UNIVERZA V LJUBLJANI
BIOTEHNIŠKA FAKULTETA

Jana Obšteter

UPORABA GENETSKEGA ALGORITMA ZA OPTIMIZACIJO REFERENČNE POPULACIJE V GENOMSKI SELEKCIJI

Seminarska naloga pri predmetu RAČUNSKA BIOLOGIJA
Doktorski študij Bioznanosti

Ljubljana, 2018

Kazalo vsebine

1 PREDSTAVITEV RAZISKOVALNEGA PROBLEMA IN CILJEV

1.1 SELEKCIJA PRI GOVEDU

Selekcija stremi h genetskemu izboljšanju populacije. V živinoreji izvajamo umetno selekcijo na gospodarsko pomembne lastnosti. V govedoreji so to predvsem proizvodne lastnosti, npr. prireja mleka ali prirast mesa, lastnosti zunanjosti živali in lastnosti zdravja. Ključen korak selekcije je odbira najboljših živali za starše naslednje generacije. Zaradi dolgih generacijskih intervalov, velikega vpliva nekaj izbranih moških živalih ter visokih stroškov testiranja in vzreje živali je odbira genetsko najboljših živali izrednega pomena. Hkrati pa je identifikacija teh živali težavna zaradi več virov variabilnosti. Med temi so genetska variabilnost za lastnost, velik okoljski vpliv, segregacija in rekombinacija pri potomcih že identificiranih elitnih živali ter vključenost različnih organizacij v selekcijsko delo. Živali so razvrščene in odbrane na podlagi plemenskih vrednosti (PV), ki so dvakratnik odstopanja potomcev od povprečja populacije (Falconer in MacKay, 1996).

Uspešnost selekcije merimo z doseženim genetskim napredkom (ΔG), za izračun katerega velja izraz

$$\Delta G = \frac{i \times r \times \sigma_A}{L} , \qquad (1)$$

kjer je ΔG = letni genetski napredek, merjen v enotah standardnega odklona na leto, i = intenzivnost selekcije, r = točnost selekcije, σ_A = genetski standardni odklon in L = generacijski interval (Lush, 1945).

V selekciji živali lahko PV napovemo na dva načina, in sicer klasično ali genomsko. Tako klasične kot tudi genomske PV napovemo z metodo mešanega modela, pri čemer je vpliv živali naključen (enačba 2). Za napoved potrebujemo vektor fenotipskih vrednosti (rešitev), matriko sistematskih vplivov in matriko sorodstva. Model za napoved definiramo z izrazom

$$y = \boldsymbol{X}b + \boldsymbol{Z}a + e , \qquad (2)$$

kjer je **X** matrika dogodkov, *b* je vektor ocen za sistematske vplive, **Z** je matrika sorodstva, *a* je vektor iskanih plemenskih vrednosti, *e* pa je vektor ostankov. Za napoved plemenskih vrednosti potrebujemo matriko sorodstva živali, ki nam omogoča izkoristek vseh razpoložljivih informacij o živalih in sorodnikih. Matriko sorodstva lahko zgradimo na podlagi rodovniških podatkov (klasična selekcija) ali na podlagi genomskih podatkov (genomska selekcija).

1.2 NAPOVED GENOMSKIH PLEMENSKIH VREDNOSTI

Razširitev genotipizacije domačih živali je omogočila uvedbo genomske selekcije (genotipizacija = pridobitev podatkov o genotipu na velikem številu genetskih označevalcev). V slednji selekcijo izvajamo na podlagi genomskih plemenskih vrednosti (gPV), ki jih pridobimo z regresijo fenotipskih podatkov na dejansko sorodstvo ocenjeno iz genomskih podatkov (Meuwissen in sod., 2001). V genomski napovedi ocenjujemo vplive posameznih označevalcev, na podlagi

katerih napovemo gPV. Za napoved gPV zaradi večjega števila spremenljivk kot podatkov uporabimo metodo L2 regularizacije (Ridge regresija) ali pa uporabimo Bayesovski pristop, kjer za vplive posameznih označevalcev uporabimo apriori distribucijo (Meuwissen in sod., 2001). Genomske PV lahko napovemo vsem genotipiziranim živalim, tudi tistim brez lastnih fenotipskih vrednosti. Pri slednjih je tako točnost ocene plemenskih vrednosti znatno večja kot pri klasični selekciji. Večja točnost omogoča odbiro staršev naslednje generacije takoj ob rojstvu, kar bistveno skrajša generacijski interval in poveča genetski napredek (Schaeffer, 2006).

Napoved gPV zahteva veliko referenčno populacijo živali z genotipskimi podatki in fenotipskimi vrednostmi. Točnost napovedi gPV je v največji meri odvisna od velikosti referenčne populacije in njene sorodnosti s selekcijskimi kandidati (napovedna populacija). Večjo točnost napovedi gPV lahko dosežemo z večjo referenčno populacijo, večjo sorodnostjo napovedne in referenčne populacije ter manjšo sorodnostjo živali v referenčni populaciji (Pszczola in sod., 2012). Zaradi majhnega števila genotipiziranih in progeno testiranih bikov je v majhnih populacijah točnost napovedanih gPV selekcijskih kandidatov manjša. Prvo možno rešitev tega problema predstavlja uporaba mednarodne oz. tuje referenčne populacije (Schöpke in Swalve, 2016), katere učinkovitost je odvisna od sorodnosti živali med populacijami. Bolj kot so si živali sorodne, večja bo točnost napovedi plemenskih vrednosti. Drugi način izboljšanja točnosti napovedi gPV predstavlja genotipizacija krav. Zaradi manjše količine fenotipskih podatkov na kravo v primerjavi s progeno testiranimi biki je relativen prispevek ene krave k referenčni populaciji manjši kot enega progeno testiranega bika (de Roos, 2011). Posledično ta način prinaša manjši prispevek k točnosti ob enakem vložku sredstev. Rejske organizacije se tako srečujejo s problemom, kako z razpoložljivimi sredstvi sestaviti referenčno populacijo za maksimiranje točnosti napovedi gPV.

1.3 RAZISKOVALNI PROBLEM

Dobra referenčna populacija lahko bistveno izboljša točnost genomske napovedi. Že prehodne študije so se ukvarjale s problemom optimizacije referenčne populacije, pri čemer so uporabile različne metode in optimizirale različne spremenljivke. V študiji Isidro in sod., 2015, so uporabili različne načine vzorčenja za maksimiranje fenotipske variance, v študiji Akdemir in sod., 2015, pa so raziskovalci uporabili genetski algoritem za rešitev kombinatoričnega problema vključitve posamezne živali v referenčno populacijo.

Optimizacija referenčne populacije je še posebej pomembna v majhnih populacijah, kje je zaradi majhnega števila živali doseganje želenih točnosti oteženo. Majhna velikost populacije z majhnim številom progeno testiranih bikov je značilnost vseh pasem govedi, ki jih redimo v Sloveniji. Zaradi malega števila moških živali s točnimi PV moramo v referenčno populacijo vključiti tudi krave. Ker pa so sredstva omejena, moramo zagotoviti, da bomo za razpoložljiva sredstva genotipizirali tiste krave oz. živali, s katerimi bomo maksimirali točnosti genomske napovedi v dani populaciji. Trenutno pri nobeni populaciji ne izvajamo napovedi gPV z lastno referenčno populacijo. Zato nas je zanimalo, ali lahko v slovenski populaciji z optimizacijo referenčne populacije izboljšamo točnost genomske napovedi in dosežemo zadostne točnosti za lastno genomsko napoved.

1.4 RAZISKOVALNI CILJI

Cilj naloge je bil:

- razviti orodje za optimizacijo referenčne populacije na podlagi sorodstva med živalmi v referenčni in napovedni populaciji z genetskim algoritmom,
- primerjati točnost napovedi genomskih plemenskih vrednosti z optimizirano in naključno izbrano referenčno populacijo.

2 MATERIALI IN METODE DELA

Delo je vključevalo tri korake, in sicer simulacijo slovenske populacije govedi, optimizacijo sestave referenčne populacije in napoved gPV z optimizirano in naključno izbrano referenčno populacijo. Shema dela je prikazana na sliki 2.

2. 1 PRIPRAVA ORODJA IN PARAMETROV ZA SIMULACIJO

Najprej smo razvili simulator populacije govedi v selekciji. Uporabnik lahko nastavi vse selekcijske parametre, kar mu omogoča simulacijo specifične populacije govedi. Simulator je zgrajen kot Python ogrodje, ki povezuje programe za izvedbo vseh korakov enega kroga selekcije:

1. AlphaSim (Faux in sod., 2016) za stohastično simulacijo rodovnika in genoma živali,
2. blupf90 (Misztal in sod., 2002) za ocenitev genomskih plemenskih vrednosti,
3. Python program za določitev staršev nove generacije potomcev.

V naši študiji so bili vsi parametri povzeti po slovenski populaciji rjave pasme govedi, ki šteje ~30.000 aktivnih živali, od tega ~10.000 krav. Simulirani genomski podatki vključujejo 10 kromosomov, iz katerih je bilo 10.000 mest izbranih za vzročna mesta za lastnost z dednostnim deležem 0,25. Da smo dosegli primerno začetno strukturo populacije, smo najprej simulirali 20 generacij selekcije z naključno odbiro staršev, nato pa še 20 generacij klasične selekcije.

2.2 OPTIMIZACIJA REFERENČNE POPULACIJE

V naslednjem koraku smo razvili orodje za optimizacijo referenčne populacije. Izbran optimizacijski algoritem je bil hevrističen genetski algoritem.

2.2.1 Genetski algoritem

Glavna lastnost hevrističnih optimizacijskih algoritmov je, da začnejo s poljubno rešitvijo, potem pa preko iteracij po nekem pravilu proizvajajo nove rešitve, ki jih ovrednotijo in sčasoma podajo najboljšo rešitev, ki so jo našli tekom iskanja. Iterativni proces je ponavadi ustavljen, ko: i) v danem številu iteracij ne uspemo izboljšati rešitve (algoritem konvergira); ii) ko je najdena rešitev dovolj

dobra; iii) ko proces doseže dovoljen računalniški čas; ali iv) ko nek interni parameter konča izvajanje procesa (Maringer, 2005).

Skupina hevrističnih algoritmov je raznolika v tipu uporabljenih metod. Prav tako obstajajo različne razvrstitve hevrističnih algoritmov, ki delijo algoritme glede na različne parametre. Ena izmed bolj uporabnih je delitev glede na število sočasno uporabljenih rešitev. Tako jih delimo na trajektorne metode, ki uporabljajo eno rešitev, in populacijske rešitve, ki delujejo nad populacijo rešitev. Za trajektorne metode je značilno, da pri iskanju opišejo tirnico v iskalnem prostoru, iskalni proces populacijskih algoritmov pa opisuje evolucijo množice točk v iskalnem prostoru. Ker bomo v nalogi uporabili algoritem, ki spada med populacijske metode, se bomo podrobneje osredotočili na to skupino. Med najbolj raziskanimi populacijskimi algoritmi sta optimizacija s kolonijami mravelj (angl. *Ant-Colony Optimisation*) in evolucijsko računanje (angl. *Evolutionary computation*). Oba posnemata fenomena iz narave: medtem ko prvi posnema orientacijo mravelj na podlagi feromonov, drugi posnema proces naravne selekcije (Korošec, 2004).

Evolucijsko računanje posnema koncept evolucije oz. naravne selekcije – imamo populacijo osebkov, v kateri okoljski pritiski ustvarijo naravno selekcijo (preživetje najmočnejšega), kar povzroči naraščanje fitnesa populacije. Najprej naključno ustvarimo populacijo osebkov ter definiramo t.i. fitnes funkcijo, ki je ciljna funkcija, s katero bomo ocenili kvaliteto osebkov. Glede na fitnes najboljše osebke odberemo kot starše prihodnje generacije, ki jo ustvarimo z mutacijo in / ali rekombinacijo dveh staršev. V naslednjem koraku tako starše kot tudi potomce ovrednotimo s fitnes funkcijo in le najboljše prenesemo v naslednjo generacijo. Vsaka iteracija algoritma ustvari novo generacijo osebkov, vsak osebek pa predstavlja eno možno rešitev. Potek evolucijskega algoritma lahko predstavimo z naslednjo shemo:

inicializiraj populacijo (naključen proces)
    osebki (kandidatne rešitve)
ovrednoti vse osebke (s fitnes funkcijo)
DOKLER ne STOP delaj:
    izberi starše iz populacije
    ustvari potomce z mutacijo in rekombinacijo staršev
    ovrednoti novorojene potomce
    zamenjaj nekatere izmed staršev s potomci

Uspeh evolucijskih algoritmov je tako močno odvisen od upravljanja populacije. V splošnem sta v evolucijskem algoritmu dve gonilni sili: selekcija, ki si prizadeva za kakovost in zmanjšuje genetsko variabilnosti populacije, in variacija – implementirana kot mutacija in rekombinacija - ki si prizadeva za novosti in povečuje genetsko variabilnost. Zato je uravnoteženje teh dveh sil ključno za uspeh evolucijskega algoritma (Eiben in Schoeanuer, 2005).

2.2.2 Opis osebka

Napoved gPV zahteva veliko referenčno populacijo. Vendar pa lahko ob istem številu živali v referenčni populaciji dosežemo različno točnost napovedi. Slednja je odvisna od sestave referenčne populacije glede na sorodnost živali v referenčni populaciji in med živalmi v referenčni in napovedni populaciji. Z genetskim algoritmom smo želeli optimizirati izbiro živali, zato je osebek izbiral, katere živali bodo vključene v referenčno populacijo.

Predmet optimizacije je bil napoved gPV v 40. generaciji simulacije. Optimizirali smo sestavo referenčne populacije, točnost napovedi pa preverjali na napovedni populaciji. Napovedna populacija je zajemala 90 novorojenih osebkov v generaciji 40, ki še nimajo lastnih meritev. Nabor kandidatnih živali za referenčno populacijo je zajemal 96 progeno testiranih bikov iz vseh prejšnjih generacij simulacije in 10.653 krav, ki so aktivne v generaciji 40. Slednje so bile rojene v generacijah 34 do 37. Skupno število kandidatnih živali je skupaj štelo 10.749 živali, za katere smo s programom AlphaRelate ustvarili matriko sorodstva na podlagi rodovniških podatkov (Preglednica 1).

Preglednica 1: Primer rodovniške matrike sorodstva **A** za šest živali.

ID živali	87257	87393	87409	88279	96191	100120
87257	1.28	0.73	0.73	0.74	0.57	0.71
87393	0.73	1.3	0.74	0.75	0.58	0.71
87409	0.73	0.74	1.3	0.75	0.58	0.71
88279	0.74	0.75	0.75	1.31	0.58	0.74
96191	0.57	0.58	0.58	0.58	1.22	0.57
100120	0.71	0.71	0.71	0.74	0.57	1.35

Točnost plemenskih vrednosti progeno testiranih bikov je zaradi velikega števila potomcev zelo visoka (~0.99), zato v referenčno populacijo prispevajo največ informacije in največ doprinesejo k točnosti napovedi. Iz tega razloga izbire progeno testiranih bikov nismo optimizirali in smo vse avtomatsko vključeni v referenčno populacijo.

V raziskavi smo optimizirali izbiro krav za vključitev v referenčno populacijo. Zaradi časovnega okvira in izvedljivosti optimizacije smo 10.653 aktivnih krav razvrstili v 100 čred, za kar smo uporabili metodo voditeljev (angl. *k-means clustering*). Razvrstitev smo izvedli na podlagi podatkov o materi ter podatkov o fenotipski in pravi genetski vrednosti krav. Uporabljena metoda voditeljev je implementirana v R-ovem paketu ’stats’. Ustvarjene črede niso štele enako število krav – velikost črede se je gibala od 45 do 205 krav, s povprečno velikostjo črede 107 krav (standardni odklon = 34 krav). Za skrajšanje časa optimizacije smo predhodno ustvarili preglednico povprečnega sorodstva med živalmi v vseh možnih kombinacijah čred krav (Preglednica 2). Na enak način smo ustvarili preglednico povprečnega sorodstva med živalmi v vseh čredah krav in

progeno testiranimi biki ter med živalmi v vseh čredah krav in živalmi v napovedni populaciji (Preglednica 3).

Preglednica 2: Primer preglednice s povprečnim sorodstvom med živalmi v vseh možnih kombinacij čred krav 1 – 100.

Čreda1	Čreda2	Povprečno sorodstvo
1	1	0.7772692308
1	2	0.7740675193
1	3	0.7749728306
1	4	0.7761943436
1	5	0.7769393904
1	6	0.772904344
1	7	0.7760591133
1	8	0.7779345173
1	9	0.7706029541
1	10	0.773313496

Preglednica 3: Primer preglednice s povprečnim sorodstvom med živalmi v vseh čredah krav s progeno testiranimi referenčnimi biki (a) in živalmi v napovedni populaciji (b).

Čreda	a) Povprečno sorodstvo s progeno testiranimi biki	b) Povprečno sorodstvo z napovedno populacijo
1	0.7422315705	0.7767094017
2	0.741169181	0.77498659
3	0.7419913026	0.7760755124
4	0.742869152	0.7758460039
5	0.7429398148	0.7769567901
6	0.7402455357	0.771718254
7	0.7431549658	0.7766940639
8	0.7448733428	0.780040404
9	0.7387739071	0.7700018215
10	0.7401029116	0.7731151272

Odbiro krav za referenčno populacijo smo tako optimizirali preko optimizacije izbire čred krav. En osebek je bil kromosom dolžine 100 (seznam) z enim elementom za vsako izmed čred krav. Prvi element je tako ustrezal čredi 1, drugi elementi čredi 2 itd. Možne vrednosti so zajemale le 0 in 1, pri čemer je 0 / 1 pomenila, da dana čreda krav ni / je vključena v referenčno populacijo. Primer osebka je prikazan na sliki 1.

```
     List size:   100
     List:        [1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0,
0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1,
0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1]
```

Slika 1: Primer osebka optimizacije kot izpis Python knjižnice Pyevolve.

2.2.3 Ciljna funkcija

Za referenčno populacijo so bile izbrane črede krav z vrednostjo 1 na odgovarjajočem elementu kromosoma / osebka. S ciljno funkcijo smo najprej za vsak osebek iz preglednice 2 in preglednice 3a izbrali elemente, ki odgovarjajo izbranim čredam. S povprečjem teh elementov smo izračunali povprečno sorodstvo živali v referenčni populaciji (izbrane črede krav + progeno testirani biki). Nato smo iz preglednice 3b izbrali elemente, ki pripadajo izbranih čredam. S povprečjem teh elementov smo izračunali povprečno sorodstvo med živalmi v referenčni in napovedni populaciji. V optimizaciji smo želeli doseči čim manjše povprečno sorodstvo samih živalih v referenčni populaciji (min) ter čim večje povprečno sorodstvo med živalmi v referenčni in napovedni populaciji (max). Ciljna funkcija je tako maksimirala razliko med povprečnim sorodstvom med referenčno in napovedno populacijo ter povprečnim sorodstvom v referenčni populaciji. zato je funkcija maksimirala razliko med slednjima. Razlike smo vnesli v t.i. reLu funkcijo, ki vrednosti funkcije pripiše 0, če je vrednost negativna, če pa je vrednost nenegativna, pa vrednosti funkcije ne spreminja. Dobljeno razliko smo tudi kvadrirali, da smo poudarili pomembnost večjih razlik v sorodnosti ter algoritmu olajšali iskanje najboljših rešitev. V zadnjem koraku je ciljna funkcija preverila odstopanje števila živali v izbrani referenčni populaciji od pred-nastavljenega želenega števila živali. Osebkom, katerih velikost referenčne populacije je za več kot 15 % odstopala od tarčne velikost, smo pripisali fitnes 0. Fitnes preostalih rešitev je ustrezal kvadrirani razliki sorodstev. V tej raziskavi je bila želena veliko referenčne populacijo 5000.

2.2.4 Parametri genetskega algoritma

Za implementacijo genetskega algoritma smo izbrali Python knjižnico Pyevolve (Perone, 2009). Za korak mutiranja smo izbrali IntegerBinary mutator iz paketa Pyevolve, ki naključno spreminja 0 in 1 elemente kromosoma oz. rešitve. Stopnjo mutacije (angl. *mutation rate*) smo nastavili na 0,01, stopnjo prekrižanja (angl. c*ross-over rate*) pa na 0.001. Za izbiro staršev naslednje generacije smo izbrali »Tournament selection«, ki za starše izbere najboljše osebke. Velikost populacije smo nastavili na 50. Kot kriterij za konec optimizacije smo nastavili število generacij, in sicer 900.

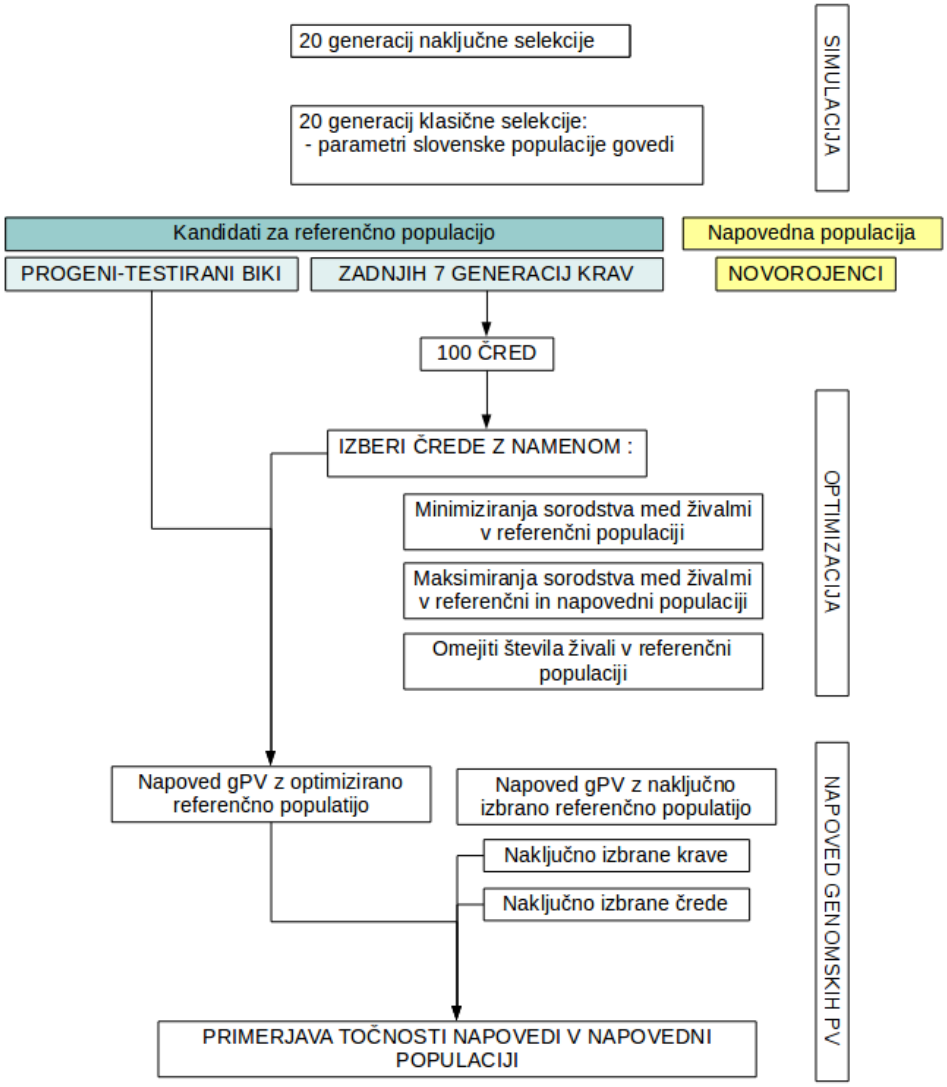
2.2.5 Napoved genomskih plemenskih vrednosti

Optimizacijo smo izvedli v desetih ponovitvah. V vsaki ponovitvi smo k optimizirani referenčni populaciji izbrali tudi dve naključno izbrani referenčni populaciji:

a) naključna izbira krav iz nabora 10.653 krav, pri čemer je bilo število izbranih krav enako nastavljeni omejitvi za število živali v referenčni populaciji;

b) naključna izbira čred krav iz nabora 100 ustvarjenih čred, pri čemer je bilo število izbranih čred enako številu izbranih čred pri optimizaciji.

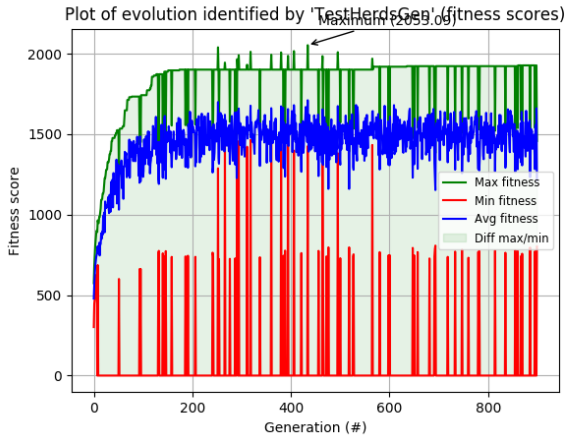
V vsaki od desetih ponovitev smo z vsako izmed optimiziranih in naključno izbranih referenčnih populacij napovedali gPV v napovedni populaciji (90 osebkov). Za napoved gPV smo uporabili program blupf90 (Misztal in sod., 2002). Točnost napovedi smo definirali kot korelacijo med pravimi genetskimi vrednostmi in napovedanimi gPV.

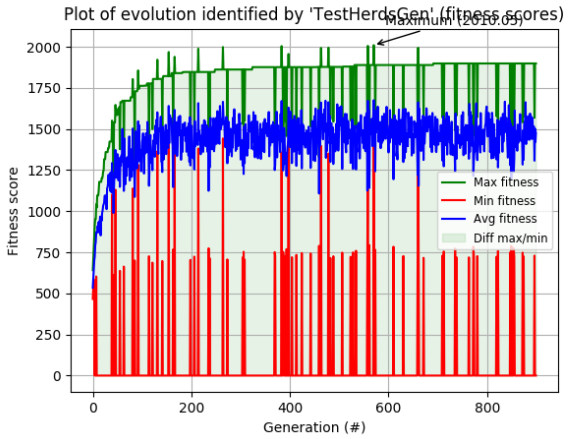


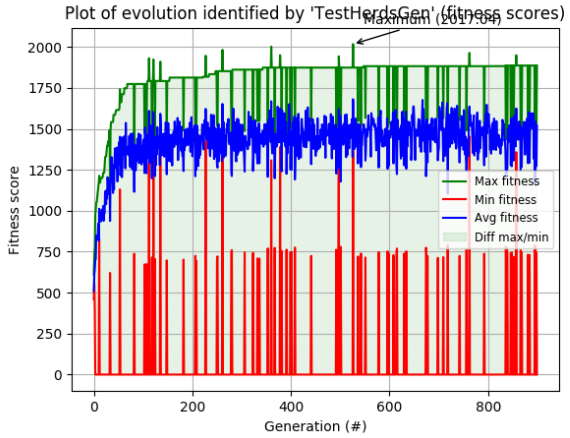
Slika 2: Shema dela.

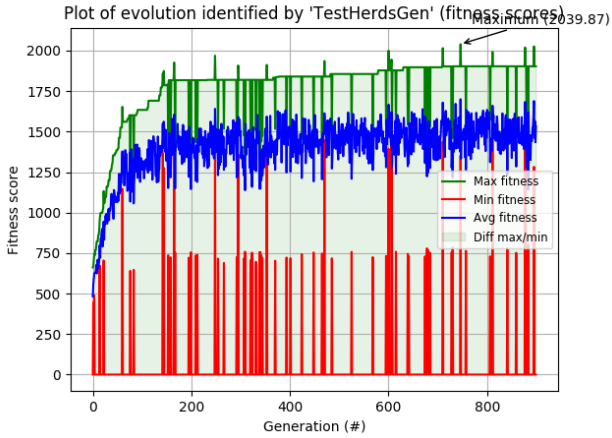
3 REZULTATI

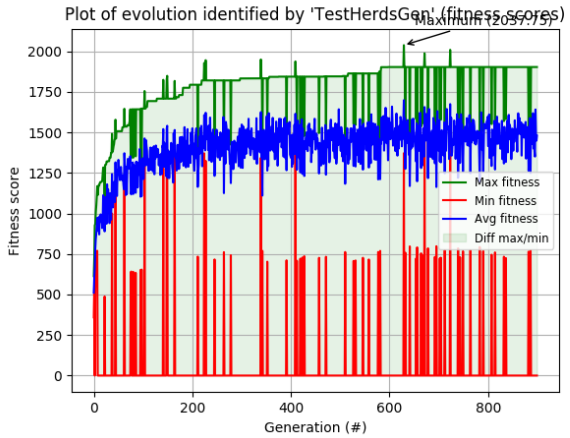
Na sliki 3 je prikazano spreminjanje povprečnega fitnesa populacije skozi iteracije algoritma za vseh 10 ponovitev optimizacije. Vidimo, da je v vseh ponovitvah povprečen fitnes populacije pri okoli 1500. V vseh optimizacijah pa vidimo tudi, da maksimalen fitnes populacije veliko bolj variira – pri čemer je bil maksimalen fitnes posameznega kromosoma dosežen v eni izmed vmesnih iteracij (slika 3).

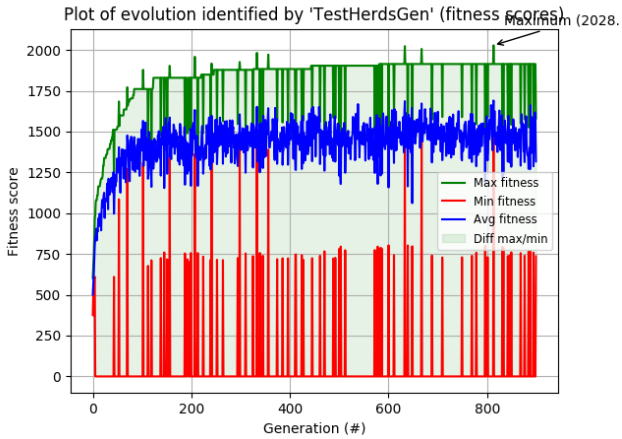

Plot of evolution identified by 'TestHerdsGen' (fitness scores)
Maximum (2059.09)
Fitness score
Generation (#)
Max fitness
Min fitness
Avg fitness
Diff max/min

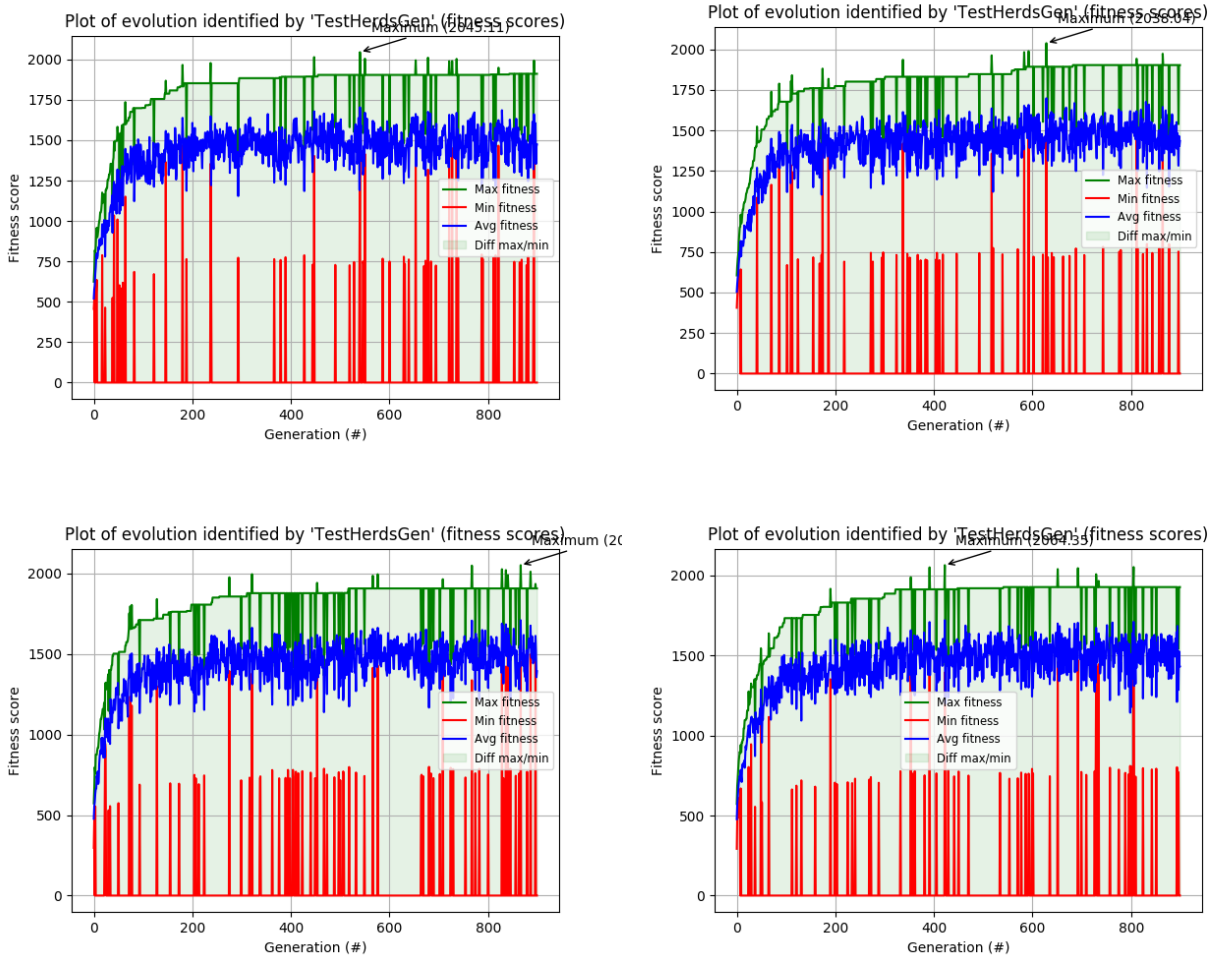

Plot of evolution identified by 'TestHerdsGen' (fitness scores)
Maximum (2010.09)
Fitness score
Generation (#)
Max fitness
Min fitness
Avg fitness
Diff max/min


Plot of evolution identified by 'TestHerdsGen' (fitness scores)
Maximum (2017.84)
Fitness score
Generation (#)
Max fitness
Min fitness
Avg fitness
Diff max/min


Plot of evolution identified by 'TestHerdsGen' (fitness scores)
Maximum (2039.87)
Fitness score
Generation (#)
Max fitness
Min fitness
Avg fitness
Diff max/min

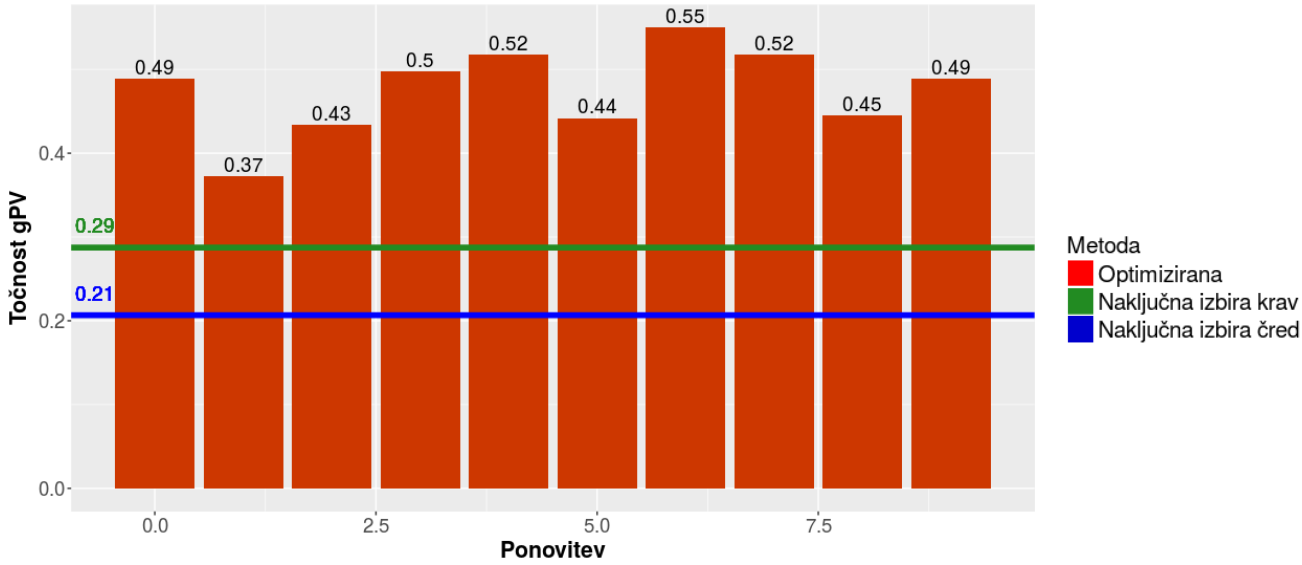

Plot of evolution identified by 'TestHerdsGen' (fitness scores)
Maximum (2037.75)
Fitness score
Generation (#)
Max fitness
Min fitness
Avg fitness
Diff max/min


Plot of evolution identified by 'TestHerdsGen' (fitness scores)
Maximum (2028.
Fitness score
Generation (#)
Max fitness
Min fitness
Avg fitness
Diff max/min



Slika 3: Spreminjanje povprečnega fitnesa populacije rešitev skozi iteracije genetskega algoritma v desetih ponovitvah.

Po optimizaciji smo v vsaki ponovitvi napovedali gPV z optimizirano ali naključno izbrano referenčno populacijo in izračunali točnosti napovedi. Slednje smo izračunali kot korelacijo med pravimi genetskimi vrednostmi in gPV in so prikazane na sliki 4. Za obe naključni izbiri, tako krav kot tudi čred, smo izračunali povprečno točnost preko 10 ponovitev. Vidimo, da smo v vseh desetih ponovitvah z optimizirano referenčno populacijo dosegli večjo točnost napovedanih gPV v napovedni oz. testni populaciji.



Slika 4: Točnost napovedanih genomskih plemenskih vrednosti z optimizirano in naključno izbrano referenčno populacijo v vsaki izmed desetih ponovitev. Optimizirana: z genetskim algoritmom optimizirana izbira čred krav za referenčno populacijo; naključno izbrane krave: naključna odbira krav za referenčno populacijo; naključno izbrane črede: naključna odbira čred krav za referenčno populacijo.

V preglednici 4 so prikazane povprečne točnosti v desetih ponovitvah optimizacije. Vidimo, da v povprečju z optimizirano referenčno populacijo dosežemo 130 % večjo točnost napovedi gPV kot z referenco naključno izbranih krav ter 65 % večjo točnost kot z referenčno populacijo naključno izbranih čred. Pri tem je bilo povprečno število izbranih čred v optimizaciji 36, povprečna velikost optimizirane referenčne populacije 4265, velikost reference z naključno izbranimi kravami je bila stalna, in sicer 5096, povprečna velikost referenčne populacije z naključno izbranimi čredami pa je bila 3853.

Preglednica 4: Povprečna točnost napovedi genomskih plemenskih vrednosti z optimizirano ali naključno izbrano referenčno populacijo.

Metoda	Povprečna točnost
Optimizirana	0,475
Naključna izbira krav	0,287
Naključna izbira čred	0,207

Optimizirana: z genetskim algoritmom optimizirana izbira čred krav za referenčno populacijo; naključno izbrane krave: naključna odbira krav za referenčno populacijo; naključno izbrane črede: naključna odbira čred krav za referenčno populacijo.

4 ZAKLJUČKI

Majhne populacije se soočajo s problemom uvedbe genomske selekcije zaradi majhnega števila živali, ki ima za posledico majhne točnosti gPV, ki ne omogočajo uporabe gPV v selekciji. Prav zato je optimizacija sestave referenčne populacije še posebnega pomena prav v majhnih

populacijah. Z optimizacijo lahko namreč dosežemo zadostne točnosti, ki nam omogočijo uporabo gPV za odbiro živali.

Že predhodne študije so razvile orodja za optimizacijo referenčne populacije, vendar pa kombinacija metod predstavljena v tem delu predstavlja nov pristop k problemu. V tej raziskavi smo za optimizacijo uporabili genetski algoritem, ki teži k minimiziranju sorodstva med živalmi v referenčni populaciji in maksimiranju sorodstva med živalmi v referenčni (trening) in napovedni (testni) populaciji. Pokazali smo, da lahko z optimizacijo referenčne populacijo za do 130 % povečamo točnosti gPV v primerjavi z napovedjo z naključno izbrano referenčno populacijo.

Še vedno pa se povprečna točnost giblje okoli 48 %, kar je premajhna točnost za praktično uporabo v selekciji. Po drugi strani pa je bila omejitev 5000 živali za referenčno populacija zelo stroga in tudi v praksi s takšno referenčno populacijo ne bi pričakovali dejanskega uspeha. Optimizacijo bi tako morali ponoviti z večjim številom za dovoljeno število živali.

V našem delu smo se ukvarjali le z napovedjo gPV v naslednji (eni) generaciji. Selekcija pa je kontinuiran proces, v katerem živali iz napovedne populacije po pridobitvi fenotipskih podatkov vstopijo v referenčno populacijo. Pri tem dosežemo ravno nasproten učinek od želenega, saj je bila referenčna populacija izbrana z namenom maksimiranja sorodstva s živalmi v napovedni populaciji – ko pa se le-te pridružijo referenčni populaciji, pa želimo, da bi bilo to sorodstvo čim manjše. V prihodnjem delu bi lahko tako z razvitim orodjem poskušali nasloviti tudi ta problem.

5 VIRI IN LITERATURA

Akdemir D., Sanchez J. I., Jannink J.-L. 2015. Optimization of genomic selection training populations with a genetic algorithm. Genetics Selection Evolution, 47, 1: 38.

de Roos A.P.W. 2011. Genomic selection in dairy cattle, doktorsko delo. Wageningen University, Wageningen, Nizozemska: 184 str.

Eiben A. in Schoenauer M. 2005. Evolutionary Computing. Information Processing Letters. 82, 1: 1-6.

Falconer D.S., Mackay, T.F.C. 1996. Introduction to Quantitative Genetics. 4. izdaja. Harlow, UK, Longman: 464 str.

Faux A.-M., Gorjanc G., Gaynor R.C., Battagin M., Edwards S.M., Wilson D.L., Hearne, S.J., Gonen S., Hickey J.M. 2016. AlphaSim: Software for Breeding Program Simulation. Plant Genome, 9, 3.

Isidro J., Jannink JL., Akdemir D., Poland J., Heslot N., Sorrells M. E. 2015. Theoretical and Applied Genetics, 128, 1: 145.

Korošec P. 2004. Magistrska naloga: Metahevristično reševanje optimizacijskega problema s kolonijami mravelj. Ljubljana, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko: 80 str.

Lush J.L. 1945. Animal Breeding Plans. 3. izdaja, Ames, Iowa, ZDA, Iowa State University Press: 442 str.

Maringer D. 2005. Portfolio Management with Heuristic Optimization. New York, ZDA, Springer-Verlag New York: 38 – 76.

Meuwissen T.H.E., Hayes B.J., Goddard M.E. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics, 157, 4: 1819–1829.

Misztal I., Tsuruta S., Strabel T., Auvray B., Druet T., Lee D.H. 2002. BLUPF90 and re-lated programs (BGF90). V: Proc. 7th World Congress on Genetics Applied to LivestockProduction. WCGALP, Montpellier, France, 19-23 avg. 2002. Montpellier, WCGALP: 1-2.

Perone C. S. 2009. Pyevolve: a Python open-source framework for genetic algorithms. ACM SIGEVOlution, 4, 1: 12 – 2.0

Pszczola M., Strabel T., Mulder H.A., Calus M.P.L. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. Journal of Dairy Science, 95, 1: 389–400.

Schaeffer L. R., 2006. Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics, 123, 4: 218-223.

Schöpke K., Swalve H.H. 2016. Review: Opportunities and challenges for small populations of dairy cattle in the era of genomics. Animal, 10, 6: 1050–1060.