Zahvaljujem svojem mentoru doc. dr. sc. Krešimiru Križanoviću na trudu, strpljenju i pomoći.

Zahvaljujem svojoj obitelji, prijateljima i kolegama na pruženoj podršci i razumijevanju tijekom studiranja i izrade ovog rada.

Sadržaj

Uvod

Genotip organizma sastoji se od niza gena. Svaki organizam ima jedinstven genotip, što objašnjava veliku raznolikost unutar pojedine vrste, ali i u živom svijetu općenito.

Geni se sastoje od DNK, složene molekule koja kodira genetske informacije za prijenos nasljednih osobina. (Rogers, n.d.) Geni kodiraju proteine koji utječu na imunološki sustav, pigmentaciju kože, boju očiju te mnoge druge osobine organizma. Alel je varijantni oblik gena. Detaljnije, svaki se gen nalazi na određenom mjestu na kromosomu u dvije kopije, pri čemu je jedan primjerak gena naslijeđen od svakog roditelja. (Gleichmann, 2020.) Te dvije kopije ne moraju nužno biti jednake. Kada se kopije gena međusobno razlikuju, one su poznate kao aleli. Dakako, određeni gen može imati više različitih alela, iako su kod svake diploidne jedinke prisutna samo dva alela. Također, u slučaju da je neki gen dupliciran, odnosno, ako se nalazi na dva ili više mjesta u genomu, jedan organizam može imati i više od dva različita alela.

Aleli rezultiraju različitim fenotipovima (uočljivim osobinama) s dominantnim alelima (koji nadjačavaju osobine drugih alela) ili, u nekim slučajevima, više alela koji djeluju kodominantno. Primjer potonjeg je krvna grupa AB u čovjeka, čiji domaćin ima jedan alel za grupu A i jedan za grupu B. (OpenStax, 2016.) Primjer dominantne izraženosti alela je tamna boja kose u čovjeka.

Poznavanje alela određenog gena važno je u mnogim slučajevima. Primjerice, kada se malarija uzrokovana *plazmodijem falciparumom* (parazitska bolest kod ljudi koju prenose zaražene ženke nekih komaraca) brzo i pravilno liječi, stopa smrtnosti iznosi 0,1%. Međutim, u nekim dijelovima svijeta parazit je razvio otpornost na najčešće korištene lijekove protiv malarije, stoga cjepiva i lijekovi variraju ovisno o geografskom položaju. (Vinayak, i dr., 2010.) Ovo je čest evolucijski fenomen koji se pojavljuje zbog toga što mutanti otporni na lijekove nastaju u populaciji i križaju se s drugim jedinkama u neposrednoj blizini. Iz tog razloga, potrebno je konstantno raditi na razvoju novih lijekova ili kombinacija lijekova za borbu protiv malarije. Da

bi se lijekovi koji utječu na određeni gen mogli razvijati, potrebno je poznavati različite alele tog gena.

Mutacije mogu uzrokovati proizvodnju proteina s krivom funkcijom ili nefunkcionalnih proteina i zato je općenito u biologiji poželjno otkriti pojavu mutacije na genima. U bioinformatici se zbog pojave mutacija javljaju problemi detekcije alela. Naime, postavlja se pitanje je li neka sekvenca za koju je, primjerice, teško odrediti pripadnost skupini već otkrivenih alela novi alel ili je samo rezultat mutacije.

Sekvenciranje DNK postupak je određivanje slijeda nukleinskih kiselina, odnosno redoslijeda nukleotida koji sačinjavaju DNK. (Nacionalno Vijeće za Istraživanje (S.A.D.), 1988.) Podaci korišteni u radu dobiveni su sekvenciranjem tehnologijom Ion Torrent. Ionsko poluvodičko sekvenciranje, poznato kao Ion Torrent sekvenciranje, metoda sekvenciranja DNK temeljena na detekciji vodikovih iona koji se oslobađaju tijekom polimerizacije DNK. (Wikipedia, Ion semiconductor sequencing, 2020.) Polimerizacija DNK je proces u kojem se jedna molekula DNK udvostručuje te nastaju dvije istovjetne kopije. (Leksikografski zavod Miroslav Krleža, 2020.) Pogreška sekvenciranja Ion Torrentom iznosi 1-3%, stoga je poželjno moći razlikovati pogrešna očitanja od alela.

Himerni geni tvore se kombinacijom dijelova dviju ili više sekvenci i tako stvaraju novi gen. (Wikipedia, Chimeric gene, 2018.) Himerna očitanja su umjetne sekvence konstruirane od dvije ili više pogrešno spojene biološke sekvence. (NCBI GenBank, Chimera Detection in 16S rRNA Sequences at NCBI, 2018.) Ovakva očitanja mogu se pojaviti prilikom sekvenciranja. Himerna očitanja nisu detektirana u radu, međutim njih bi također bilo korisno izuzeti iz analize.

Uzimajući u obzir navedene izazove, cilj je ovog rada razviti nove i implementirati poznate algoritme nenadziranog učenja, točnije grupiranja (engl. *clustering*) varijanti gena u skupine koje će označiti reprezentativni alel određene vrste. Algoritmi bi idealno trebali iz skupa podataka sekvenciranih gena detektirati alele u tom skupu i uspješno zanemariti pogreške nastale unutar organizma ili pri

sekvenciranju. Naravno, ti su problemi vrlo složeni, ali odabirom prikladnih algoritama i metoda pokušat će se približiti idealnom cilju.

1. Korištena tehnologija i alati

1.1. Algoritmi poravnanja

U radu je korištena vlastita implementacija algoritama poravnanja dviju sekvenci (engl. *Pairwise sequence alignment*). Poravnanje bioloških sekvenci vrlo je važan korak u bioinformatičkoj analizi te zbog svoje široke primjene predstavlja jedan od najstarijih i najviše istraživanih problema u bioinformatici. (Šikić & Domazet-Lošo, 2013.) Ovi algoritmi koriste se u raznim područjima i u raznim izvedbama, međutim svima je zajednički pojam dinamičkog programiranja. Dinamičko programiranje je algoritamska tehnika rješavanja problema rastavljanjem istoga na jednostavnije potprobleme i korištenjem činjenice da optimalno rješenje početnog problema ovisi o optimalnom rješenju njegovih potproblema. (Paljak & Hadviger, 2015.) Tri algoritma za računanje poravnanja dviju sekvenci su:

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

Cilj računanja poravnanja je pronaći optimalno poravnanje – ono za koje će udaljenost uređivanja (engl. *edit distance*) dviju sekvenci biti najmanja, odnosno sličnost najveća. Udaljenost uređivanja označava broj operacija nad jednim znakom potrebnih da bi se jedna sekvenca pretvorila u drugu. Operacija nad znakom može biti umetanje, brisanje i zamjena. Kod umetanja, znak će se dodati u sekvencu s_1 da bi se dobila sekvenca s_2 (npr. iz ATG u AT**G**G). Brisanje se obavlja nad znakom u sekvenci s_1 da bi se dobila sekvenca s_2 na način da se jedan znak izuzme (npr. GC**T**A u GCA). Kod zamjene jedan će se znak u sekvenci s_1 zamijeniti na određenoj poziciji tako da odgovara sekvenci s_2 (npr. G**T**A u G**C**A).

Algoritmi poravnanja svode se na računanje elemenata matrice dimenzije (n+1) * (m+1), gdje su n i m duljine nizova čije se poravnanje računa. Matrica poravnanja prikazuje se kao mreža čija dva susjedna brida označavaju sekvence u izvornom poretku. Ako su sekvence postavljene na gornji i lijevi brid matrice, točka u matrici

na poziciji (*x*, *y*) označava par *x*-te pozicije prve sekvence i *y*-te pozicije druge sekvence (gdje je pozicija (*0*, *0*), odnosno praznine prije početka svake sekvence u gornjem lijevom kutu matrice). Iz svake točke mreže moguće je kretati se u 3 smjera: po duljini prve sekvence, kada se iz točke (*x*, *y*) prelazi na poziciju (*x+1*, *y*). Ovo kretanje za sekvencu *s* čiji indeksi označavaju redove matrice predstavlja umetanje, *s* obzirom na to da ona zadržava svoj indeks, dok se kretanje obavilo po drugoj sekvenci. Drugi je smjer kretanje duž druge sekvence kada se iz pozicije (x, y) prelazi na poziciju (*x*, *y+1*). Ovakvo kretanje za sekvencu *s* predstavlja brisanje jednog elementa. Posljednje kretanje je dijagonalno (duž obiju sekvenci) kada se s pozicije (*x*, *y*) prelazi na poziciju (*x+1*, *y+1*). Ono predstavlja slaganje ili neslaganje između dva elementa niza. Svakom od ovih kretanja pridaje se težina koja predstavlja trošak te operacije. Cilj je pronaći poravnanje za koje će ukupni trošak biti najmanji. To se postiže tako da se svakom elementu matrice pridaje minimalna vrijednost (zbroj težine kretanja i vrijednosti elementa matrice iz kojeg se kretanje dogodilo). Počevši od pozicije (*0*, *0*) može se odrediti trošak puta do svakog elementa matrice i konačno odabrati optimalan put (onaj čiji je trošak minimalan).

1.2. Bioparser

Za parsiranje podataka koji su analizirani u radu korišten je bioparser[1]. Bioparser je biblioteka implementirana u jeziku C++, a omogućava parsiranje raznih formata korištenih za prikaz podataka u bioinformatici (FASTA, FASTQ, MHAP, PAF, SAM).

1.3. Spoa

Partial Order Alignment (POA) omogućuje izgradnju i analizu poravnanja više sekvenci kao usmjerenih acikličkih grafova koji sadrže složenu strukturu grananja. (Lee, 2003)

Vrijednosti čvorova u grafu prvobitno su postavljene na nulu. Počinje se umetati sekvence u graf i vrijednosti čvorova se ažuriraju. Pri biranju puta potrebno je odabrati onaj s najvećom težinom, tako je vrijednost čvora jednaka maksimalnoj

[1] https://github.com/rvaser/bioparser

sumi težine brida i vrijednosti čvora iz koje je brid usmjeren. Konsenzusna sekvenca očitava se iz grafa prateći težine bridova između odgovarajućih čvorova, a pripadajući elementi čvorova slijedno čine konsenzus. (Wikipedia, Multiple sequence alignment, 2020.)

Višestruko poravnanje sekvenci (engl. MSA, *Multiple Sequence Alignment*) je poravnanje triju ili više bioloških sekvenci. (Lipman, Altschul, & Kececioglu, 1989.)

U radu je za potrebe generiranja konsenzusne sekvence i višestrukog poravnanja korišten alat Spoa[2] (engl. *SIMD partial order alignment*).

1.4. Googletest

Za testiranje komponenata programa korišten je Googletest[3], okolina za testiranje koja među inim podržava *unit* testove.

1.5. CMake

Biblioteke bioparser, SPOA i googletest integrirane su u projekt pomoću alata CMake[4].

1.6. R

Grafovi prikazani u radu kreirani su koristeći programski jezik R[5] te razvojnu okolinu RStudio[6].

[2] https://github.com/rvaser/spoa
[3] https://github.com/google/googletest
[4] https://cmake.org/overview
[5] https://www.r-project.org/about.html
[6] https://rstudio.com/about/

2. Podaci

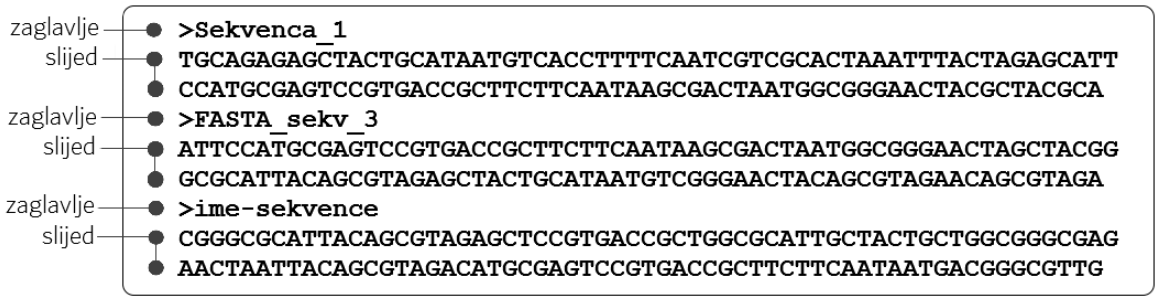
2.1. Formati podataka

Podaci analizirani u radu zapisani su u FASTA i FASTQ tekstualnom formatu.

2.1.1. FASTA

FASTA tekstualni format služi za prikaz sljedova proteina ili nukleinskih kiselina pri čemu je svaki nukleotid ili aminokiselina prikazan jednim slovom. (Šikić & Domazet-Lošo, 2013.) Zapis u FASTA formatu počinje jednolinijskim opisom iza kojeg slijede redovi samog slijeda. Definicija slijeda (zaglavlje) razlikuje se od slijeda znakom veće-od (>) na početku reda. Riječ koja slijedi nakon znaka „>“ (bez razmaka između) je identifikator slijeda, a ostatak linije je opcionalan i označava opis slijeda. (Hosseini, Pratas, & Pinho, 2016)

Primjer zapisa u FASTA formatu prikazan je na slici (Slika 2.1).



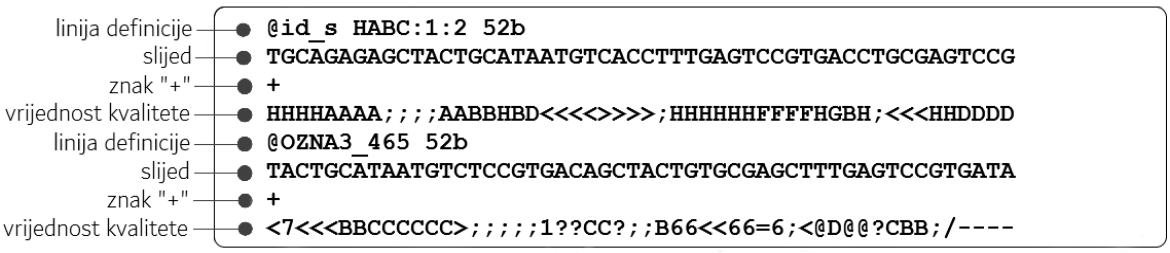
Slika 2.1 Primjer zapisa u FASTA formatu

2.1.2. FASTQ

FASTQ format podataka koristi se za prikaz podataka na izlazu uređaja za sekvenciranje. (Šikić & Domazet-Lošo, 2013.) FASTQ format nalikuje FASTA formatu, no može sadržavati eventualne dodatne informacije (dodatnu liniju identifikacije, ocjenu kvalitete svake baze). Linija definicije započinje znakom „@“ iza čega slijedi identifikator zapisa i dodatne informacije. Sljedeća linija ili više njih sadržavaju sami slijed. Treća linija može služiti za opis dodatnih informacija uz identifikator koji dolazi nakon znaka „+“, a može biti i prazna uz znak „+“ na početku. Posljednja linija

sadrži vrijednost kvalitete slijeda. Broj znakova ove linije mora biti jednak broju znakova u slijedu. (Hosseini, Pratas, & Pinho, 2016)

Primjer zapisa u FASTQ formatu prikazan je na slici (Slika 2.2).



Slika 2.2 Primjer zapisa u FASTQ formatu

2.2. Prikupljeni podaci

Gen čije je alele potrebno pronaći dio je gena MHC (engl. *the Major Histocompatibility Complex*). MHC je skupina gena koji pomažu imunološkom sustavu pri prepoznavanju stranih tvari, a nalaze se u organizmima svih viših kralježnjaka. (Britannica, n.d.) Kao što je već spomenuto, uzorci gena dobiveni su sekvenciranjem metodom Ion Torrent čija pogreška iznosi 1-3%.

Podaci analizirani u radu dobiveni su na projektu „Interakcija nositelj-parazit: odnos tri različita tipa nositelja prema invaziji metiljem *Fascioloides magna*“[7]. Uz podatke za analizu, na raspolaganju su i dva uzorka (J29_B_CE_IonXpress_005 i J30_B_CE_IonXpress_006) s već pronađenim varijantama gena MHC za jelena običnog.

Aleli gena MHC jelena običnog u uzorku J29_B_CE_IonXpress_005:

>jelenref05 J29B-1_M13F-pUC 249bp

CTGTATGCTAAGAGCGAGTGTCATTTCTCCAACGGGACGCAGCGGGTGGGGTTCCTG
GACAGATACTTCTATAACGGAGAAGAGTTCGTGCGCTTCGACAGCGACTGGGGCGAG
TACCGGGCGGTGACAGAGCTGGGGCGGCCGGTGGCCGAGTACCTGAACAGCCAGAAG
GAGTACATGGAGCAGACGCGGGCCGAGGTGGACACGTACTGCAGACACAACTACGGC
GGCGTTGAGAGTTTCACTGTG

[7] https://pdb.irb.hr/project/irb:004557

>jelenref06 J29B-3_M13F-pUC 249 bp

CTGTATACTACGAGCGAGTGTCATTTCTCCAACGGGACGCAGCGGGTGGGGTTCCTG
GACAGATACTTCTATAACGGAGAAGAGTACGTGCGCTTCGACAGCGACTGGGGCGAG
TACCGGGCGGTGACAGAGCTGGGGCGGCCGTCCGCCAAGTACTGGAACAGCCAGAAG
GAGTACATGGAGCAGACGCGGGCCGAGGTGGACAGGTACTGCAGACACAACTACGGG
GTTCTTGACAGTTTCGCTGTG

>jelenref07 J29B-6_M13F-pUC 249bp

GAGCATCATAAGTGCGAGTGTCATTTCTCCAACGGGACGGAGCGGGTGCAGTTCCTG
CAGAGATACATCTATAACCGGGAAGAGTACGTGCGCTTCGACAGCGACTGGGGCGAG
TACCGGGCGGTGACAGAGCTGGGGCGGCCGTCCGCCAAGTACTATAACAGCCAGAAG
GAGCTCCTGGAGCAGAAGCGGGCCGCGGTGGACAGGTACTGCAGACACAACTACGGG
GTCGTTGAGAGTTTCACTGTG

Aleli gena MHC jelena običnog u uzorku J30_B_CE_IonXpress_006:

>jelenref02 J16B-1_M13F-pUC

GAGTATGCTAAGAGCGAGTGTCATTTCTCCAACGGGACGCAGCGGGTGCGGTTCCTG
GACAGATACTTCTATAACCGGGAAGAGTACGTGCGCTTCGACAGCGACTGGGGCGAG
TTCCGGGCGGTGACCGAGCTGGGGCGGCCGTCCGCCAAGTACTGGAACAGCCAGAAG
GATTTCATGGAGCAGAAGCGGGCCGAGGTGGACACGGTGTGCAGACACAACTACGGG
GTTATTGAGAGTTTCACTGTG

>jelenref01 J16B-4_M13F-pUC

GAGCATCTTAAGGCCGAGTGTCATTTCTTCAACGGGACGGAGCGGATGCAGTTCCTG
GCGAGATACTTCTATAACGGAGAAGAGTACGCGCGCTTCGACAGCGACGTGGGCGAG
TTCCGGGCGGTGACCGAGCTGGGGCGGCCGGACGCCAAGTACTGGAACAGCCAGAAG
GAGATCCTGGAGCAGCACCGGGCAGAGGTGGACAGGTACTGCAGACACAACTACGGG
GTCGGTGAGAGTTTCACTGTG

>jelenref04 J20B-10_M13F-pUC

ATGTATACTAAGAAAGAGTGTCATTTCTCCAACGGGACGCAGCGGGTGGGGCTCCTG
GACAGATACTTCTATAACGGAGAAGAGTTCGTGCGCTTCGACAGCGACTGGGGCGAG
TTCCGGGCGGTGACCGAGCTGGGGCGGCCGGCGGCCGAGGGCTGGAACAGCCAGAAG
GAGCTCCTGGAGCAGAGGCGGGCCGCGGTGGACACGTACTGCAGACACAACTACGGG
GTTATTGAGAGTTTCACTGTG

Podaci koji su također na raspolaganju dobiveni su na istraživačkom projektu HRZZ-a pod nazivom „DNA kao dokaz o distribuciji i vitalnosti ugrožene Balkanske divokoze“[8].

[8] http://balkcham.agr.hr/

3. Analiza uzoraka

3.1. Analiza poznatih alela

Pri analizi uzoraka početno su promatrani već poznati aleli iz uzoraka J29B_expected i J30B_expected. Kako bi algoritmi grupiranja u konačnici trebali pronaći konsenzuse grupa koje označavaju alele, već poznati aleli su međusobno uspoređeni kako bi se dobila ideja o udaljenosti alela i vrijednostima parametara koji će se koristiti u daljnjoj analizi (primjerice, koja udaljenost će označavati pripadnost pojedinom klasteru, a koja formiranje novog klastera).

Za usporedbu alela iz uzoraka već poznatih korišten je algoritam Needleman-Wunsch, globalno poravnanje, s parametrima:

- 1 (podudaranje)
- 0 (zamjena)
- -1 (umetanje ili brisanje)

Udaljenost dvaju alela tada je izračunata kao razlika između veličine manjeg alela i vrijednosti dobivene algoritmom međusobnog poravnanja. Analiza je provedena nad svim parovima alela u uzorcima J29B_expected i J30B_expected i rezultati su prikazani u tablici (Tablica 3.1). Aleli koji pripadaju uzorku J29B_expected su jelenref05, jelenref06 i jelenref07, a aleli iz uzorka J30B_expected su jelenref02, jelenref01 i jelenref04.

Moguće je primijetiti da najmanja razlika dobivena korištenjem prethodno opisanog postupka iznosi 16, između alela jelenref05 i jelenref06. Najveća se pak razlika pojavljuje između alela jelenref01 i jelenref05, a iznosi 36. Može se reći da je alel jelenref01 najrazličitiji od svih pronađenih alela što potencijalno ukazuje na to da će ovaj alel biti najlakše pronaći.

Ipak, mnogi drugi faktori utjecat će na pronalazak alela u testnim uzorcima, kao što je primjerice zastupljenost pojedinog alela u uzorku.

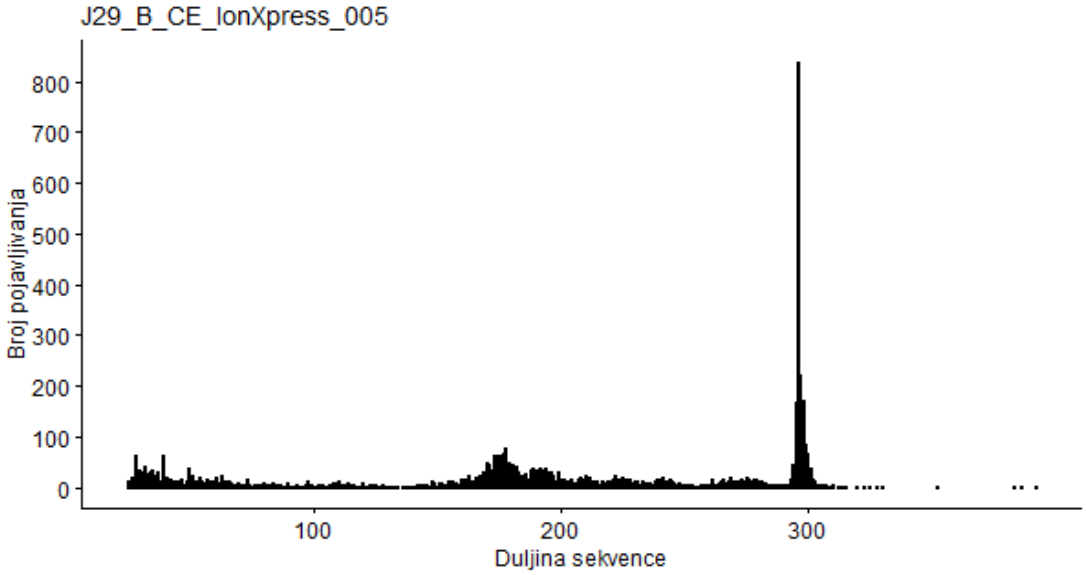
Tablica 3.1 Međusobna usporedba alela iz uzoraka J29B_expected i J30B_expected

	jelenref05	**jelenref06**	**jelenref07**	**jelenref02**	**jelenref01**	**jelenref04**
jelenref05	0	16	30	24	36	21
jelenref06	16	0	26	19	34	22
jelenref07	30	26	0	23	25	31
jelenref02	24	19	23	0	31	23
jelenref01	36	34	25	31	0	35
jelenref04	21	22	31	23	35	0

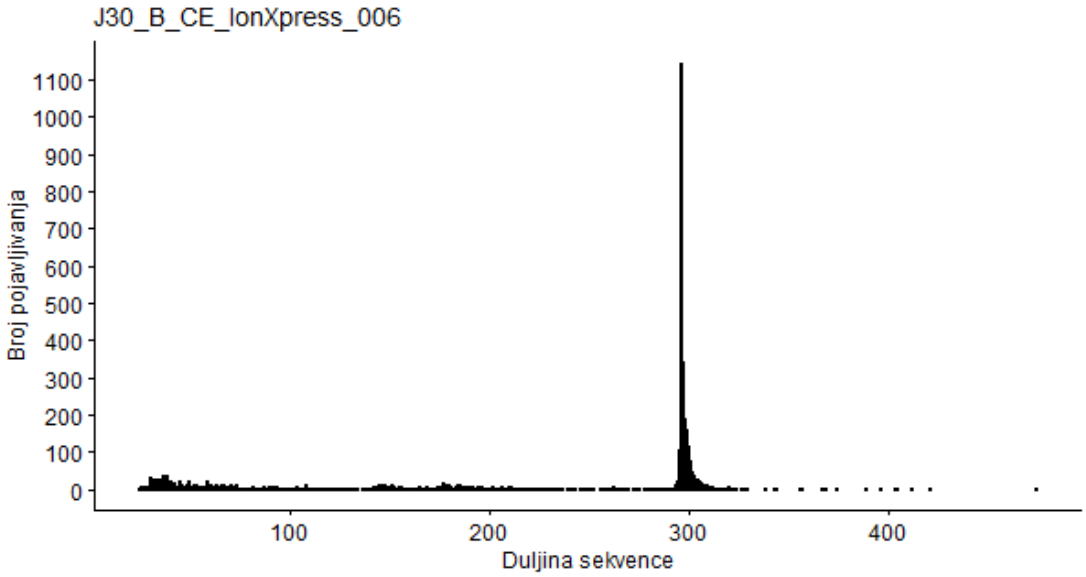
3.2. Analiza neoznačenih podataka

Da bi se ispitala zastupljenost pojedinog alela u uzorku provedena je analiza nad očitanjima u uzorcima J29_B_CE_IonXpress_005 i J30_B_CE_IonXpress_006. Analiza se temelji na usporedbi svih očitanja iz spomenutih uzoraka s referentnim alelima u uzorcima J29B_expected i J30B_expected. Udaljenost između dvije sekvence računa se na sličan način kao kod usporedbe poznatih alela, no razlika je u tome da se sada koristi polu-globalno poravnanje koje ne kažnjava praznine na početku i na kraju. Ova promjena je napravljena zato što se na očitanjima gena na izlazu iz uređaja za sekvenciranje nalaze početnice i završnice – sekvence specifične za pojedini gen koje se nalaze na početku i završetku svakog gena. Njih prilikom usporedbe nije poželjno uzeti u obzir i stoga se za izračun poravnanja koristi spomenuti algoritam. Analizom duljina sekvenci (Slika 3.1, Slika 3.2) pronađen je podatak da najčešća

duljina iznosi 296 nukleotidnih baza s 838 pojavljivanja u uzorku J29_B_CE_IonXpress_005 te 1144 pojavljivanja u uzorku J30_B_CE_IonXpress_006.



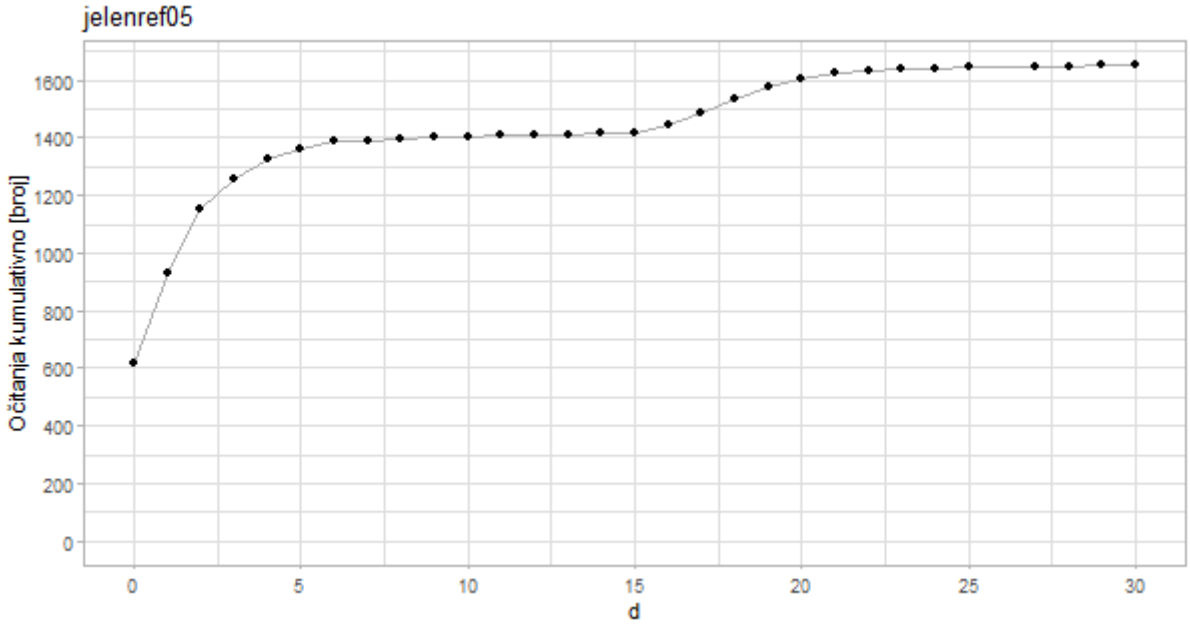
Slika 3.1 Broj pojavljivanja sekvenci u ovisnosti o duljini za uzorak J29_B_CE_IonXpress_005



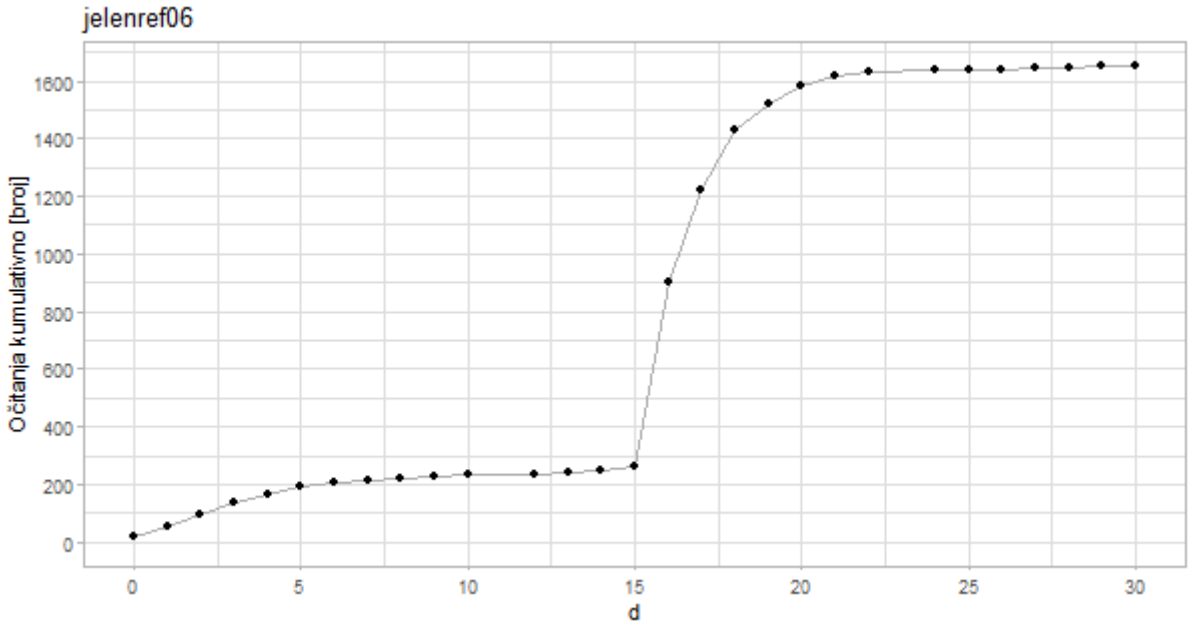
Slika 3.2 Broj pojavljivanja sekvenci u ovisnosti o duljini za uzorak J30_B_CE_IonXpress_006

Zanimljivo je primijetiti kako u uzorku J29_B_CE_IonXpress postoji puno više kraćih očitanja nego što je slučaj za uzorak J30_B_CE_IonXpress. U početnu su analizu uključena sva očitanja najčešće duljine (296) +/- 5 nukleotidnih baza.

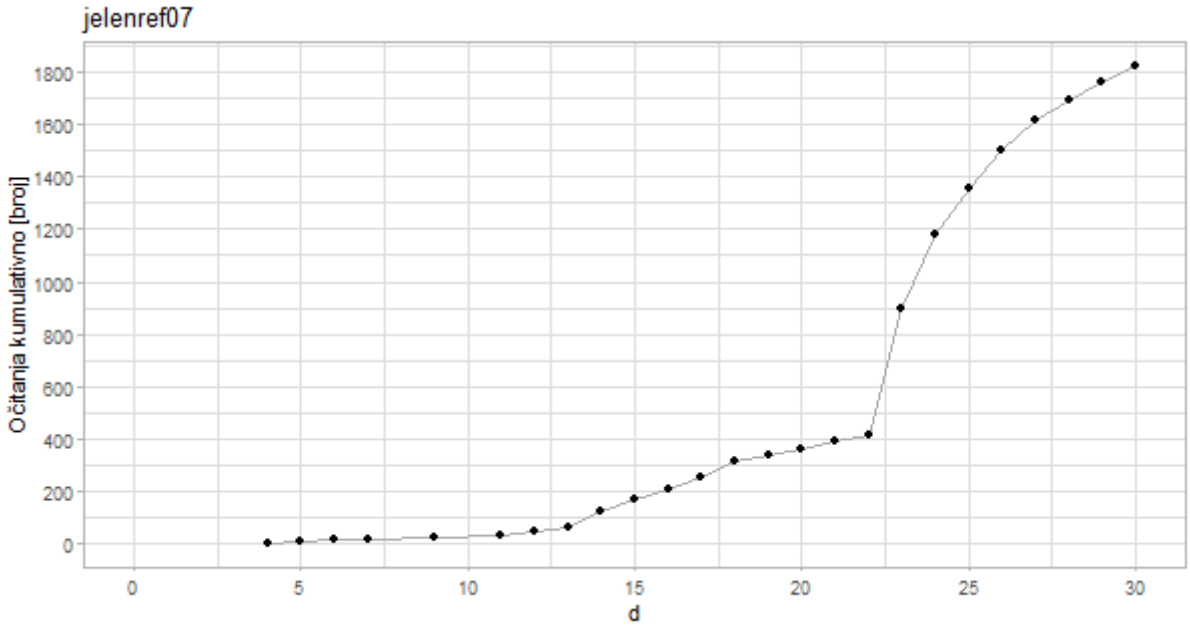
Rezultati opisane analize uzorka J29_B_CE_IonXpress_005 za alele jelenref05, jelenref06 i jelenref07 prikazani su redom na slikama (Slika 3.3, Slika 3.4, Slika 3.5).



Slika 3.3 Broj očitanja iz uzorka J29_B_CE_IonXpress_005 u ovisnosti o udaljenosti d od alela jelenref05



Slika 3.4 Broj očitanja iz uzorka J29_B_CE_IonXpress_005 u ovisnosti o udaljenosti d od alela jelenref06



Slika 3.5 Broj očitanja iz uzorka J29_B_CE_IonXpress_005 u ovisnosti o udaljenosti d od alela jelenref07

Kako je iz prethodne analize poznato da udaljenost između alela jelenref05 i jelenref06 iznosi 16, što je ujedno najmanja udaljenost između parova alela iz oba uzorka, to možemo zaključiti da za udaljenost $d = 7$ neće doći do slučaja u kojem će neko očitanje pripasti u dvije grupe istovremeno. Za $d = 7$ broj sekvenci bliskih alelu jelenref05 iznosi 1391, broj sekvenci bliskih alelu jelenref06 iznosi 215, dok broj sekvenci bliskih alelu jelenref07 iznosi tek 4.

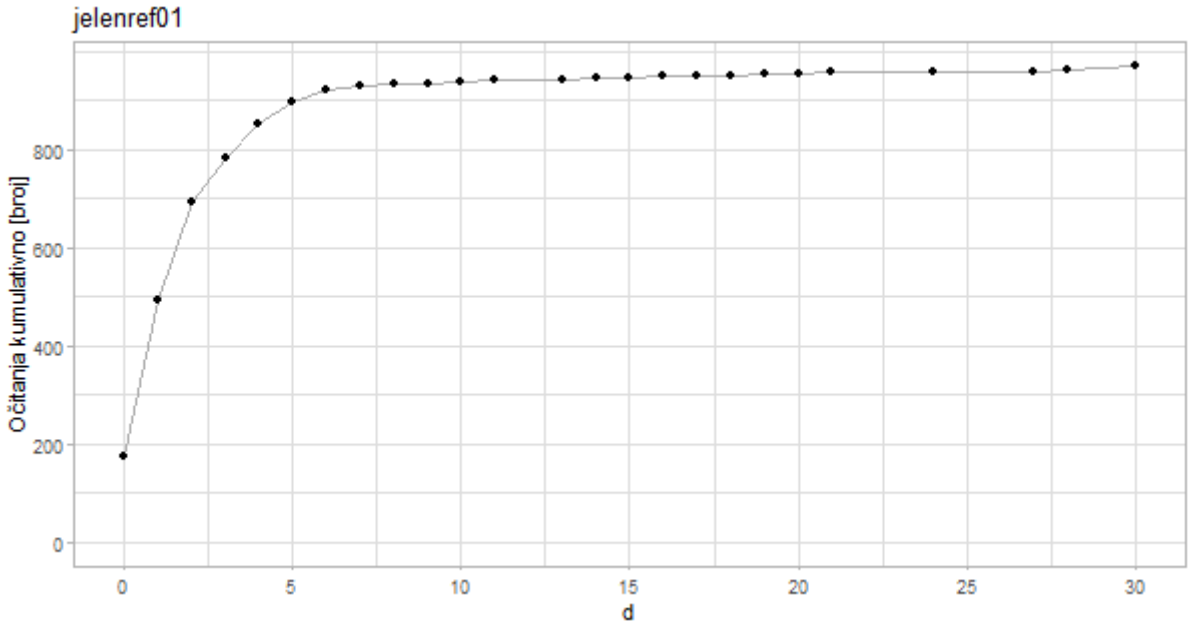
Alel jelenref05 najzastupljeniji je u uzorku i s obzirom na to da i za udaljenosti $d < 7$ očitanja bliskih ovom alelu ima mnogo, pretpostavlja se da će detekcija ovog alela biti jednostavna. Također, vrlo je velik broj očitanja upravo jednakih ovom alelu (čak 617).

Detekcija alela jelenref06 također bi trebala biti ostvarena, no kod ovog alela nailazi se na veći broj očitanja udaljenih za malu vrijednost parametra d ($d = 1, 2, 3, 4\ldots$) nego onih udaljenih za $d = 0$. Ovakva statistika mogla bi rezultirati pronalaskom alela jelenref06 s minimalnim pogreškama.

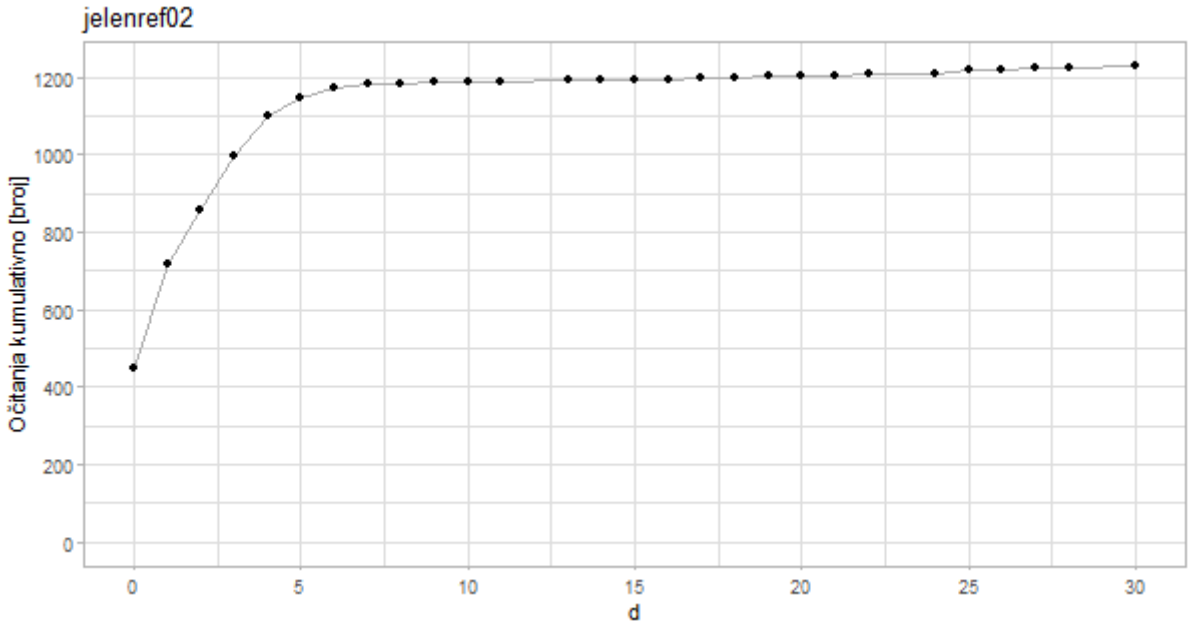
S druge strane, jelenref07 vrlo je slabo zastupljen u ovom uzorku. Od četiri sekvence za koje je udaljenost $d <= 7$, 2 su očitanja na udaljenosti $d = 1$, a 2 na udaljenosti $d =$

7. Uz ovakve rezultate, pretpostavlja se da je detekcija ovog alela algoritmima grupiranja upitna.

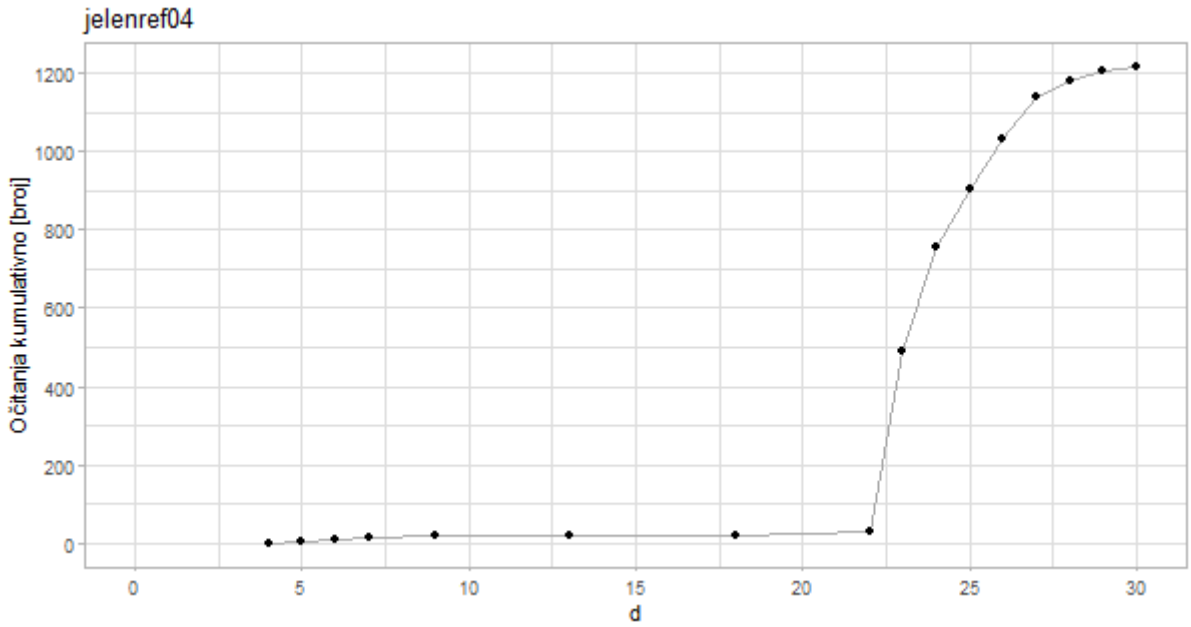
Rezultati analize uzorka J30_B_CE_IonXpress_006 prikazani su na slikama (Slika 3.6, Slika 3.7, Slika 3.8).



Slika 3.6 Broj očitanja iz uzorka J30_B_CE_IonXpress_006 u ovisnosti o udaljenosti d od alela jelenref01



Slika 3.7 Broj očitanja iz uzorka J30_B_CE_IonXpress_006 u ovisnosti o udaljenosti d od alela jelenref02



Slika 3.8 Broj očitanja iz uzorka J30_B_CE_IonXpress_006 u ovisnosti o udaljenosti d od alela jelenref04

Analiza uzorka J30_B_CE_IonXpress_006 pokazuje slične rezultate. Alel jelenref02 najzastupljeniji je u uzorku s čak 449 očitanja jednakih njemu samom. Za udaljenosti *d* <= 7 postoji 1181 očitanje blisko alelu jelenref02, 930 bliskih alelu jelenref01 te 14 očitanja bliskih alelu jelenref04. Ovaj alel je najmanje zastupljen u uzorku, međutim može se pretpostaviti da će ga biti jednostavnije pronaći nego alel jelenref07.

4. Metode grupiranja i rezultati

4.1. Prvi algoritam

4.1.1. Opis algoritma

Prvi algoritam implementiran u radu najjednostavniji je od svih metoda, a sekvence koje su rezultat višestrukog poravnanja uspoređuje međusobno i na taj se način određuje pripadnost pojedinoj grupi.

Algoritam je opisan koracima:

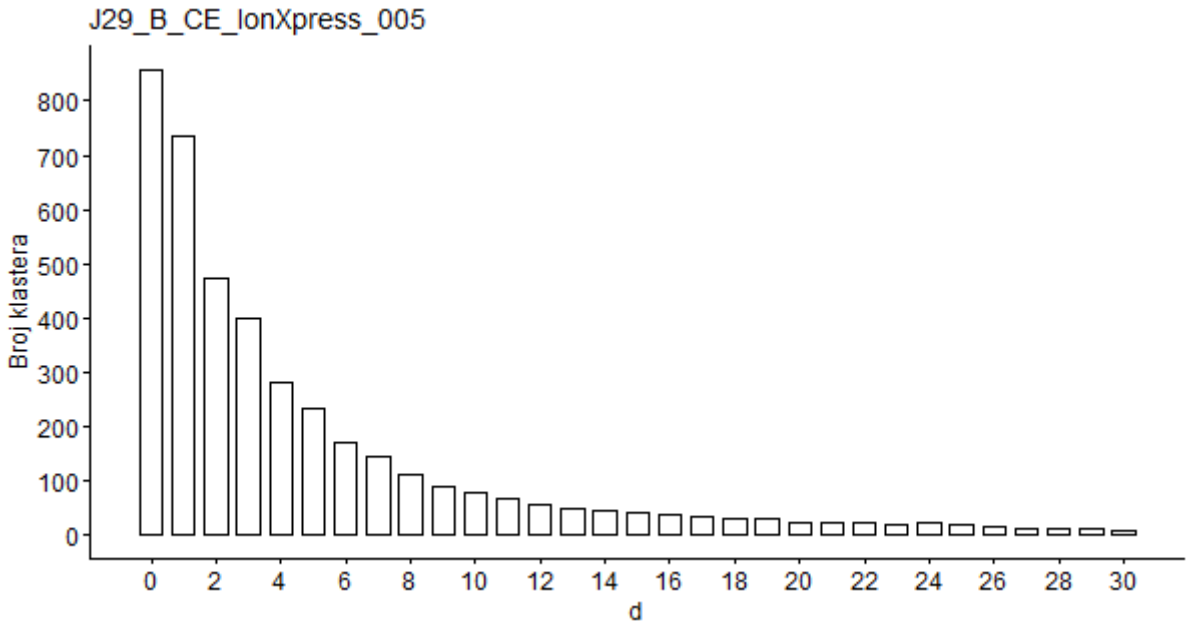
1. Filtriranje očitanja po duljini
2. Određivanje višestrukog poravnanja sekvenci
3. Odabir početnog konsenzusa iz višestrukog poravnanja
4. Za svaku sekvencu iz višestrukog poravnanja
    a. Usporedba sa svim sekvencama u grupama dok se ne pronađe grupa kojoj pripada
    b. Ako grupa nije pronađena, stvaranje nove grupe
5. Filtriranje grupa po veličini
6. Generiranje konsenzusnih sekvenci zadržanih grupa

U ovoj analizi sudjeluju sekvence najčešće duljine uz odstupanje iznosa +/- 5 baza. Stoga je potrebno sekvence početno filtrirati po duljini. Nakon toga, obavlja se izgradnja višestrukog poravnanja sekvenci pomoću alata Spoa[9] te se u daljnjoj usporedbi koriste poravnate sekvence/nizovi, točnije oni nizovi koje u višestrukom poravnanju predstavljaju originalne sekvence iz uzorka. Ovi nizovi, izuzev nukleotidnih baza, mogu dodatno sadržavati jedan ili više znakova „-“. Početno se kreira jedna grupa koja sadrži prvi niz iz liste poravnatih nizova, a zatim kreće usporedba svih ostalih. Računa se *Hammingova* udaljenost (broj pozicija na kojima se nizovi razlikuju, tj. broj pozicija na kojima nizovi imaju različite simbole (Pandžić, i dr., 2012.)) između niza koji predstavlja sekvencu u višestrukom poravnanju i svih nizova u grupi. Ako je niz dovoljno sličan (u ovisnosti o parametru *d*) svim nizovima u početnoj grupi, promatrani niz se dodaje u grupu, a originalna sekvenca koju ovaj

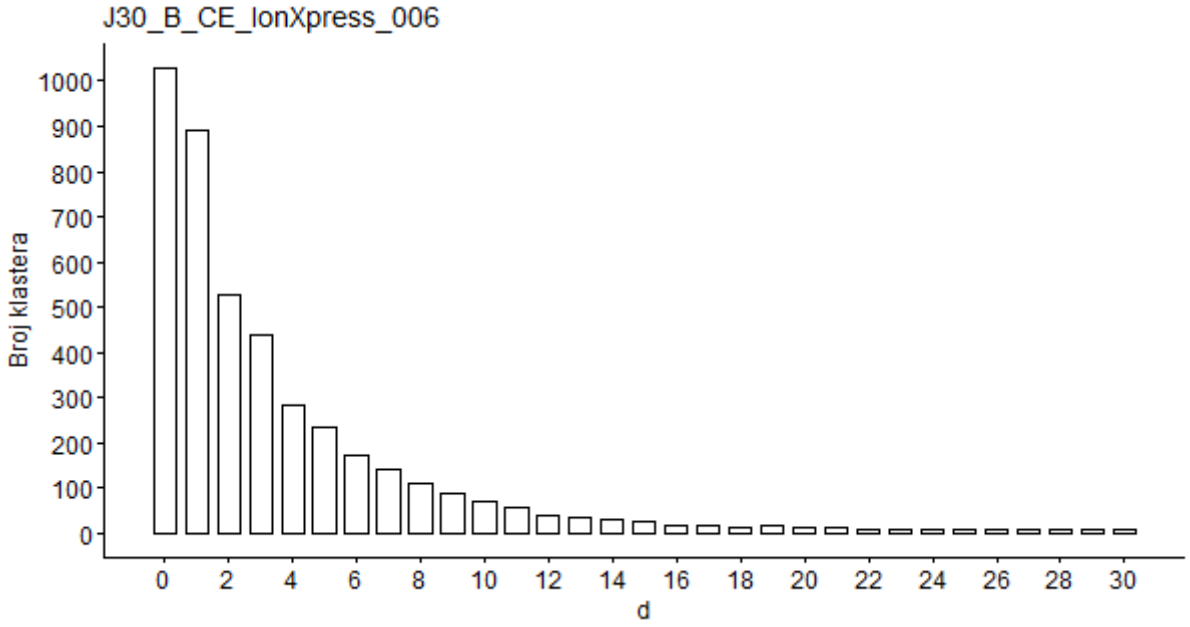
[9] https://github.com/rvaser/spoa

niz predstavlja također se pohranjuje radi kasnijeg generiranja konačne konsenzusne sekvence. Ako je niz udaljeniji od bilo kojeg niza u svakoj od do tada pronađenih grupa za vrijednost veću od parametra *d*, stvara se nova grupa čiji je jedini početni član promatrani niz. Postupak se ponavlja za sve sekvence u uzorku. Prilikom odabira parametra *d* u obzir je uzeta analiza prethodno u tablici (Tablica 3.1), koja zapravo daje vrijednosti *Hammingove* udaljenosti između pojedinih alela.

Također, napravljena je analiza ovisnosti broja formiranih grupa o parametru *d*. Rezultati su prikazani grafovima na slikama (Slika 4.1, Slika 4.2).



Slika 4.1 Graf ovisnosti broja generiranih klastera o parametru d za uzorak J29_B_CE_IonXpress_005



Slika 4.2 Graf ovisnosti broja generiranih klastera o parametru d za uzorak J30_B_CE_IonXpress_006

Iz analize *Hammingove* udaljenosti alela i ovisnosti broja klastera u parametru usporedbe, za parametar *d* odabrana je vrijednost 15.

4.1.2. Rezultati

Uz parametre: 1 za preklapanje, 0 za zamjenu, -1 za brisanje ili umetanje te globalno poravnanje, 15 za udaljenost i 18 za veličinu klastera, rezultat pokretanja prvog algoritma implementiranog u radu za uzorak J29_B_CE_IonXpress_005 prikazan je u nastavku.

```
First Algorithm Results:
Size of cluster: 1356
Allele (296)
GATCCTCTCTCTGCAGCACATTTCCTGCTGTATGCTAAGAGCGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGGGGTTCCTGGACAGATACTTCTATAACGGAGAAGAGTTCGTGCGCTTCG
ACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGGCCGGTGGCCGAGTACCTG
AACAGCCAGAAGGAGTACATGGAGCAGACGCGGGCCGAGGTGGACACGTACTGCAGACACAA
CTACGGCGGCGTTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA

Size of cluster: 207
Allele (298)
GATCCTCTCTCTGCAGCACATTTCCTGCTGTATACTACGAGCGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGGGGTTCCTGGACAGATACTTCTATAACGGAGAAGAGTACGTGCGCTTCG
ACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGGCCGTCCGCCAAGTACTGG
AACAGCCAGAAGGAGTACATGGAGCAGACGCGGGCCGAGGTGGACAGGTACTGCAGACACAA
CTACGGGGTTCTTGACAGTTTCGCTGGTGCAGCGGTCGAGGTGACGCGAA
```

Algoritam u uzorku J29_B_CE_IonXpress_005 pronalazi 41 grupu te očekivano pronalazi alele jelenref05 u potpunosti i jelenref06 uz potrebno jedno brisanje, ali uz zadane parametre ne pronalazi alel jelenref07.

Na analizi uzorka J30_B_CE_IonXpress_006 prvi algoritam detektira 25 grupa te pronalazi alel jelenref02 u potpunosti, a alel jelenref01 uz potrebnu jednu zamjenu. Alel jelenref04 sa zadanim parametrima kao za analizu prethodnog uzorka pronađen je sa sljedećim potrebnim izmjenama: 1 brisanje, 2 umetanja i 1 zamjenu. Ovaj klaster sadrži samo 18 sekvenci. Analizom dodatnih podataka, klaster bi se mogao povećati, međutim potrebno je uzeti u obzir to da je postavljanjem minimalne veličine na ovako malenu vrijednost moguće pronaći konsenzuse koji uopće ne predstavljaju alel. Rezultat pokretanja prvog algoritma za uzorak J30_B_CE_IonXpress_006 prikazan je u nastavku.

```
First Algorithm Results:
Size of cluster: 1172
Allele (297)
GGATCCTCTCTCTGCAGCACATTTCCTGGAGTATGCTAAGAGCGAGTGTCATTTCTCCAACG
GGACGCAGCGGGTGCGGTTCCTGGACAGATACTTCTATAACCGGGAAGAGTACGTGCGCTTC
GACAGCGACTGGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGTCCGCCAAGTACTG
GAACAGCCAGAAGGATTTCATGGAGCAGAAGCGGGCCGAGGTGGACACGGTGTGCAGACACA
ACTACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA

Size of cluster: 923
Allele (296)
GATCCTCTCTCTGCAGCACATTTCCTGGAGCATCTTAAGGCCGAGTGTCATTTCTTCAACGG
GACGGAGCGGATGCAGTTCCTGGCGAGATACTTCTATAACGGAGAAGAGTACGCGCGCTTCG
ACAGCGACGTGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGGACGCCAAGTACTGG
AACAGCCAGAAGGAGATCCTGGAGCAGCACGGGGCAGAGGTGGACAGGTACTGCAGACACAA
CTACGGGGTCGGTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA

Size of cluster: 18
Allele (295)
GATCCTCTCTCTGCAGCACATTTCCTGATGTATACTAAGAAAGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGGGGCTCCTGGACAGATACTTCTATAACGGAGAAGAGTTCGTGCGCTTCG
ACAGCGACTGGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGGACGCCGAGGCTGGA
ACAGACAGAAGGAGCTCCTGGAGCAGAGGCGGGCCGCGGTGGACACGTACTGCAGACACAAC
TACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA
```

Zaključno, ovaj algoritam sa zadanim parametrima daje bolje rezultate na uzorku J30_B_CE_IonXpress_006.

Odabir drugačije vrijednosti parametra *d* rezultirao je dakako drugačijim rezultatima. Primjerice, odabirom vrijednosti manje od 15, algoritam je rezultirao boljim vrijednostima na skupu podataka J29_B_CE_IonXpress_005, međutim lošijim vrijednostima na uzorku J30_B_CE_IonXpress_006. Ovakav rezultat može se objasniti primjedbom da su aleli iz uzorka J29_B_CE_IonXpress_005 međusobno sličniji nego oni u uzorku J30_B_CE_IonXpress_006. Tako u drugom spomenutom uzorku dolazi do detekcije klastera čije sekvence potencijalno predstavljaju alele uz određene pogreške te se u prevelikoj mjeri razlikuju od sekvenci brojčano vodećih grupa.

4.2. Drugi algoritam

4.2.1. Opis algoritma

Druga metoda u radu za usporedbu sekvenci koristi algoritam poravnanja. Za razliku od prvog algoritma, ovaj algoritam uspoređuje originalne sekvence iz uzorka, a konsenzus grupe određuje se korištenjem biblioteke Spoa. Koraci drugog algoritma:

1. Filtriranje očitanja po duljini
2. Postavljanje početnog očitanja iz uzorka za početnu grupu
3. Za svaku sekvencu iz uzorka:
    a. Usporedba s konsenzusima grupa dok se ne pronađe ona kojoj pripada (u ovisnosti o udaljenosti *d*)
    b. Pripajanje sekvence odgovarajućoj grupi i računanje konsenzusa nakon dodavanja sekvence
    c. Ako sekvenca ne pripada nijednoj grupi, stvaranje nove grupe čiji je konsenzus promatrana sekvenca
4. Filtriranje grupa u ovisnosti o parametru veličine *v*

Ovaj algoritam dodavanjem sekvence u grupu odmah određuje novu konsenzusnu sekvencu te svaku sekvencu za koju je potrebno odrediti pripadnost grupi uspoređuje s konsenzusom. Na kraju tako nije potrebno određivati reprezentativne sekvence pojedinih grupa, već su one generirane pri samoj analizi.

4.2.2. Rezultati

Uz parametre 1, 0 i -1 za preklapanje, zamjenu i brisanje ili umetanje respektivno, 7 za granicu razlike među sekvencama pri određivanju pripadnosti grupi te polu-globalno poravnanje, 2. algoritam na uzorku J29_B_CE_IonXpress_005 pronalazi 53 grupe. Alel jelenref05 je pronađen u potpunosti, jelenref06 uz potrebno jedno brisanje, no alel jelenref07 nije pronađen.

```
Second algorithm results:

Size of cluster: 1374
Allele (296)
GATCCTCTCTCTGCAGCACATTTCCTGCTGTATGCTAAGAGCGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGGGGTTCCTGGACAGATACTTCTATAACGGAGAAGAGTTCGTGCGCTTCG
ACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGGCCGGTGGCCGAGTACCTG
AACAGCCAGAAGGAGTACATGGAGCAGACGCGGGCCGAGGTGGACACGTACTGCAGACACAA
CTACGGCGGCGTTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA

Size of cluster: 38
Allele (298)
GATCCTCTCTCTGCAGCACATTTCCTGCTGTATACTACGAGCGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGGGGTTCCTGGACAGATACTTCTATAACGGAGAAGAGTACGTGCGCTTCG
ACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGGCCGTCCGCCAAGTACTGG
AACAGCCAGAAGGAGTACATGGAGCAGACGCGGGCCGAGGTGGACAGGTACTGCAGACACAA
CTACGGGGTTCTTGACAGTTTCGCTGGTGCAGCGGTCGAGGTGACGCGAA
```

Zanimljivo je napomenuti da je analizom odbačenih klastera primijećeno da je pronađen klaster s konsenzusnom sekvencom vrlo sličnom alelu jelenref07. Točnije, sekvenca se od alela jelenref07 razlikuje za 1 zamjenu, 1 umetanje i 1 brisanje.

```
Size of cluster: 3
Allele (296)
GATCCTCTCTCTGCAGCACATTTCCTGGAGCATCATAAGTGCGAGTGTCATTTCTCCAACGG
GACGGAGCGGGTGCAGTTCCTGCAGAGATACATCTATAACCGGGAAGAGTACGTGCGCTTCG
ACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGCCGTCTCGCCAAGTACTAT
AACAGCCAGAAGGAGCTCCTGGAGCAGAAGCGGGCCGCGGTGGACAGGTACTGCAGACACAA
CTACGGGGTCGTTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA
```

Nažalost, ova grupa sadrži samo 3 sekvence, a s obzirom na to da je pronađeno nekoliko većih, a još uvijek relativno malih grupa sličnijih alelima jelenref05 i jelenref06, ovu grupu bilo je potrebno odbaciti. No, može se pretpostaviti da bi ovaj algoritam dao prihvatljivije rezultate na većem skupu podataka za analizu. S druge strane, algoritam ima veliku vremensku složenost s obzirom na to da za svaku sekvencu iz testnog skupa računa poravnanje s barem jednim konsenzusom iz skupa

klastera, a nakon dodavanja svake sekvence u grupu određuje i novi konsenzus promatrane grupe.

Na uzorku J30_B_CE_IonXpress_006 metoda uz jednake parametre daje drugačije rezultate. Sekvence su grupirane u 34 grupe. Pronađene su dvije velike grupe (sadržeći 669 i 502 sekvence) koje se razlikuju u samo jednoj praznini, odnosno dodatnoj bazi na početku i određuju alel jelenref02 u potpunosti. Također, pronađena je grupa veličine 924 sekvence koja određuje alel jelenref01 uz potrebnu jednu zamjenu. Alel jelenref04 je pronađen uz potrebno 1 brisanje, 2 umetanja i 1 zamjenu, baš kao i kod prve metode.

```
Second algorithm results:

Size of cluster: 924
Allele (296)
GATCCTCTCTCTGCAGCACATTTCCTGGAGCATCTTAAGGCCGAGTGTCATTTCTTCAACGG
GACGGAGCGGATGCAGTTCCTGGCGAGATACTTCTATAACGGAGAAGAGTACGCGCGCTTCG
ACAGCGACGTGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGGACGCCAAGTACTGG
AACAGCCAGAAGGAGATCCTGGAGCAGCACGGGGCAGAGGTGGACAGGTACTGCAGACACAA
CTACGGGGTCGGTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA

Size of cluster: 669
Allele (296)
GATCCTCTCTCTGCAGCACATTTCCTGGAGTATGCTAAGAGCGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGCGGTTCCTGGACAGATACTTCTATAACCGGGAAGAGTACGTGCGCTTCG
ACAGCGACTGGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGTCCGCCAAGTACTGG
AACAGCCAGAAGGATTTCATGGAGCAGAAGCGGGCCGAGGTGGACACGGTGTGCAGACACAA
CTACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA

Size of cluster: 502
Allele (297)
GGATCCTCTCTCTGCAGCACATTTCCTGGAGTATGCTAAGAGCGAGTGTCATTTCTCCAACG
GGACGCAGCGGGTGCGGTTCCTGGACAGATACTTCTATAACCGGGAAGAGTACGTGCGCTTC
GACAGCGACTGGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGTCCGCCAAGTACTG
GAACAGCCAGAAGGATTTCATGGAGCAGAAGCGGGCCGAGGTGGACACGGTGTGCAGACACA
ACTACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA

Size of cluster: 16
Allele (295)
GATCCTCTCTCTGCAGCACATTTCCTGATGTATACTAAGAAAGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGGGGCTCCTGGACAGATACTTCTATAACGGAGAAGAGTTCGTGCGCTTCG
ACAGCGACTGGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGGACGCCGAGGCTGGA
ACAGACAGAAGGAGCTCCTGGAGCAGAGGCGGGCCGCGGTGGACACGTACTGCAGACACAAC
TACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA
```

Analizom odbačenih klastera u rezultatima prva dva algoritma, zaključak je da se često formiraju manje grupe koje sadrže sekvence koje su drugačije od onih koje pripadaju najvećim grupama (konačno neodbačenima), ali u konačnici rezultiraju konsenzusima koji su ili vrlo slični poznatim alelima, ili potpuno jednaki. Svakako je moguće da parametri odabrani za izgradnju klastera nisu idealni, međutim ovom bi se problemu moglo doskočiti na način da se u konačnici takve grupe spoje i promatraju kao jedan klaster. Kako opisane grupe u pravilu sadrže mali broj sekvenci, one vjerojatno neće previše utjecati na formiranje konsenzusa velike grupe, ali ovo pripajanje moglo bi biti korisno u vidu otkrivanja manje zastupljenih alela.

4.3. Treći algoritam

4.3.1. Opis algoritma

Treća metoda implementirana u radu razlikuje se od prve dvije jer počinje s pronalaskom sekvenci koje će označavati početne centroide klastera, a zatim se ostale sekvence pridružuju svakoj grupi od čijih se članova sekvenca razlikuje za manje od parametra *d*. Na posljetku se određuju konačni centroidi svake grupe te se vrlo slične grupe spajaju i računa se novi konsenzus spojenih grupa. Koraci treće metode:

1. Filtriranje očitanja po duljini
2. Određivanje višestrukog poravnanja sekvenci
3. Odabir prve sekvence iz liste višestruko poravnatih za početni konsenzus
4. Za svaku poravnatu sekvencu
    a. Usporedba sa svim centroidima računanjem *Hammingove* udaljenosti
    b. Ako je udaljena više od parametra *c* od svih centroida, postavlja se za novi centroid
5. Za svaku poravnatu sekvencu:
    a. Popunjavanje grupa na temelju udaljenosti *d* svake sekvence od svakog centroida
6. Detekcija sličnih konsenzusa u ovisnosti o parametru *g* korištenjem algoritma poravnanja

7. Spajanje grupa sličnih konsenzusa
8. Određivanje konsenzusa spojenih grupa
9. Filtriranje grupa u ovisnosti o parametru veličine *v*

U analizi se za generiranje konačnih konsenzusa koriste izvorne sekvence iz uzorka.

4.3.2. Rezultati

Uz parametre: 1 za preklapanje, 0 za zamjenu, -1 za umetanje ili brisanje te polu-globalno poravnanje, 16 za kreiranje novog klastera, 14 (< 15) za dodavanje u klaster i 15 (< 16) za granicu sličnosti centroida, rezultati za uzorak J29_B_CE_IonXpress_005 prikazani su u nastavku.

```
Third algorithm results:
Size of cluster: 1080
Allele (296)
GATCCTCTCTCTGCAGCACATTTCCTGCTGTATGCTAAGAGCGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGGGGTTCCTGGACAGATACTTCTATAACGGAGAAGAGTTCGTGCGCTTCG
ACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGGCCGGTGGCCGAGTACCTG
AACAGCCAGAAGGAGTACATGGAGCAGACGCGGGCCGAGGTGGACACGTACTGCAGACACAA
CTACGGCGGCGTTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA

Size of cluster: 231
Allele (299)
GATCCTCTCTCTGCAGCACATTTCCTGCTGTATACTACGAGCGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGGGGTTCCTGGACAGATACTTCTATAACGGAGAAGAGTACGTGCGCTTCG
ACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGGCCGTCCGCCAAGTACTGG
AACAGCCAGAAGGAGTACATGGAGCAGACGCGGGCCGAGGTGGACAGGTACTGCAGACACAA
CTACGGGGTTCTTGACAGTTTCGCTGGTGCAGCGGTCGAGGTGACGCGAAA

Size of cluster: 4
Allele (295)
GATCCTCTCTCTGCAGCACATTTCCTGGAGCATCATAAGTGCGAGTGTCATTTCTCCAACGG
GACGGAGCGGGTGCAGTTCCTGCAGAGATACATCTATAACCGGGAAGAGTACGTGCGCTTCG
ACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGCCGTCCGCCAAGTACTATA
ACAGCCAGAAGGAGCTCCTGGAGCAGAAGCGGGCCGCGGTGGACAGGTACTGCAGACACAAC
TACGGGGTCGTTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA

Size of cluster: 4
Allele (296)
GATCCTCTCTCTGCAGCACATTTCCTGGAGTATCATAAGAGCGAGTGTCATTTCTTCAACGG
GACCGAGCGGGTGCGGTTCCTGGACAGATACTTCCATAATGGAGAAGAGTTCGTGCGCTTCA
ACAGCGACTGGGGCGAGTACCGGGCGGTGGCCGAGCTGGGGCGGCCGGCCGCCGAGCACTGG
AACAGCCAGAAGGAGATTCTGGAGCAGAGGCGGGCCGAGGTGGACACGGTGTGCAGACACAA
CTACGGGGTCGTTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA
```

3. algoritam pokazuje uspjeh u pronalasku najmanje zastupljenog alela u uzorku J29_B_CE_IonXpress_005 (jelenref07) uz potrebno jedno umetanje. Neuspjeh leži u tome što je pronađena jedna grupa jednake veličine kao grupa najmanje zastupljenog alela (veličine 4 sekvence) čiji je konsenzus vrlo različit od svih poznatih alela. Može se pretpostaviti da je ovaj „alel“ rezultat pogreške u konkretnom uzorku i da postoji manja vjerojatnost da bi bio neodbačen na većem uzorku, međutim, u ovom slučaju ne može se zanemariti. U ostalim grupama pronađena je sekvenca koja u potpunosti odgovara alelu jelenref05 te sekvenca kojoj je potrebno jedno brisanje da bi odgovarala alelu jelenref06.

Na uzorku J30_B_CE_IonXpress_006 3. algoritam pronalazi alel jelenref02 u potpunosti, alel jelenref01 uz potrebnu jednu zamjenu te alel jelenref04 uz potrebno 1 brisanje, 2 umetanja i 1 zamjenu.

```
Third algorithm results:

Size of cluster: 752
Allele (296)
GATCCTCTCTCTGCAGCACATTTCCTGGAGTATGCTAAGAGCGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGCGGTTCCTGGACAGATACTTCTATAACCGGGAAGAGTACGTGCGCTTCG
ACAGCGACTGGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGTCCGCCAAGTACTGG
AACAGCCAGAAGGATTTCATGGAGCAGAAGCGGGCCGAGGTGGACACGGTGTGCAGACACAA
CTACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA


Size of cluster: 215
Allele (296)
GATCCTCTCTCTGCAGCACATTTCCTGGAGCATCTTAAGGCCGAGTGTCATTTCTTCAACGG
GACGGAGCGGATGCAGTTCCTGGCGAGATACTTCTATAACGGAGAAGAGTACGCGCGCTTCG
ACAGCGACGTGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGGACGCCAAGTACTGG
AACAGCCAGAAGGAGATCCTGGAGCAGCACGGGGCAGAGGTGGACAGGTACTGCAGACACAA
CTACGGGGTCGGTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA

Size of cluster: 18
Allele (295)
GATCCTCTCTCTGCAGCACATTTCCTGATGTATACTAAGAAAGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGGGGCTCCTGGACAGATACTTCTATAACGGAGAAGAGTTCGTGCGCTTCG
ACAGCGACTGGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGGACGCCGAGGCTGGA
ACAGACAGAAGGAGCTCCTGGAGCAGAGGCGGGCCGCGGTGGACACGTACTGCAGACACAAC
TACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA
```

4.4. Algoritam k-srednjih vrijednosti

4.4.1. Opis algoritma

Algoritam k-srednjih vrijednosti (engl. *k-means algorithm*) obavlja particijsko grupiranje u *K* čvrstih grupa. (Šnajder, 2019.) Broj *K* mora se unaprijed zadati. Ideja na kojoj se ovaj algoritam temelji je da svaka od grupa ima svoju srednju vrijednost (centroid), a svaki primjer pripada onoj grupi čiji mu je centroid najbliži. Pseudokod algoritma k-srednjih vrijednosti implementiranog u radu prikazan je u nastavku. Primjer iz domene sekvenci iz uzorka označen je s „**x**“, centroid iz skupa centroida označen je s „**c**“.

```
inicijaliziraj centroide
ponavljaj
      za svaki x
            pronađi c za koji je udaljenost(x, c) minimalna
            pridruži x grupi pronađenog c-a
      za svaku grupu
            konstruiraj nove centroide c
dok svi c ne konvergiraju
```

Algoritam k-srednjih vrijednosti pohlepno pretražuje te pronalazi lokalni optimum funkcije pogreške (Šnajder, 2019.), stoga njegova uspješnost uvelike ovisi o izboru početnih centroida. Primjerice, kada bi se za početne centroide odabrale dvije jednake sekvence, algoritam vjerojatno ne bi detektirao ispravne grupe, već bi se grupe ispreplitale ili bi jedna od grupa ostala prazna.

Implementacija algoritma k-srednjih vrijednosti ostvarena je kao biblioteka koja je uključena u program uz pomoć alata CMake. Definirane su dvije metode unutar biblioteke, od kojih jedna prima željene početne centroide, a druga u svojoj definiciji samostalno izračunava iste.

4.4.2. Rezultati

Koristeći algoritam *k-means* dobiveni su različiti rezultati ovisno o tome koja od definiranih funkcija je korištena.

4.4.2.1 Loš slučaj odabira početnih centorida

Početno je napravljena analiza nad sekvencama najčešće duljine +/- 5 nukleotidnih baza. Parametri poravnanja kao i izgradnje POA grafa su: 1 za podudaranje, -1 za zamjenu te -1 za umetanje ili brisanje. Zadani broj klastera iznosi 3. Početni centroidi pojedinog klastera birani su na temelju udaljenosti od konsenzusne sekvence svih sekvenci (dobivene korištenjem Spoa grafa). Naime, one sekvence koje su najrazličitije od konsenzusa, točnije čiji je rezultat poravnanja između njih samih i konsenzusne sekvence najlošiji u odnosu na sve ostale, odabrani su za početne centroide. Ovom metodom provjerilo se ponašanje algoritma u lošem slučaju, odnosno u slučaju kada su početni centroidi potencijalno rezultat pogreške u sekvenciranju ili mutacije. Najzastupljeniji alel jelenref05 u uzorku J29_B_CE_IonXpress_005 pronađen je u potpunosti, alel jelenref06 uz potrebno jedno brisanje, no zadnjem klasteru je pripala samo jedna sekvenca – centroid koja je vrlo različita od svih poznatih alela te alel jelenref07 nije pronađen. Slični su rezultati za uzorak J_30_B_CE_IonXpress_006. Alel jelenref02 pronađen je u potpunosti, alel jelenref01 uz potrebnu jednu zamjenu, a treći klaster sadrži samo jednu sekvencu-centroid koji ne odgovara nijednom od poznatih alela.

Sumarno, algoritam k-srednjih vrijednosti čak i uz početne sekvence odabrane lošim postupkom uspijeva pronaći zastupljenije alele u uzorku.

4.4.2.2 Smisleno određivanje početnih centroida

Drugi način postavljanja početnih centroida slijedi metodu opisanu u trećem algoritmu. Naime, prva sekvenca iz uzorka postavlja se za početni centroid prvog klastera, a nakon toga se svaka sekvenca uspoređuje sa svim centroidima (početno samo prvim centroidom) na način opisan u trećem poglavlju rada (3.2) te ako se razlikuje od svih centroida za barem parametar *c*, kreira se novi klaster čiji je centroid jednak promatranoj sekvenci. Parametri korišteni u provođenju grupiranja su 1 za podudaranje, 0 za zamjenu, -1 za umetanje ili brisanje, polu-globalno poravnanje, 30 za parametar razlike stvaranja novog klastera te 6 za graničnu veličinu klastera. U analizi su promatrane sekvence najčešće duljine +/-5 nukleotidnih baza. Algoritam ovim načinom odabira početnih centroida pokazuje

značajan uspjeh u otkrivanju slabo zastupljenih alela. U uzorku J29_B_CE_IonXpress_005 pronađena su dva klastera, jedan veličine 1406, a drugi 6 sekvenci, koji u potpunosti određuju alel jelenref05.

Alel jelenref06 pronađen je uz potrebno jedno brisanje, a klaster koji pripada ovom alelu sadrži 234 sekvence. Također, pronađen je klaster veličine 6 sekvenci koji označava alel jelenref07 uz potrebno jedno umetanje.

```
K-means algorithm results with fair centroids initialization:
File J29_B_CE_IonXpress_005

Size of cluster: 234
Allele (298)
GATCCTCTCTCTGCAGCACATTTCCTGCTGTATACTACGAGCGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGGGGTTCCTGGACAGATACTTCTATAACGGAGAAGAGTACGTGCGCTTCG
ACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGGCCGTCCGCCAAGTACTGG
AACAGCCAGAAGGAGTACATGGAGCAGACGCGGGCCGAGGTGGACAGGTACTGCAGACACAA
CTACGGGGTTCTTGACAGTTTCGCTGGTGCAGCGGTCGAGGTGACGCGAA

Size of cluster: 1406
Allele (296)
GATCCTCTCTCTGCAGCACATTTCCTGCTGTATGCTAAGAGCGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGGGGTTCCTGGACAGATACTTCTATAACGGAGAAGAGTTCGTGCGCTTCG
ACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGGCCGGTGGCCGAGTACCTG
AACAGCCAGAAGGAGTACATGGAGCAGACGCGGGCCGAGGTGGACACGTACTGCAGACACAA
CTACGGCGGCGTTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA

Size of cluster: 6
Allele (295)
GATCCTCTCTCTGCAGCACATTTCCTGGAGCATCATAAGTGCGAGTGTCATTTCTCCAACGG
GACGGAGCGGGTGCAGTTCCTGCAGAGATACATCTATAACCGGGAAGAGTACGTGCGCTTCG
ACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGCCGTCCGCCAAGTACTATA
ACAGCCAGAAGGAGCTCCTGGAGCAGAAGCGGGCCGCGGTGGACAGGTACTGCAGACACAAC
TACGGGGTCGTTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA

Size of cluster: 6
Allele (295)
GATCCTCTCTCTGCAGCACATTTCCTGCTGTATGCTAAGAGCGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGGGGTTCCTGGACAGATACTTCTATAACGGAGAAGAGTTCGTGCGCTTCG
ACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGGCCGGTGGCCGAGTACCTG
AACAGCCAGAAGGAGTACATGGAGCAGACGCGGGCCGAGGTGGACACGTACTGCAGACACAA
CTACGGCGGCGTTGAGAGTTTCACTGTGCAGCGGCGAGGTTGACGAC
```

U uzorku J_30_B_CE_IonXpress_006 algoritam uz jednake parametre uspijeva detektirati alel jelenref02 u potpunosti, alel jelenref01 uz potrebnu jednu zamjenu te alel jelenref04 uz potrebno 1 brisanje, 2 umetanja i 1 zamjenu, kao i većina ostalih algoritama. Rezultati su prikazani u nastavku.

```
K-means algorithm results with fair centroids initialization:
File J30_B_CE_IonXpress_006

Size of cluster: 951
Allele (298)
GATCCTCTCTCTGCAGCACATTTCCTGGAGCATCTTAAGGCCGAGTGTCATTTCTTCAACGG
GACGGAGCGGATGCAGTTCCTGGCGAGATACTTCTATAACGGAGAAGAGTACGCGCGCTTCG
ACAGCGACGTGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGGACGCCAAGTACTGG
AACAGCCAGAAGGAGATCCTGGAGCAGCACGGGGCAGAGGTGGACAGGTACTGCAGACACAA
CTACGGGGTCGGTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAAAA

Size of cluster: 1196
Allele (298)
GGATCCTCTCTCTGCAGCACATTTCCTGGAGTATGCTAAGAGCGAGTGTCATTTCTCCAACG
GGACGCAGCGGGTGCGGTTCCTGGACAGATACTTCTATAACCGGGAAGAGTACGTGCGCTTC
GACAGCGACTGGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGTCCGCCAAGTACTG
GAACAGCCAGAAGGATTTCATGGAGCAGAAGCGGGCCGAGGTGGACACGGTGTGCAGACACA
ACTACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAAA

Size of cluster: 19
Allele (296)
GATCCTCTCTCTGCAGCACATTTCCTGATGTATACTAAGAAAGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGGGGCTCCTGGACAGATACTTCTATAACGGAGAAGAGTTCGTGCGCTTCG
ACAGCGACTGGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGGACGCCGAGGCTGGA
ACAGACAGAAGGAGCTCCTGGAGCAGAGGCGGGCCGCGGTGGACACGTACTGCAGACACAAC
TACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAAA
```

4.4.2.3 Postavljanje željenih sekvenci za početne centroide

Na kraju su metodi algoritma k-srednjih vrijednosti za početne centroide predane sekvence dobivene kao rezultat 3. algoritma. Korišteni parametri jednaki su onima navedenima u prethodnom potpoglavlju. U samo dvije iteracije, algoritam uspijeva ispraviti sekvencu koja predstavlja alel jelenref07 te ga u potpunosti pronaći. Ostale sekvence dobivene na ovaj način ne razlikuju se od rezultata 3. algoritma. Rezultati uz parametre: 1 za podudaranje, 0 za zamjenu te -1 za umetanje ili brisanje te polu-globalno poravnanje za uzorak J29_B_CE_IonXpress_005:

```
K-means algorithm results:
File J29_B_CE_IonXpress_005

Size of cluster: 1418
Allele (297)
GATCCTCTCTCTGCAGCACATTTCCTGCTGTATGCTAAGAGCGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGGGGTTCCTGGACAGATACTTCTATAACGGAGAAGAGTTCGTGCGCTTCG
ACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGGCCGGTGGCCGAGTACCTG
AACAGCCAGAAGGAGTACATGGAGCAGACGCGGGCCGAGGTGGACACGTACTGCAGACACAA
CTACGGCGGCGTTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAAA

Size of cluster: 232
Allele (298)
GATCCTCTCTCTGCAGCACATTTCCTGCTGTATACTACGAGCGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGGGGTTCCTGGACAGATACTTCTATAACGGAGAAGAGTACGTGCGCTTCG
ACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGGCCGTCCGCCAAGTACTGG
AACAGCCAGAAGGAGTACATGGAGCAGACGCGGGCCGAGGTGGACAGGTACTGCAGACACAA
CTACGGGGTTCTTGACAGTTTCGCTGGTGCAGCGGTCGAGGTGACGCGAA

Size of cluster: 7
Allele (296)
GATCCTCTCTCTGCAGCACATTTCCTGGAGCATCATAAGTGCGAGTGTCATTTCTCCAACGG
GACGGAGCGGGTGCAGTTCCTGCAGAGATACATCTATAACCGGGAAGAGTACGTGCGCTTCG
ACAGCGACTGGGGCGAGTACCGGGCGGTGACAGAGCTGGGGCGGCCGTCCGCCAAGTACTAT
AACAGCCAGAAGGAGCTCCTGGAGCAGAAGCGGGCCGCGGTGGACAGGTACTGCAGACACAA
CTACGGGGTCGTTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAA
```

Rezultati uz jednake parametre za uzorak J30_B_CE_IonXpress_006:

```
K-means algorithm results:
File J30_B_CE_IonXpress_006

Size of cluster: 1199
Allele (298)
GGATCCTCTCTCTGCAGCACATTTCCTGGAGTATGCTAAGAGCGAGTGTCATTTCTCCAACG
GGACGCAGCGGGTGCGGTTCCTGGACAGATACTTCTATAACCGGGAAGAGTACGTGCGCTTC
GACAGCGACTGGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGTCCGCCAAGTACTG
GAACAGCCAGAAGGATTTCATGGAGCAGAAGCGGGCCGAGGTGGACACGGTGTGCAGACACA
ACTACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAAA

Size of cluster: 949
Allele (298)
GATCCTCTCTCTGCAGCACATTTCCTGGAGCATCTTAAGGCCGAGTGTCATTTCTTCAACGG
GACGGAGCGGATGCAGTTCCTGGCGAGATACTTCTATAACGGAGAAGAGTACGCGCGCTTCG
ACAGCGACGTGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGGACGCCAAGTACTGG
AACAGCCAGAAGGAGATCCTGGAGCAGCACGGGGCAGAGGTGGACAGGTACTGCAGACACAA
CTACGGGGTCGGTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAAAA

Size of cluster: 19
Allele (296)
GATCCTCTCTCTGCAGCACATTTCCTGATGTATACTAAGAAAGAGTGTCATTTCTCCAACGG
GACGCAGCGGGTGGGGCTCCTGGACAGATACTTCTATAACGGAGAAGAGTTCGTGCGCTTCG
ACAGCGACTGGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGGACGCCGAGGCTGGA
ACAGACAGAAGGAGCTCCTGGAGCAGAGGCGGGCCGCGGTGGACACGTACTGCAGACACAAC
TACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAAA
```

4.5. Analiza na kraćim očitanjima

Intuitivno je pretpostaviti da će kraća očitanja uz korištenje polu-globalnog poravnanja „pomoći“ u određivanju traženih alela. Kako je poznato iz analize duljina sekvenci (Slika 3.1, Slika 3.2), u uzorku J29_B_CE_IonXpress_005 postoje mnoga očitanja kraća od najčešće duljine. Za analizu je upotrijebljen 3. algoritam te su promatrana očitanja duljine [200, 296+5] nukleotidnih baza, gdje gornja granica označava najčešću duljinu uz odstupanje iznosa 5 baza. Sekvence dulje od najčešće duljine uz spomenuto odstupanje nisu uzete u obzir jer se pretpostavlja da su one gotovo sigurno himerna očitanja.

Na uzorku J29_B_CE_IonXpress_005 pronađen je alel jelenref05 u potpunosti te se brojnost klastera koji pripada ovom alelu povećala na čak 2217 očitanja. S druge strane, pronalazak alela jelenref06 pokazao se manje uspješnim te su sekvenci koja predstavlja ovaj alel potrebna 2 brisanja, 3 umetanja i 8 zamjena. Sekvenca koja je

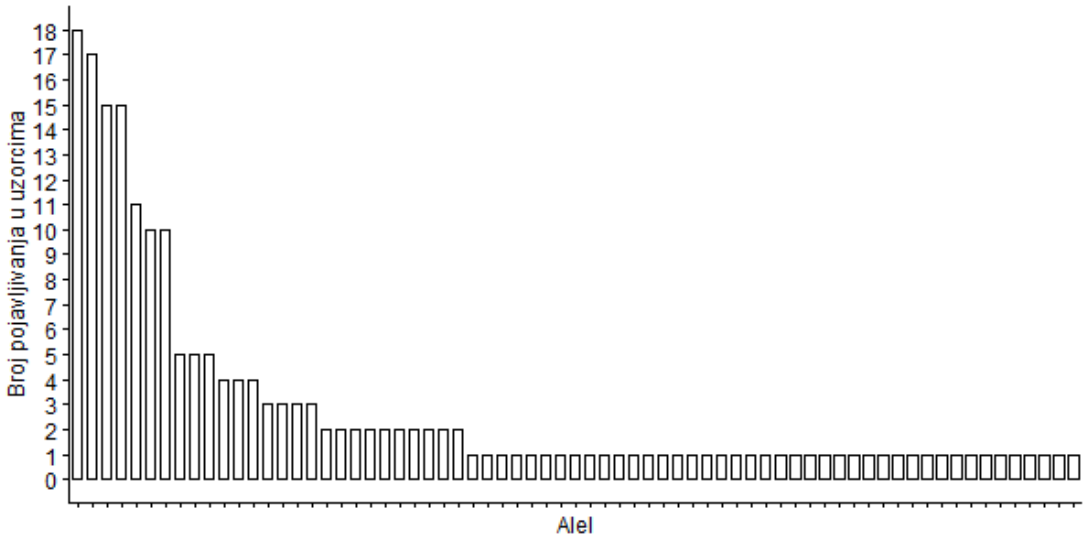
najsličnija alelu jelenref07 zahtjeva 3 umetanja i 3 zamjene da bi u potpunosti odgovarala ovom alelu. Nadalje, pronađen je klaster koji sadrži 720 sekvenci, a ne pripada nijednom alelu, odnosno u velikoj se mjeri razlikuje od svih poznatih.

Metoda se pokazala uspješnijom pri analizi uzorka J30_B_CE_IonXpress_006 u odnosu na prethodni uzorak, međutim daje rezultate u potpunosti jednake analizi provedenoj na očitanjima najčešće duljine +/- 5, uz drugačije veličine pojedinih klastera.

4.6. Analiza svih dostupnih uzoraka

U završnoj analizi korišteni su svi dostupni uzorci gena jelena običnog.

Broj uzoraka u kojima je pojedini alel detektiran prikazan je histogramom na slici (Slika 4.3).



Slika 4.3 Histogram broja pojavljivanja alela u uzorcima

Alel koji je najčešće detektiran je alel jelenref02, pronađen u čak 18 uzoraka, kao i u 3 uzorka koji se razlikuju od sekvence pronađene u spomenutih 18 uzoraka u jednoj dodatnoj bazi završnice.

U 17 uzoraka pronađen je alel jelenref01. Isti alel uz potrebne 2 zamjene pronađen je u 4 uzorka, uz potrebne 4 zamjene u 3 uzorka te uz potrebne 2 zamjene u 3 uzorka.

Jedan nepozanti alel kao i alel jelenref05 pojavljuju se u 15 uzoraka.

Alel jelenref06 uz potrebno jedno brisanje detektiran je u 11 uzoraka te u 4 uzorka. Sekvence koje predstavljaju ovaj alel razlikuju se u jednom umetanju/brisanju u dijelu sekvence koji predstavlja završnicu.

Još dva nepoznata alela pronađena su u 10 uzoraka.

U 5 uzoraka pronađen je jelenref07 u potpunosti te alel jelenref04 uz potrebna 2 umetanja, 1 brisanje i 1 zamjenu.

U manjem broju uzoraka također su pronađene varijacije pojedinih poznatih i nepoznatih alela te se postavlja pitanje koliko je zapravo specifičnih alela, a koliko je samo posljedica minornih modifikacija istovrsnog alela.

5. Rasprava

Iz rezultata prikazanih u prethodnom poglavlju moguće je primijetiti kako različite metode daju različite rezultate u vidu pronađenih alela i veličine pripadnih klastera. Najuspješnijom se pokazala 3. metoda u kombinaciji sa „ispravljanjem“ alela predavanjem konačnih rješenja algoritmu k-srednjih vrijednosti za početne centroide klastera, iz razloga što je na taj način u potpunosti pronađen najmanje zastupljen alel jelenref07 iz uzorka J29_B_CE_IonXpress_005.

Analiza koristeći kraća očitanja za uzorak J29_B_CE_IonXpress_005 nije se pokazala pretežito uspješnom, no moguće je da u uzorku postoji prevelik broj pogrešnih očitanja za uspješnu detekciju alela. S druge strane, u uzorku J30_B_CE_IonXpress_006 analiza nije rezultirala boljim rješenjem od analize koristeći samo očitanja najčešće duljine uz dozvoljenu pogrešku, ali brojnost sekvenci u pojedinim klasterima se povećala, stoga se može pretpostaviti da bi ovakva analiza u nekim slučajevima pridonijela detekciji slabo zastupljenih alela.

Skupna analiza uzoraka daje zanimljive rezultate. Svakako bi bilo korisno dodatno analizirati alele pronađene u manjem broju uzoraka i promotriti u kojoj se mjeri razlikuju od ostalih pronađenih sekvenci.

Nadalje, skupnu analizu bilo bi zanimljivo provesti na način da se svi uzorci spoje te promatraju kao jedan skup neoznačenih podataka. Time bi se potencijalno izbjeglo pojavljivanje vrlo sličnih alela, posebice korištenjem 3. algoritma koji u svojoj implementaciji ovakve konsenzusne sekvence detektira te pripadne klastere spaja u jedan.

Zaključak

Nakon razvoja algoritama, vrlo važan korak predstavilo je određivanje parametara poravnanja, udaljenosti, veličine prihvatljivih klastera itd. Događalo se da pojedini parametri dobro funkcioniraju na jednom skupu podataka, dok na drugom ne daju prihvatljive rezultate.

Čak i uz temeljitu analizu podataka prije provođenja nenadziranog strojnog učenja, pri samom grupiranju često se događaju neočekivane pojave i put do konačnog rješenja zahtjeva malo kreativnosti i mnogo pokušaja i pogrešaka. Stoga prostora za dodatne analize i daljnji rad ima u izobilju.

Pronalaženje varijanti gena vrlo je zahtjevan, ali koristan posao. Rješenja opisana u radu postigla su donekle prihvatljive rezultate, no trebala bi se isprobati na dodatnim skupovima podataka s poznatim očekivanim rezultatima da bi se odredila ispravnost i općenitost postupaka.

Sljedeći korak u radu na ovom području mogao bi biti razvoj i implementacija dodatnih metoda grupiranja. Jedna od informacija koju bi te metode idealno proizvele je vjerojatnost pripadnosti određene sekvence svakoj od pronađenih grupa. Ta bi vjerojatnost bila korisna pri grupiranju jer nije nemoguće da neke sekvence imaju predispoziciju za smještanje u više grupa s različitom vjerojatnošću. Ovakvo pitanje bi odgovor moglo pronaći u implementaciji modela Gaussove mješavine[10] i algoritma maksimizacije očekivanja[11]. Također, moguće je da sekvenca pripadne pojedinoj grupi samo zato što je od nje najmanje udaljena (iako je ta udaljenost i dalje neprihvatljiva). Na taj način grupama pripadaju sekvence koje u konačnici „kvare“ konsenzus. U nekim pak algoritmima ovakve sekvence formiraju nove grupe kojima se kasnije pridruže slične sekvence i zbog svoje brojnosti otežavaju distinkciju nezastupljenih alela od pogrešaka. Stoga stršeće vrijednosti ne bi bilo na odmet u potpunosti ukloniti iz analize, pa bi bilo korisno iskoristiti

[10] https://en.wikipedia.org/wiki/Mixture_model#Gaussian_mixture_model
[11] https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

algoritam koji detektira i ne uzima u obzir sekvence koje su vrlo različite od svih ostalih.

Literatura

Britannica, E. o. (n.d.). *Major histocompatibility complex*. Preuzeto 5. Lipanj 2020. iz https://www.britannica.com/science/major-histocompatibility-complex

Gleichmann, N. (9. Ožujak 2020.). *Neuroscience News & Research, Gene vs Allele: Definition, Difference and Comparison*. (Technology networks) Preuzeto 5. Lipanj 2020. iz https://www.technologynetworks.com/neuroscience/articles/gene-vs-allele-definition-difference-and-comparison-331835

Hosseini, M., Pratas, D., & Pinho, A. J. (Listopad 2016). A Survey on Data Compression Methods for Biological Sequences. *Information, 56*(10.3390).

Lee, C. (22. svibanj 2003). Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics, 19*(8), str. 999-1008.

Leksikografski zavod Miroslav Krleža. (2020.). *Hrvatska enciklopedija.* Preuzeto 12. Lipanj 2020. iz https://www.enciklopedija.hr/natuknica.aspx?ID=52496

Lipman, D. J., Altschul, S. F., & Kececioglu, J. D. (Lipanj 1989.). A tool for multiple sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, str. 4412-4415. Preuzeto 5. Lipanj 2020. iz https://www.ebi.ac.uk/Tools/msa/

Nacionalno Vijeće za Istraživanje (S.A.D.), N. (1988.). *Mapping and Sequencing the Human Genome.* National Academies Press.

NCBI GenBank, Chimera Detection in 16S rRNA Sequences at NCBI. (18. Travanj 2018.). Preuzeto 9. Lipanj 2020. iz https://www.ncbi.nlm.nih.gov/genbank/rrnachimera/

OpenStax. (2016.). *Biology.* OpenStax CNX. Preuzeto 5. Lipanj 2020. iz https://cnx.org/contents/GFy_h8cu@10.8:4qg08nt-@7/Characteristics-and-Traits

Paljak, I., & Hadviger, A. (11. Studeni 2015.). *Uvod u dinamičko programiranje.* Preuzeto 9. Lipanj 2020. iz https://www.fer.unizg.hr/_download/repository/prezentacija[7].pdf

Pandžić, I. S., Bažant, A., Ilić, Ž., Vrdoljak, Z., Kos, M., & Sinković, V. (2012.). *Uvod u teoriju informacije i kodiranje* (2. izd.). Zagreb: Element.

Rogers, K. (n.d.). *Britannica, What's the Difference Between a Gene and an Allele?* Preuzeto 5. Lipanj 2020. iz https://www.britannica.com/story/whats-the-difference-between-a-gene-and-an-allele

Šikić, M., & Domazet-Lošo, M. (Prosinac 2013.). *Bioinformatika - skripta.* Preuzeto 5. Lipanj 2020. iz https://www.fer.unizg.hr/_download/repository/bioinformatika_skripta_v1.2.pdf

Šnajder, J. (1. Listopad 2019.). *Nastavni materijali iz Strojnog učenja, Grupiranje.* Preuzeto 5. Lipanj 2020. iz https://www.fer.unizg.hr/_download/repository/SU-2019-19-Grupiranje[1].pdf

Vinayak, S., Alam, M. T., Mixson-Hayden, T., McCollum, A. M., Sem, R., Shah, N. K., . . . Wongsricha, C. e. (26. Ožujak 2010.). Origin and Evolution of Sulfadoxine Resistant Plasmodium falciparum. (Sibley, & L. David, Ur.) *PLoS Pathog, 6*(3). Dohvaćeno iz https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1000830

Wikipedia, Chimeric gene. (28. Prosinac 2018.). Preuzeto 5. Lipanj 2020. iz https://en.wikipedia.org/wiki/Chimeric_gene

Wikipedia, Ion semiconductor sequencing. (28. Siječanj 2020.). Preuzeto 12. Lipanj 2020. iz https://en.wikipedia.org/wiki/Ion_semiconductor_sequencing

Wikipedia, Multiple sequence alignment. (24. Svibanj 2020.). Preuzeto 5. Lipanj 2020. iz https://en.wikipedia.org/wiki/Multiple_sequence_alignment

Sažetak

Pronalaženje varijanti gena grupiranjem pomoću algoritama dinamičkog programiranja

Aleli su inačice gena koje se kod većine višestaničnih organizama nalaze na istom mjestu u homolognim parovima kromosoma. Par alela čini nasljednu osobinu. Poznavanje inačica pojedinog gena vrlo je korisno, no alele nije uvijek jednostavno prepoznati u uzorku. Pojedini aleli mogu biti slabo zastupljeni u populaciji, pa tako i u uzorku za istraživanje. Cilj ovog rada je razviti metode i algoritme grupiranja za prepoznavanje različitih alela u uzorku sekvenciranih gena MHC jelena običnog. Implementirana su četiri algoritma u programskom jeziku C++ koristeći alate Spoa i bioparser te algoritme dinamičkog programiranja međusobnog poravnanja sekvenci. Uzorci gena promatrani u radu dobiveni su sekvenciranjem metodom Ion Torrent i pohranjeni u datotekama FASTQ formata. U radu su opisani spomenuti algoritmi i rezultati dobiveni koristeći raspoložive podatke. Metode uspješno detektiraju vrlo zastupljene alele u uzorcima te ilustriraju probleme koji nastaju pri određivanju slabo zastupljenih alela kao i prisutnosti nevažećih podataka u uzorku.

Ključne riječi: alel, gen, grupiranje, poravnanje, dinamičko programiranje, C++, Spoa

Abstract

Detecting gene variants using clustering and dynamic programming algorithms

Alleles are gene variants found at the same place in homologous chromosome pairs in most multicellular organisms. A pair of alleles constitutes a hereditary trait. Knowing the variants of a particular gene is very useful, but alleles are not always easy to recognize in a sample. Particular alleles may be poorly represented in the population, thus in the research sample. The aim of this paper is to develop clustering methods and algorithms for detecting different alleles in a sample of sequenced MHC deer genes. Four algorithms were implemented using C++ programming language, spoa and bioparser tools, as well as dynamic programming algorithms for pairwise sequence alignment. The gene samples observed in the paper were obtained using Ion Torrent sequencing method and stored in FASTQ format files. The paper describes the mentioned algorithms and the results obtained using available data. The methods successfully detect highly represented alleles in the sample and illustrate the problems which arise when determining poorly represented alleles as well as the presence of invalid data in the sample.

Key words: allele, gene, clustering, alignment, dynamic programming, C++, Spoa