCAGCACATTTCCTGGAGTATGCTAAGAGCGAGTGTCATTTCTCC
AACGGGACGCAGCGGGTGCGGTTCCTGGACAGATACTTCTATAACCGGGAAGAGTAC
GTGCGCTTCGACAGCGACTGGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCG
TCCGCCAAGTACTGGAACAGCCAGAAGGATTTCATGGAGCAGAAGCGGGCCGAGGTG
GACACGGTGTGCAGACACAACTACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCGA
GGTGACGCGAA
```

Analiziranjem i pozivanjem poravnanja dviju sekvenci utvrdili smo da druga ispisana sekvenca u potpunosti odgovara testnom alelu jelenref02, a prva ispisana sekvenca odgovara alelu jelenref01, ali uz pogrešku. Međusobnim poravnanjem dobivenog i očekivanog alela dobijemo sljedeći CIGAR ispis: `27S 189= 1X 59= 20S`, što znači da je razlika između sekvenci u jednoj bazi. Metoda velikog kažnjavanja u konačnici je rezultirala točnim pronalaskom oba očekivana alela iz testnog uzorka J29_B_CE_IonXpress_005 (aleli jelenref04 i jelenref05). Iako smo u poglavlju 4.2 pretpostavili pronalazak sva tri testna uzorka, ovom metodom pronašli smo ih samo dva. S druge strane, druga metoda opisana u radu, gruba metoda,

daje bolje rezultate i pronalazi sve tri varijante MHC gena za uzorak J30_B_CE_IonXpress_006. Rezultati su prikazani u nastavku.

```
Consensus (299)

GATCCTCTCTCTGCAGCACATTTCCTGGAGCATCTTAAGGCCGAGTGTCATTTCTTC
AACGGGACGGAGCGGATGCAGTTCCTGGCGAGATACTTCTATAACGGAGAAGAGTAC
GCGCGCTTCGACAGCGACGTGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCG
GACGCCAAGTACTGGAACAGCCAGAAGGAGATCCTGGAGCAGCACCGGGCAGAGGTG
GACAGGTACTGCAGACACAACTACGGGGTCGGTGAGAGTTTCACTGTGCAGCGGCGA
GGTGACGCGAAAAA

Consensus (298)

GGATCCTCTCTCTGCAGCACATTTCCTGGAGTATGCTAAGAGCGAGTGTCATTTCTC
CAACGGGACGCAGCGGGTGCGGTTCCTGGACAGATACTTCTATAACCGGGAAGAGTA
CGTGCGCTTCGACAGCGACTGGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCC
GTCCGCCAAGTACTGGAACAGCCAGAAGGATTTCATGGAGCAGAAGCGGGCCGAGGT
GGACACGGTGTGCAGACACAACTACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCG
AGGTGACGCGAAA

Consensus (295)

GATCCTCTCTCTGCAGCACATTTCCTGATGTATACTAAGAAAGAGTGTCATTTCTCC
AACGGGACGCAGCGGGTGGGGCTCCTGGACAGATACTTCTATAACGGAGAAGAGTTC
GTGCGCTTCGACAGCGACTGGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCG
GACGCCGAGGCTGGAACAGACAGAAGGAGCTCCTGGAGCAGAGGCGGGCCGCGGTGG
ACACGTACTGCAGACACAACTACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCGAG
GTGACGCGAA
```

Prve dvije ispisane sekvence u potpunosti odgovaraju alelima jelenref02 i jelenref01 (respektivno). Treći uzorak nije u potpunosti identičan alelu jelenref04. CIGAR prikaz njihovih poravnanja je: `27S 145= 2X 7= 1I 9= 1X 84= 20S`. Iako razlika iznosi tri zamjene i jedno umetanje, pronalazak ovog alela svakako se može smatrati uspjehom. Metoda za uzorak J29_B_CE_IonXpress_005 pronalazi jelenref05, dok alel jelenref06 pronalazi s jednom razlikom u bazi (CIGAR: `27S 247= 1D 2= 22S`). Dvije su osnovne razlike između ove i prethodne metode koje su doprinijele boljim rezultatima. Korišten je veći broj očitanja i vrši se ponovno klasteriranje konsenzusnih sekvenci.

Metoda pridruživanja najbližem klasteru također pronalazi najmanje zastupljen alel u uzorku J30_B_CE_IonXpress_006. Prva ispisana sekvenca odgovara alelu jelenref02 s CIGAR ispisom `28S 110= 1D 11= 1D 17= 1D`

111= 21S, što znači da je potrebno obaviti tri operacije brisanja kako bi te dvije sekvence bile podudarne. CIGAR između druge sekvence i alela jelenref01 je: 27S 189= 1X 3= 1D 56= 23S, a poravnanje između najmanje zastupljene sekvence i alela jelenref04 glasi: 27S 145= 2X 7= 1I 9= 1X 84= 20S. U ovoj metodi također koristimo i sekvence duljine +/- 5 baza od duljine najčešćih sekvenci. Već je opisano da se klasteriranje vrši na način da se očitanje svrstava u najbliži klaster, gdje je blizina određena aritmetičkom udaljenošću od svih već postojećih očitanja u klasteru. Zbog toga je broj sekvenci u pojedinim klasterima nakon prve faze klasteriranja bio veći nego što je to bilo slučaj za nakon prve faze klasteriranja kod prošle metode. Slijedi prikaz dobivenih varijanti gena.

Consensus (301)

GGATCCTCTCTCTGCAGCACATTTCCTGGAGTATGCTAAGAGCGAGTGTCATTTCTC
CAACGGGACGCAGCGGGTGCGGTTCCTGGACAGATACTTCTATAACCGGGAAGAGTA
CGTGCGCTTCGACAGCGACTGGGGGCGAGTTCCGGGGCGGTGACCGAGCTGGGGGCG
GCCGTCCGCCAAGTACTGGAACAGCCAGAAGGATTTCATGGAGCAGAAGCGGGCCGA
GGTGGACACGGTGTGCAGACACAACTACGGGGTTATTGAGAGTTTCACTGTGCAGCG
GCGAGGTGACGCGAAA

Consensus (300)

GATCCTCTCTCTGCAGCACATTTCCTGGAGCATCTTAAGGCCGAGTGTCATTTCTTC
AACGGGACGGAGCGGATGCAGTTCCTGGCGAGATACTTCTATAACGGAGAAGAGTAC
GCGCGCTTCGACAGCGACGTGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCG
GACGCCAAGTACTGGAACAGCCAGAAGGAGATCCTGGAGCAGCACGGGGGCAGAGGT
GGACAGGTACTGCAGACACAACTACGGGGTCGGTGAGAGTTTCACTGTGCAGCGGCG
AGGTGACGCGAAAAA

Consensus (295)

GATCCTCTCTCTGCAGCACATTTCCTGATGTATACTAAGAAAGAGTGTCATTTCTCC
AACGGGACGCAGCGGGTGGGGCTCCTGGACAGATACTTCTATAACGGAGAAGAGTTC
GTGCGCTTCGACAGCGACTGGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCG
GACGCCGAGGCTGGAACAGACAGAAGGAGCTCCTGGAGCAGAGGCGGGCCGCGGTGG
ACACGTACTGCAGACACAACTACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCGAG
GTGACGCGAA

Kod uzorka J29_B_CE_IonXpress_005 pronađen je alel jelenref05, dok je varijanta jelenref06 pronađena uz grešku od jedne baze: 27S 247= 1D 2= 22S. Slijedi prikaz.

Consensus (298)

GATCCTCTCTCTGCAGCACATTTCCTGCTGTATGCTAAGAGCGAGTGTCATTTCTCC
AACGGGACGCAGCGGGTGGGGTTCCTGGACAGATACTTCTATAACGGAGAAGAGTTC
GTGCGCTTCGACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGGCCG
GTGGCCGAGTACCTGAACAGCCAGAAGGAGTACATGGAGCAGACGCGGGCCGAGGTG
GACACGTACTGCAGACACAACTACGGCGGCGTTGAGAGTTTCACTGTGCAGCGGCGA
GGTGACGCGAAAA

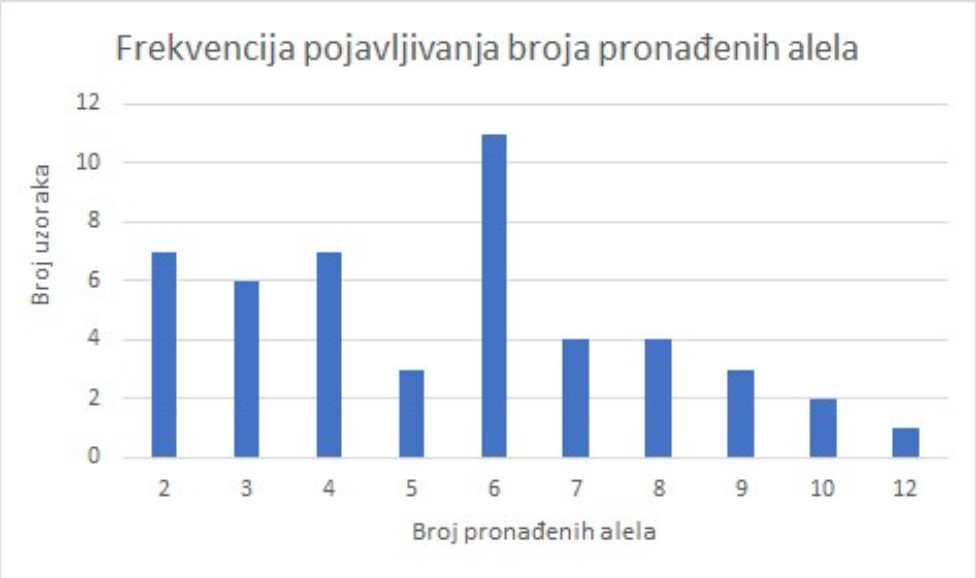
Consensus (299)

GATCCTCTCTCTGCAGCACATTTCCTGCTGTATACTACGAGCGAGTGTCATTTCTCC
AACGGGACGCAGCGGGTGGGGTTCCTGGACAGATACTTCTATAACGGAGAAGAGTAC
GTGCGCTTCGACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGGCCG
TCCGCCAAGTACTGGAACAGCCAGAAGGAGTACATGGAGCAGACGCGGGCCGAGGTG
GACAGGTACTGCAGACACAACTACGGGGTTCTTGACAGTTTCGCTGGTGCAGCGGTC
GAGGTGACGCGAAA

Napravljena je skupna analiza uzoraka korištenjem druge (grube) metode i metode pridruživanja najbližem klasteru. U analizi je sudjelovalo 48 uzoraka. Za svaki uzorak kroz iteraciju u programa spremane su informacije o broju alela za uzorak i broju klastera. U konačnici smo dobili rezultate o tome koliko ima ukupno različitih uzoraka u svim očitanjima, u koliko se uzoraka nalazi određeni alel i koji su to uzorci, te naravno popis alela po uzorku.

Korištenjem grube metode pronađeno je 260 alela, od kojih je 48 različitih. Za očekivati bi bilo da će se aleli ponavljati u uzorcima, i to doista i je tako. Tri najfrekventnija alela ponavljaju se u 22, 20 i 18 uzoraka.

Metoda pridruživanja najbližem klasteru daje nešto drugačije rezultate. Pronađeno je 266 alela, od kojih je 48 različitih. Najfrekventniji aleli ponavljaju se u 23, 21 i 19 uzoraka. Na slici 8 prikazano je koliko je pronađeno različitih alela u koliko uzoraka. Primjerice, u 11 uzoraka broj pronađenih alela je 6.



Slika 8 Frekvencija pojavljivanja broja pronađenih alela po broju uzoraka. Možemo primijetiti da u čak 11 uzoraka broj pronađenih (različitih) alela iznosi 6.

Iako smo uspjeli razviti metode koje detektiraju dobre varijante gena iz testnih uzoraka čije rezultate znamo, postoje faktori koji nam stvaraju poteškoće. Jedan je svakako podzastupljenost određenog alela u uzorku te malen broj očitanja na temelju kojih bi tu varijantu gena uspjeli pronaći. Naš program trenutno ne razlikuje pogrešno sekvencirane podatke od alela te će, ukoliko je više međusobno sličnih očitanja s pogreškom, njih grupirati zajedno i naposljetku prikazati njihov konsenzus kao varijantu gena.

Budući da su rezultati metoda nad ispitnim uzorcima ipak bili točniji korištenjem druge (grube) metode, vjerojatnije je da su i rezultati nad uzorcima divokoza točniji pri korištenju upravo te metode. Lako je moguće da bi uklanjanjem početnica u završnica u očitanjima rezultati bili drugačiji. Vjerojatno je da na ovaj način kada i oni sudjeluju u poravnanju, doprinose pogrešci.

8 Zaključak

Problem koji se vrlo brzo pojavio u implementaciji, bio je vezan uz veličinu parametra k. Postavljalo se pitanje po kojem bi parametru k bilo *najprirodnije* raditi klasteriranje, tj. koja je udaljenost ta koja će utvrditi grupe očitanja koji su biološkim procesima došli od istog alela.

Razvoj svake dobre metode traje te putu do onog konačnog rješenja prethodi mnogo pokušaja i pogrešnih metoda. Tako ni razvoj ovih metoda nije konačan, nego mjesta za napredak i daljnji rad ima napretek. Kada se razriješe problemi dobivanja istinitih alela iz očitanja najčešćih duljina, tematika istraživanja koja se nastavlja na ovu, mogla bi biti dobivanje alela iz svih očitanja u uzorku, budući da su rezultati doista dali bolje podatke za veći broj korištenih očitanja u analizi, a ne samo onih koji variraju oko najčešće duljine. Razvitak novih metoda klasteriranja i ponovnog klasteriranja također nije na odmet. Tu su i varijacije na generiranje međusobnog višestrukog poravnanja i konsenzusa koristeći razne parametre i vrste poravnanja (globalno, lokalno, polu-globalno).

9 Literatura

1. Populacijska genetika, https://hr.wikipedia.org/wiki/Populacijska_genetika
2. Vaser R., Sović I., Nagarajan N., Šikić M., Fast and accurate de novo genome assembly from long uncorrected reads, *Genome Res*. 2017;27(5):737–746. doi:10.1101/gr.214270.116
3. Bujanić, M., Raznolikost gena glavnoga sustava tkivne podudarnosti jelena običnoga (*Cervus elaphus*) u odnosu na invaziju metiljem *Fascioloides magna*, doktorska disertacija, Veterinarski fakultet, Zagreb, 2019.
4. Christopher Lee, Catherine Grasso, Mark F. Sharlow, Multiple sequence alignment using partial order graphs , *Bioinformatics*, Volume 18, Issue 3, ožujak 2002, stranice 452–464, https://doi.org/10.1093/bioinformatics/18.3.452
5. Christopher Lee, Generating consensus sequences from partial order multiple sequence alignment graphs, *Bioinformatics*, Volume 19, Issue 8, 22. svibanj 2003, stranice 999–1008, https://doi.org/10.1093/bioinformatics/btg109
6. Šikić, M., Domazet-Lošo, M., Bioinformatika: Optimalno poravnanje sljedova, Zagreb, prosinac 2013.

Pronalazak varijanti gena iz podataka dobivenih sekvenciranjem

Sažetak

Aleli predstavljaju dva alternativna gena koja određuju istu osobinu. Pronalazak alela iz višestrukih očitanja gena nije trivijalan. Događa se da su određeni aleli u uzorcima podzastupljeni te je teško programskim rješenjima detektirati njihovo pojavljivanje.

Cilj ovog rada bio je implementirati metode za pronalazak varijanti gena iz uzoraka dobivenih sekvenciranjem. Problem je riješen korištenjem programskih jezika C++ i Python, alata SPOA i Bioparser te algoritmima za poravnanje sekvenci. Razvijene metode čitaju uzorke iz FASTQ datoteka, koriste višestruko poravnanje sekvenci te vrše grupiranje sličnih očitanja u klastere. Nakon toga, odabiru se najveći klasteri na temelju broja očitanja u njima i računa se konsenzusna sekvenca koja naposljetku predstavlja pronađeni alel. U radu su se koristila očitanja MHC gena divokoza.

Ključne riječi: C++, Python, SPOA, višestruko poravnanje sekvenci, FASTQ, konsenzus, aleli

Discovering gene variants from sequencing data

Abstract

Alleles represent two alternative genes that determine the same feature of an organism. Finding alleles from multiple readings is not trivial. It is possible that certain alleles in the samples are underrepresented and it is difficult to detect their appearance using software solutions.

The aim of this paper was to implement methods for finding gene variants from sequencing data. The problem was solved using C++ and Python programming languages, SPOA and Bioparser tools as well as algorithms for pairwise alignment. Developed methods read samples from FASTQ files, use multiple sequence alignment and perform clustering. Furthermore, the largest clusters are selected based on the number of readings in the clusters and a consensus sequence is generated. A consensus sequence represents a found allele. In this paper the MHC gene samples of chaimos were used.

Key words: C++, Python, SPOA, multiple sequence alignment, FASTQ, consensus, alleles