Vyhodnocování multilokusových dat II

verze 2013-12-18 (T. Fér, F. Kolář)

1. PGDSpider – převod mezi různými formáty dat
2. FSTAT - zjištění základních populačně genetických parametrů kodominantních dat
3. Arlequin - AMOVA (zjištění rozdělení genetické variability)
4. Geneland – Bayesovská analýza genetické struktury s geografickými priors
5. NewHybrids – Bayesovská analýza hybridních jedinců a jejich rodičů
6. K-means clustering – neparametrické určení genetické struktury

1. PGDSpider – převod mezi různými formáty dat

Program slouží k převodu mezi zhruba 30 různými formáty dat používanými pro kodování populačních molekulárních dat. Mimo jiné zvládá konverzi mezi Arlequin, BAPS, FSTAT, GENELAND, IM, MSA, NewHybrids, NEXUS, PHYLIP, Structure ad. Program lze stáhnout z http://www.cmpg.unibe.ch/software/PGDSpider/ a vyžaduje pro svůj běh nainstalovanou Javu (http://www.oracle.com/technetwork/java/javase/downloads/index.html; stahujte JRE, nikoliv JDK!). Grafickou verzi programu spustíme kliknutím na `PGDSpider2.exe`. Vybereme *Data Input File* | *File format:* (např. MSA) a pomocí *Select input file* vybereme příslušný soubor. Stejně tak zvolíme *Data Output File* | *File format:* (např. NEWHYBRIDS). Poté klikneme na *Convert* a program se nás zeptá na doplňující informace ohledně typu dat nebo formátu vstupního/výstupního souboru. Kliknutím na *Apply* je provedena vlastní konverze a uložena do příslušného souboru. (Program použijeme pro konverzi z MSA formátu do formátu pro NewHybrids a FSTAT.)

2. FSTAT - zjištění základních populačně genetických parametrů kodominantních dat

Program FSTAT (http://www2.unil.ch/popgen/softwares/fstat.html) je určen pro práci s alelickými (kodominantními) daty a umí pro jednotlivé populace a lokusy spočítat počty a frekvence alel, Nei´s gene diversity, koeficient inbreedingu, F-statistiku a také její odhady dle Weir & Cockerham (1984; F, θ, f). Kromě toho také pomocí randomizací počítá signifikanci odchylek od Hardy-Weibergovy rovnováhy.
Program FSTAT má vlastní formát dat, případně umí konvertovat i data v GENPOP formátu (tento formát je možné vytvořit pomocí programu MSA a v programu FSTAT použít *Utilities* → *File Conversion* → *Genepop->Fstat*). Data v FSTAT formátu vypadají následovně:

```
6 9 146 3
Loc_1
Loc_2
…
1   088105 139143 095098 143143 089096 088088 053000 046050 139140
1   105105 143143 091098 134143 089096 088088 053000 046050 139140
1   105105 144146 098098 136143 096096 088088 053000 045049 139140
2   089105 139143 091098 134143 089096 088088 053000 046051 139140
2   088105 138143 095098 134143 089096 088088 053000 046051 139140
2   090105 139143 095098 134143 089096 088088 053000 046050 139140
…
```

Čísla na prvním řádku jsou: počet populací, počet lokusů, nejvyšší číslo alely, 3=alela má 3 znaky. Pak následují názvy lokusů a řádky s daty (nejprve číslo populace, pak alely). Pozn. FSTAT označuje populace jako '*sample*'.

Data načteme pomocí *File* → *Open*. Dále zaklikneme, co všechno chceme spočítat a klikneme *Run*. Program zapíše veškeré výsledky do souboru s koncovkou *.out. Ve výsledcích nalezneme např.:

- počty a frekvence jednotlivých alel v jednotlivých lokusech – pro každý lokus je uvedena frekvence alel v jednotlivých populacích a průměrná frekvence přes populace
- genové diverzity, tj. očekávané heterozygosity (H_e) pro jednotlivé lokusy a populace
- počty alel v jednotlivých populacích a lokusech
- inbreeding koeficient pro lokus a populaci (F_{IS})
- odhady heterozygosity podle Nei (H_o, H_S, H_T, G_{ST} ad.)
- odhady F_{IT} (CapF – F), F_{ST} (theta – θ) and F_{IS} (smallF – f) pro každý lokus a alelu a přes všechny lokusy, zjištěny jsou také jackknife odhady přes populace (průměr a SE) a bootstrapy přes populace s 99% a 95% konfidenčním intervalem

3. Arlequin - AMOVA (zjištění rozdělení genetické variability)

Analýza molekulární variance (AMOVA) je používána ke zjištění, jaká část z celkové variability je uvnitř populací a jaká mezi populacemi, případně mezi skupinami populací. To je možné vypočítat např. pomocí programu Arlequin (freeware), verze 3.5 je stažitelná z adresy http://cmpg.unibe.ch/software/arlequin35. Vstupní soubor = datová matice s informací o příslušnosti jedinců do populací/skupiná je poměrně komplikovaný, pro AFLP data je umí naformátovat např. FAMD nebo AFLPDat, pro mikrosatelitová data např. MSA. Automaticky vygenerovaný soubor *.arp obsahuje většinou pouze rozdělení jedinců do populací, pokud chcete (např. kvůli AMOVA) do dat přidat další úroveň hierarchie (skupiny populací), je nutné tak učinit ručně v poslední sekci souboru (níže označena tučně).
Datový soubor může vypadat například následovně (místo textu napsaného *kurzívou* je třeba doplnit vlastní údaje – pozor! – čísla nejsou v uvozovkách, názvy naopak v uvozovkách být musí! „…“ znamená pokračování datového souboru). V části [Data] jsou údaje o jedincích ([[HaplotypeDefinition]]), populacích ([[Samples]]) a skupinách populací ([[Structure]]). Soubor je třeba uložit s koncovkou *.arp.

```
[Profile]
Title="jméno"
NbSamples=počet populací
DataType=RFLP
GenotypicData=0
LocusSeparator=NONE
CompDistMatrix=1
[Data]
[[HaplotypeDefinition]]
HaplListName="jméno"
HaplList= {
sample1            101110101...
sample2            101110101...
sample3            100110011...
...
}
[[Samples]]
SampleName="jméno populace1"
SampleSize=počet vzorků v populaci1
SampleData= {
sample1  počet
sample2  počet
...
```

```
}
SampleName="jméno populace2"
SampleSize=počet vzorků v populaci2
SampleData= {
sample3  počet
sample4  počet
}
[[Structure]]
StructureName="jméno"
NbGroups=počet skupin
#"jméno skupiny1"
Group={
"jméno populace1"
"jméno populace2"
...
}
#jméno skupiny2
Group={
"jméno populace3"
"jméno populace4"
}
```

1. Po spuštění programu  je třeba načíst data (*Open Project*).
2. V levé části okna na záložce *Project* se zobrazí seznam vzorků (*Samples*) a struktura datového souboru (*Groups*).
3. Na kartě *Settings* postupně zaškrtneme, co všechno chceme spočítat, pro AMOVA je to *Genetic structure* → *AMOVA*. V případě mikrosatelitových dat můžeme zvolit metodu výpočtu matice vzdáleností mezi jedinci: *number of different alleles* (počet rozdílných alel, tj. F_{ST} analog) nebo *sum of squared differences* (součet čtverců rozdílů mezi délkami alel, tj. R_{ST} analog). *Locus by locus* AMOVA je vhodná pro případy s množstvím chybějících dat.
4. Zaškrtneme *AMOVA computations* a můžeme kliknout na *Start*.
5. Po provedení výpočtů program vygeneruje webovou stránku v html formátu s výsledky, ve kterých se dá přehledně orientovat pomocí menu v levé části (v IE je třeba povolit spuštění aktivního obsahu!), AMOVA výsledky najdeme pod *Run* → *Genetic structure* → *AMOVA*. Tabulka nazvaná „*AMOVA design and results*“ ukazuje rozdělení celkové variability na jednotlivé úrovně (Source of variation, Percentage of variation). Více viz manuál k programu Arlequin.

Další užitečné hodnoty počítané programem Arlequin

- párové mezipopulační F_{ST} distance – jedná se o míry divergence mezi populacemi, které lze použít jako aproximace genetických distanci mezi populacemi (získanou matici pak možné vynést např. jako distanční strom mezi populacemi): Na kartě *Settings* zaškrtneme *Genetic structure* → *Population comparisons* → *Compute pairwise FST* (analýza založená na párových rozdílech mezi AFLP profily). Ve výsledné html stránce nalezneme matici FST distancí.
- *Slatkin´s linearized* F_{ST}‘s – zkonvertuje párovou matici mezipopulačních koeficientů jako $F_{ST}/(1- F_{ST})$. To se často používá ke korelaci s párovou maticí logaritmů geografických vzdáleností mezi populacemi, kdy při *isolation-by-distance* je vztah přibližně lineární

4. Geneland – Bayesovská analýza genetické struktury s geografickými priors

Program Geneland (http://www2.imm.dtu.dk/~gigu/Geneland/) provádí Bayesovské rozdělování dat do skupin, přičemž zohledňuje vzájemnou geografickou polohu studovaných jedinců a populací.

Vstupní soubor: pro AFLP = binární matice vypadá takto (1 řádek na jedince, uložen jako textový soubor). ! Pozor, v souboru nesmí být žádné monomorfní lokusy (tj. lokusy mající všude jen 1 nebo 0 – zkontrolujeme např. v Excelu):

```
0     1     1     0     1     1     1     1     1
0     1     1     0     1     1     1     0     1
0     1     0     1     1     0     0     0     1
1     1     0     1     1     0     0     0     1
0     0     1     0     1     1     1     1     0
0     0     1     0     1     1     1     0     0
...
```

Pro mikrosatelity = matice alel (1 řádek na jedince, každý lokus = 2 sousední sloupce, 000 = missing value, uložen jako textový soubor):

```
198 000 358 362 141 141 179 000 208 224 243 243 278 284 86 88 120 124
200 202 000 358 141 141 183 183 218 224 237 243 276 278 88 88 120 124
...
```

Vstupní soubor souřadnic vypadá takto (opět txt soubor, první je *longitude* druhá je *latitude*). Pokud máme více jedinců z populace, přiřadíme jim stejné koordináty. Pořadí jedinců musí být stejné jako v datové matici!

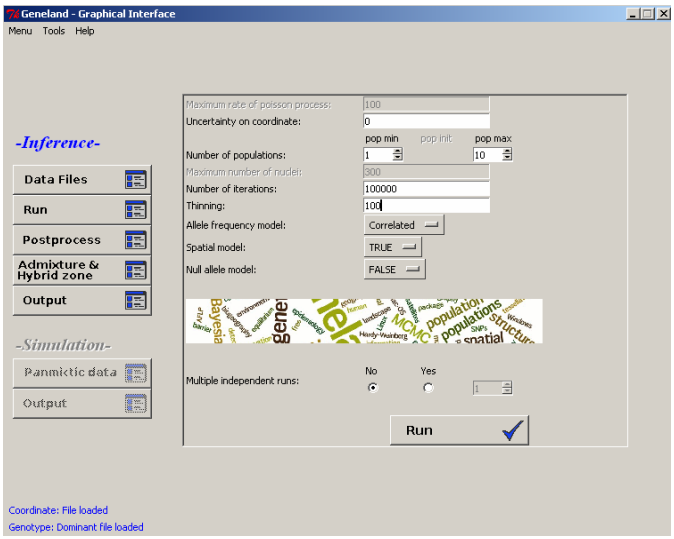
```
13.59327    58.48050
13.59327    58.48050
13.59327    58.48050
13.50051    58.65538
13.50051    58.65538
14.88409    56.79258
15.01303    56.53843
14.84552    56.25931
14.84552    56.25931
```

1. Spustíme R, změníme pracovní adresář na místo, kde máme vstupní soubory (*File* → *Change dir*), a napíšeme následující příkazy:

   ```
   library (Geneland)
   Geneland.GUI()
   ```

2. Otevře se klikací rozhraní. V něm v kartě *Data Files* nejprve zvolíme výstupní adresář a vstupní soubory (*Coordinate file* a *Codominant* nebo *Dominant markers file*).

3. Nastavíme parametry analýzy v kartě *Run*:



- *Uncertainity of coordinate*: pokud je 0 jsou všichni jedinci přiřazeni k téže populaci a také k téže finálně určené skupině (pokud očekáváme smíšené populace, je lepší stanovit >0, např. 1)
- *Number of populations* = rozsah počtu skupin, které budou odhadovány (tzn. K)
- *Number of iterations* = počet MCMC hledání
- *Thinning* = každá kolikátá iteration bude uložena (v našem příkladu bude uložen každý stý výsledek hledání, tzn. 100000/100 = 1000 výsledků celkem)
- *Allele frequency model* = možné zvolit, zda correlated nebo uncorrelated

4. Analýza běží, pokud se zobrazilo okno *Please wait* a doběhne až se otevře okno s nápisem *Terminated with success → OK*.
5. V kartě *Postprocess* nastavíme parametry výstupu, tedy především burnin tzn. kolik prvních uložených iterations bude vyhozeno (v našem příkladu burnin = 200 vyloučí prvních 20% z celkem 1000 uložených iterations).
6. V kartě *Output* prozkoumáme výsledky
   - *Number of populations* – zobrazí histogram, kde by modální hodnota měla odpovídat nejlepšímu K (blíže viz manuál programu!)
   - *Map of estimated population membership* – zobrazí mapu rozdělení jedinců/populací do skupin při vybraném nejlepším K
   - *Map of proba. of pop. membership* (se zaškrtnutým *Save to file?*) – zobrazí a uloží (jako postscript) postupně mapy posterior probabilities příslušnosti do jednotlivých skupin
   - *Fst* and *Fis* zobrazí F-statistiky včetně matice párových Fst distancí jednotlivých identifikovaných skupin (ty je pak možné např. načíst do PASTu a vynést si UPGMA dendrogram jednotlivých skupin)

5. NewHybrids – Bayesovská analýza hybridních jedinců a jejich rodičů

Tento program (http://ib.berkeley.edu/labs/slatkin/eriq/software/software.htm) je určen pro identifikaci hybridů mezi dvěma druhy. Pomocí Bayesovského clusterování je pomocí MCMC vypočtena pro každého jedince pravděpodobnost, se kterou patří do některé z předem

definovaných hybridních tříd (F1, F2, různé typy zpětných kříženců) a/nebo čistých druhů. Jedinci mohou být přiřazeni i do více tříd. Není potřeba dopředu specifikovat, co jsou čisté rodičovské druhy. Vstupní soubor pro NewHybrids připravíme např. pomocí PGDSpider. Datový soubor pro mikrosatelity vypadá následovně:

```
NumIndivs 114
NumLoci 9
Digits 3
Format Lumped
1 088105 139143 095098 143143 089096 088088 053000 046050 139140
2 105105 143143 091098 134143 089096 088088 053000 046050 139140
3 105105 144146 098098 136143 096096 088088 053000 045049 139140
4 105105 143143 091098 143143 096096 088088 053000 050051 139140
5 105105 143143 095098 134143 096096 088088 053000 050051 139140
…
```

V případě AFLP dat jsou tato kódována jako `+/–` . Soubor s datovou maticí nakopírujeme do složky s programem, spustíme soubor `NewHybrids_PC_1_1.exe`. Vepíšeme název vstupního souboru a při otázce na '*genotype frequency classes*' a '*prior allele frequency information*' vložíme '0'. Zadáme dvě čísla pro generator náhodných čísel (neovlivní výsledek analýzy), stiskneme *Enter* a otevře se okno (*Info Window*), kde po stisknutí mezerníku spustíme analýzu. Stiskem '1' se otevře mnoho dalších oken, které ukazují průběh MCMC simulací a jejich průměry pro nejrůznější parametry. Dvě okna vlevo nahoře ukazují pravděpodobnosti zařazení jedinců do jednotlivých tříd – *Category Probabilities* (pro daný MCMC krok a průměr). Okno dole (*Data LogL Trace*) ukazuje graf logaritmu pravděpodobnosti MCMC kroků – je potřeba ho naškálovat kliknutím na něj a stisknutím '*v*', potom již uvidíme, jak po stabilizaci řetězce osciluje hodnota logaritmu pravděpodobnosti modelu okolo určité hodnoty. Burn-in hodnoty odstraním z průměru tak, že poté co po určité době uvidíme, že je řetězec stabilizován, klikneme na 'Info Window' a stiskneme '*e*'. Tím vynulujeme počítání průměru a všechny průměrné hodnoty jsou počítány až z hodnot řetězce následujících po stisku '*e*'. Legendu v každém okně vyvoláme stisknutím '*L*'. Další všemožná nastavení jsou možná pomocí menu, které vyvoláme pravým stiskem myši v každém okně. Program zapisuje výsledky do několika souborů, nejdůležitější je `aa-PofZ.txt`, kde jsou pro každý vzorek zapsány posteriorní pravděpodobnosti příslušnosti do každé z definovaných hybridních tříd.

6. K-means clustering – neparametrické určení genetické struktury

Nehirarchické K-means klastrování se snaží o rozdělení našeho souboru do K skupin tak, aby byla variance mezi těmito skupinami nejvyšší. Opakované odhady K-means clustering pro AFLP data lze spočítat např. pomocí R skriptu publikovaného v Arrigo et al. 2010 New Phytol. Vstupní data uložíme v následujícím formátu jako textový soubor:

```
vz1         pop1        1     1     1     0     0     1     1
vz2         pop1        1     1     1     1     0     1     1
vz3         pop1        1     1     1     1     0     1     1
vz4         pop1        1     1     1     0     0     1     1
vz5         pop1        1     1     1     1     1     1     1
vz1         pop2        1     1     1     0     0     1     1
vz2         pop2        1     1     1     0     0     1     1
```

1. V R nastavíme pracovní adresář (*File → Change dir)* tam, kde máme vstupní soubor a zároveň skript `KMNruns.r`. Skript načteme *File → Source R code*

2. načteme data pomocí následujících po sobě jdoucích příkazů

```
data=read.delim("jmenodatasetu.txt",header=F,row.names=1)
matm=data[,-1]
matm=matm[,!is.na(colSums(matm))]
matm=matm[,colSums(matm)>1] > KMNruns(matm,nreps=50000,kmax=10,'A')
```

3. spustíme výpočet, nastavujeme maximální počet skupin (kmax) a počet opakování výpočtu (nreps) (výpočet trvá poměrně dlouho, řádově až hodiny, a během něj se na obrazovce nic neděje, proto je předem vhodné zkusit na malém počtu opakování, zda se skript spouští správně)

```
KMNruns(matm,nreps=50000,kmax=10,'A')
```

4. skript vygeneruje následující soubory
   - `KMNruns50000A.txt` = výsledek rozdělení do skupin. Pro každé K je zobrazen jen nejlepší výsledek, tj. ten s nejvyšší mezi-skupinovou variancí (zde charakterizovaná jako „inertia“), tato variance je na třetím řádku.
   - `KMNruns_distri50000A.txt` = všechny hodnoty meziskupinové variance („inertia“)
   - `KMNruns_Inertia_SD_50000A.txt` = směrodatné odchylky hodnot meziskupinových variancí pro každé K
   - `pdf` soubory zobrazující (i) vzrůstající hodnoty meziskupinové variance („inertia“) a (ii) jejich druhé derivace
5. výběr nejlepšího K lze učinit např. pomocí delta K obdobně jako u STRUCTURE výsledků, pro tento účel stačí do tabulky `DiagGraphs.xls` doplnit hodnoty „inertia“ ze souboru `KMNruns50000A.txt` a jejich směrodatné odchylky ze souboru `KMNruns_Inertia_SD_50000A.txt`