Fylogenetické analýzy sekvenačních dat – Lekce II.

(Tomáš Fér, Pavel Škaloud, Eliška Záveská, Filip Kolář – PřF UK v Praze)

24. října 2013

Analýza sekvenačních pomocí různých metod k rekonstrukci fylogeneze a následná inspekce fylogenetických stromů a jejich topologií:

I. konstrukce fylogenet. stromů na základě genet. vzdáleností [MEGA, PAUP]
II. konstrukce fylogenet. stromů pomocí Maximum Likelihood [Garli, PAUP]
III. konstrukce fylogenet. stromů pomocí Bayesovské analýzy[MrBayes]
IV. konstrukce fylogenet. stromů pomocí Maximální parsimonie [PAUP]
V. testy alternativních topologií stromů využívající ML [PAUP]
VI. testy inkongruence – ILD test [PAUP]
VII. vizualizace stromů [Treeview, FigTree, Splitstree, Dendroscope]

I. Výpočet fylogenetických stromů na základě genetických vzdáleností

Neighbor-Joining – program MEGA

- programy:
  - MEGA: http://www.megasoftware.net/
- zdrojové alignmenty (fylogeneze druhů krásivkového rodu *Micrasterias*):
  - Micrasterias.fas
- Konverze Fasta formátu do formátu MEGA
  - Otevřít soubor v programu MEGA
  - Data – Export Alignment – MEGA Format (OK, No)
  - Zavřít Fastaalignment, otevřít alignment ve formátu MEGA
- NJ analýza
  - Phylogeny – Construct/Test Neighbor-JoiningTree… – Yes
    - Test Phylogeny – Bootstrapmethod (1000 replikací)
    - Model/Method – vybrat podle model testu
    - RatesamongSites – vybrat podle model testu
    - Compute
- Úprava výsledného stromu
  - Zakořenění stromu: vybrat větev a kliknout na ikonku 
  - Optimalizace tvaru stromu: View – Arrange Taxa – ForBalancedShape
  - Úprava velikosti stromu, délky větví: View – Options – karta Tree
  - Export stromu: Image – Safe as PDF/TIFF file (vektor/rastr)

Neighbor-Joining – program PAUP

- programy:
  - PAUP
- zdrojové alignmenty (fylogeneze druhů krásivkového rodu *Micrasterias*):
  - Micrasterias.nex
- NJ analýza
  - Na konec Nexus souboru přidat „PAUP_NJ.block“:

```
Begin paup;
    set autoclose=yes warnreset=no increase=auto;
    log start = yes file = NJ.log replace=yes;
    factory;

    set criterion=distance;
    dset distance=jc objective=me subst=all negbrlen=setzero;
    nj showtree=no breakties=random BioNJ=yes;

    savetrees brlens=yes maxDecimals=4 file=NJ.tree replace=yes;
    bootstrap search = NJ nreps = 1000 conlevel = 50;
    log stop;
end;
```

automatické zavření okna po dokončení analýzy
vytvoření logu – záznamu o průběhu analýzy
reset nastavení PAUPu

počítání distančních analýz
nastavení distanční matice (JC, F81, K2P, GTR, …)
spočítání NJ stromu, možnost výběru BioNJ

uložení výsledného NJ stromu
bootstrapping (1000 pseudoreplikací)
ukončení a uložení logu

II. konstrukce fylogenetických stromů pomocí Maximum Likelihood

Maximum-Likelihood

- programy:
  - Garli: https://code.google.com/p/garli/
  - PAUP
- zdrojové alignmenty (fylogeneze druhů krásivkového rodu *Micrasterias*):
  - Micrasterias.nex - konkatenovaný dataset
    - 1-1782: 18S rDNA
    - 1783-2702: psaA
    - 2703-3285: coxIII
- ML analýza
  - v případě partitions je nutné na konec Nexus souboru přidat informaci o rozdělení datasetu:

```
beginsets;

    charsetSSU      = 1-1782;
    charsetpsaA     = 1783-2702;
    charsetcox      = 2703-3285;

    charpartition marker = SSU:SSU,psaA:psaA,cox:cox;

end;
```

příkaz pro rozdělení alignmentu na partitions

definice první partition
definice druhé partition
definice třetí partition

vytvoření partitions (rozdělení se v tomto případě jmenuje „marker“

- nastavení analýzy probíhá úpravou konfiguračního souboru „Garli.conf“, tento soubor lze najít ve složce programu /example/basic (pro data bez partitions) a /example/partition/templateConfigs (pro data s partitions)
    - garli.3diffModels.bigData.conf = partitions s odlišnými substitučními modely
    - garli.oneModelType.bigData.conf = partitions se stejným modelem
    - garli.mixedDnaMkv.conf = analýza DNA a binárních znaků (např. gap coding)
    - garli.mkv.conf = analýza morfologických znaků
- použijeme konfigurační soubor pro analýzu dat s partitions; následuje tabulka s nastavením důležitých parametrů (ostatní se měnit nemusí):

datafname = Micrasetrias.nex	*název analyzovaného alignmentu*
ofprefix = Garli	*prefix názvu výstupního souboru (nemusí se měnit)*
genthreshfortopoterm = 100000	*podmínka ukončení analýzy (počet generací, kdy se nezlepšil likelihood. Je dobré zvýšit na 100000*
outgroup = 1 7 10 11	*specifikace outgroupu (pořadí sekvencí) pro nastavení orientace výsledného stromu*
searchreps = 1	*počet ML replikací (pro bootstrapování stačí 1 replikace, pro výpočet jediného stromu je vhodné ho zvýšit alespoň na 5*
bootstrapreps =100	*bootstrapping: počet pseudoreplikací , pro výpočet jediného stromu se zadá 0*
[model1], [model2],[model3],...	*specifikace evolučních modelů všech partitions*
ratematrix = 6rate	*nastavení subst. modelu = 1rate (JC69, F81), 2rate (K80, HKY), 6rate (GTR)*
statefrequencies = estimate	*nastvení frekvencí bází = equal (JC69, K80), estimate (u modelů s různými frekvencemi bází*
ratehetmodel = gamma	*nastavení gamma distribuce = none, gamma*
invariantsites = estimate	*nastavení proporce nevariabilních míst = none, estimate*

- pro analýzu stačí spustit program „Garli-2.0.exe“ve složce s konfiguračním souborem a alignmentem
- v případě bootstrappingu je nutné v PAUPu nakonec vytvořit konsenzuální strom na základě 100 získaných stromů. Použijeme následující příkazy:

Paup Micrasterias.nex	*Načtení analyzované hoalignmentu*
log file=MLboot.txt;	*vytvoření logu*
gettrees file=Garli.boot.tre StoreTree Wts=yes;	*získání stromů vygenerovaných v GARLI*
contree all/strict=no majrule=yes usetreewts=yes;	*vytvoření konsenzuálního stromu*
log stop;	*ukončení logu*

III. konstrukce fylogenet. stromů pomocí Bayesovské analýzy

Bayesian inference

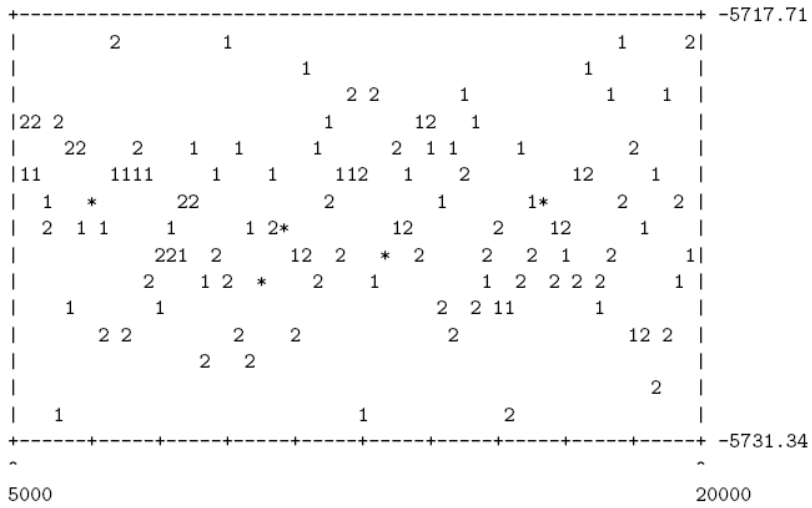
- programy:
  - MrBayes: http://mrbayes.sourceforge.net/
- zdrojové alignmenty (fylogeneze druhů krásivkového rodu *Micrasterias*):
  - Micrasterias.nex - konkatenovaný dataset
    - 1-1782: 18S rDNA
    - 1783-2702: psaA
    - 2703-3285: coxIII
- Bayesovská analýza
  - na konec Nexus souboru přidáme blok pro MrBayes:

```
beginmrbayes;                                          vytvoření logu
charset 18S= 1-1782;                                   definice první partition
charset psa = 1783-2702;                               definice druhé partition
charsetcox= 2703-3285;                                 definice třetí partition
partitionmarker = 3:18S,psa,cox;                       definice dělení alignmentu na partitions
set partition = marker;                                aplikace partitions

prsetapplyto=(all) ratepr=variable;                    definice priors (partitions se definují v závorkách);
                                                       ratepr=variable: partitions mohou mít různé
                                                       evoluční rychlosti
lsetapplyto=(1)  nst=6  rates=invgamma;                definice substitučních modelů (partitions opět
lsetapplyto=(2)  nst=6  rates=propinv;                 v závorkách); nst=1 (JC, F81), nst=2 (K80, HKY),
lsetapplyto=(3)  nst=2  rates=equal;                   nst=6 (GTR); rates=equal / gamma (Γ) / propinv
                                                       (I) /invgamma (Γ + I)

unlinkstatefreq=(all) revmat=(all) tratio=(all) shape=(all)  každá partition může mít různé parametry
pinvar=(all);                                          substitučních modelů

mcmcngen=5000000 samplefreq=100relburnin=no;           nastavení mcmc analýzy: ngen=počet generací;
                                                       samplefreq=míra zapisování parametrů modelů

end;
```

  - mimo výše zmíněných parametrů je analýza založena na velkém množství defaultně nastavených parametrů. Jejich výpis lze získat příkazem „help“:
    - helpprset = výpis priors
    - helplset = výpis parametrů modelů
    - helpmcmc = výpis parametrů mcmc analýzy
  - Bayesovskou analýzu spustíme příkazem „mrbayes“ a poté „exeMicrasterias.nex“

- vyhodnocení analýzy
    - na konci analýzy se MrBayes zeptá, jestli chceme pokračovat v analýze. Rozhodneme se podle hodnoty „average standard deviationof split frequencies“, která vypovídá o konvergenci dvou nezávislých analýz. Pokud je hodnota vyšší než 0.05, měli bychom pokračovat v analýze. Hodnoty pod 0.01 vypovídají o velmi dobré konvergenci
    - tvorba konsenzuálního stromu:

sump**burnin=1000**	*summary parametrů modelu. Burnin = počet generací, které jsou odstraněny před výpočtem konsenzuálního stromu. Číslo burnin navyšujeme tak dlouho, než dojde ke stacionárnímu stavu (zobrazený graf obsahuj enáhodně umístěné 1, 2 a *) – viz obrázek níže*
-nebo-	
sumprelburnin=yes**burninfrac=0.1**	
sumt**burnin=1000**contype=allcompat	*vytvoření konsenzuálního Bayesovského stromu, s použitím hodnoty burnin zjištěné výše.*



Ideální výstup příkazu sump

pokud nedojde k promíchání jedniček a dvojek, či má graf vzestupnou či sestupnou tendenci, je nutné Bayesovský výpočet opakovat či výrazně navýšit počet generací.

Dále zkontrolujeme hodnoty PSRF+ vypsané pod grafem. Hodnoty by se měly oscilovat okolo 1.

IV. konstrukce fylogenet. stromů pomocí Maximální parsimonie

<u>Maximum parsimony</u>

- programy:
  - PAUP
- zdrojové alignmenty (fylogeneze druhů krásivkového rodu *Micrasterias*):
  - Micrasterias.nex - konkatenovaný dataset
    - 1-1782: 18S rDNA
    - 1783-2702: psaA
    - 2703-3285: coxIII
- výpočet MP stromu
  - Na konec Nexus souboru přidat „PAUP_MP.block“:

```
beginpaup;
    set autoclose=yeswarnreset=no increase=auto;
    log start = yesfile = log.logreplace=yes;
    factory;

    set criterion=parsimonymaxtrees=10000 increase=no;

    hsearchdstatus=5      start=stepwiseaddseq=randomnreps=100
swap=tbr status=no;

    savetreesbrlens=yesmaxDecimals=4 file=output.trereplace=yes;
    log stop;
end;
```

automatické zavření okna po dokončení analýzy
vytvoření logu – záznamu o průběhu analýzy
reset nastavení PAUPu

počítání MP analýz, nastavení maximálního počtu stromů v paměti (proti zahlcení PC)

nastavení získání startingtree (start, addseq, nreps), nastavení heuristiky (swap=TBR/NNI/SPR)

uložení výsledného MP stromu

- MP bootstrapping
  - Na konec Nexus souboru přidat „PAUP_MPboot.block“:

```
beginpaup;
    set autoclose=yeswarnreset=no increase=auto;
    log start = yesfile = log.logreplace=yes;
    factory;

    set criterion=parsimonymaxtrees=1000 increase=no;

    bootstrapnreps=1000 cutoffpct=0 keepall=yessearch=heuristic;

    savetreesbrlens=yesmaxDecimals=4 file=boot.trereplace=yes
    from=1 to=1 savebootp=nodelabelsmaxdecimals=0;

    log stop;
end;
```

automatické zavření okna po dokončení analýzy
vytvoření logu – záznamu o průběhu analýzy
reset nastavení PAUPu

počítání MP analýz, nastavení maximálního počtu stromů v paměti (proti zahlcení PC)

nastavení bootstrappingu (1000 pseudorepikací)

uložení výsledného MP stromu

- wMPanalýza (vážená parsimonie)
    - Na konec Nexus souboru přidat „PAUP_wMP.block“:

```
beginpaup;
    set autoclose=yeswarnreset=no increase=auto;
    log start = yesfile = log.logreplace=yes;
    factory;

    set criterion=parsimonymaxtrees=10000 increase=no;
    hsearch start=stepwiseaddseq=randomnreps=100 swap=tbr status=no;

    reweight index=rcbasewt=1000 fit=mean;
    hsearch start=stepwiseaddseq=randomnreps=100 swap=tbr status=no;

    savetreesbrlens=yesmaxDecimals=4 file=output.trereplace=yes;

    log stop;
end;
```

automatické zavření okna po dokončení analýzy
vytvoření logu – záznamu o průběhu analýzy
reset nastavení PAUPu

heuristický výpočet MP stromu

nový výpočet MP stromu na základě převážených hodnot parsimonie pro jednotlivé substituce (použit rescaled consistency index)

uložení wMP stromu

V. testy alternativních topologií stromů využívající ML

Pokud jsme získali analýzou našich dat několik různých stromů (např. se liší stromy získané pomocí NJ, MP a ML) nebo chceme porovnat strom získaný z molekulárních dat s jinou evoluční hypotézou (známe např. topologii stromu získanou na základě morfologických dat), můžeme k porovnání topologií použít několik různých testů.

Pro porovnání dvou alternativních hypotéz (topologií stromu) na základě jednoho alignmentu se nejčastěji používají tzv. „Sitewise“ nebo také „Paired-sites“ testy. Obecně tyto testy počítají věrohodnost nulové a alternativní topologie pro každý znak v alignmentu a nakonec vypočítají rozdíl celkových likelihood pro nulovou a alternativní topologii dle vzorce **δ = lnL1 - lnL0**. Podle rozložení statistiky **δ** je spočtena hodnota **p** s jakou můžeme zavrhnout nulovou hypotézu **H0 (tj. topologii L0).**

Jedním z klasických testů tohoto typu je tzv. **KH test** (**Kishino-Hasegawa test**), který počítá rozložení δ analyticky. Tento test je možné spočítat např. pomocí programu **PAUP**.

- programy:
    - PAUP
- zdrojové soubory:
    - matK.nex - alignment cpDNA pro fylogenezi čeledi *Zingiberaceae*
    - matk.tre – soubor s topologiemi dvou stromů, které chceme porovnat (nexus formát)
    - soubory by měly být umístěny ve společné složce s **exe** souborem programu PAUP

- Do souboru s alignmentem ve formátu nexus vložíme příkazy pro PAUP pro výpočet testu

Begin Paup;	
Set criterion=likelihood;	*Nastavení kriteria „likelihood“*
Lset Base=(0.3543 0.1287 0.1395) Nst=6 Rmat=(1.2095 1.4937 0.3784 0.1832 1.4937) Rates=gamma Shape=0.9913 Pinvar=0.4045;	*Nastavení vhodných parametrů modelu evoluce DNA (potřeba vypočítat pro každý dataset zvlášť např. pomocí ModelTest)*
gettrees file=matk.tre allblocks=yes;	*Načtení souboru s porovnávanými stromy*
lscores all / Khtest=normal;	*Vypočtení likelihood scores a provedení KH testu*
end;	

Výsledkem testu je p-hodnota, podle které přijímáme nebo zamítáme nulovou hypotézu (obvykle druhý strom v seznamu/souboru představuje topologii L0, tj. nulovou hypotézu H0).

VI. testy inkongruence – ILD test [PAUP]

Pokud analyzujeme alignment, který vznikl spojením dat z několika různých DNA markerů, tj. různých lokusů DNA, které mohly prodělat odlišnou evoluční historii, je potřeba otestovat, jestli jednotlivé části alignmentu nepodporují jinou evoluční hypotézu (tj. nenavrhují jinou topologii stromu). Pokud v rámci jednoho alignmentu je více partition, které podporují jiné evoluční hypotézy, je potřeba analyzovat jednotlivé partition separátně.
Existuje mnoho testů, které lze pro tyto účely použít, klasickým příkladem je např. **Incongruence lenght difference test** (někdy též „**Partitions homogeneity test**“), zkráceně **ILD test**. Tento test provádíme na alignmentu konkatenovaných sekvencí, ve kterém definujeme hranice jednotlivých úseků pomocí tzv. partition (např. 1. partition – gen z nDNA, 2. partition – cpDNA). Test můžeme provést např. v programu PAUP.

- programy:
  - PAUP
- zdrojové soubory:
  - CHS_matK_concatenated.nex – konkatenovaný alignment úseku CHS (nDNA) a matK (cpDNA) zástupců čeledi *Zingiberaceae*

- Do souboru s alignmentem ve formátu nexus vložíme příkazy pro rozdělení datasetu na partitions a příkazy pro výpočet ILD testu

```
Begin sets;
 charset CHS_exon = 1-959;
 charset matK = 960-3872;
 charpartition genes = CHS_exon:CHS_exon, matK:matK;
End;

Begin PAUP;
hompart partition=genes nreps=500 / start=stepwise addseq=random nreps=10 savereps=no randomize=addseq rstatus=no hold=1 swap=tbr multrees=yes;
log stop;
end ;
```

Rozdělení datasetu na 2 partitions („CHS_exon“ a „matK“)

Příkazy pro spuštění ILD testu (pro urychlení analýzy lze nastavit nižší počet replikací „nreps=100“, ale pro korektní analýzu je vyšší počet spolehlivější)

Po skončení analýzy získáme **p hodnotu**, podle které zamítáme či přijímáme nulovou hypotézu - H0: dataset je homogenní, tj. všechny části datasetu preferují stejnou evoluční hypotézu, stejnou topologii stromu. Pokud je $p < 0.05$, porovnávané partition preferují signifikantně jinou evoluční hypotézu, tj. jednotlivé partition jsou inkongruentní.

VII. vizualizace stromů [Treeview, FigTree, Splitstree, Dendroscope]

Pro zobrazení stromů vypočtených pomocí různých metod rekonstrukce evoluce můžeme využít nejrůznějších programů, které mají každý své klady a zápory ☺.

- programy:
  - Treeview http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
  - FigTree http://tree.bio.ed.ac.uk/software/figtree/
  - Splitstree http://splitstree.org/
  - Dendroscope http://ab.inf.uni-tuebingen.de/software/dendroscope/

Všechny uvedené programy umí zobrazovat stromy ve formátu nexus, tj. formátu který využívají jako vystupní formát programy PAUP, MrBayes apod.

```
#NEXUS

BEGIN TREES;
        TRANSLATE
                1       Z118_FCM_301_Boesenbergia_curtisii,
                2       Z125_FCM_283_Newnania_othostachys,
                3       Z330_ko_FCM_416_Newmania_serpens,
                4       Z436_JLS_1574_Newmania_sp,
                5       Z437_JLS_1581_Newmania_sp_nov,
                6       Z460_JLS_1646_Newmania_sp,
                7       Z13_FCM_170_Riedelia_aff_corallina,

                33      Z499_FCM_032_Tamijia,
                34      Z263_FCM_363_Siphonochilus_decorus,
                35      FCM_102_Globba_marantina,
                36      Z342_M4_Globba_patens,
                37      Z161_FCM_351_Caulokaempferia_appendiculata,
                38      Z212_FCM_469_Hedychium_ellipticum,
                39      Z65_FCM_278_Hedychium_hasselti,
                40      Z20_FCM_188_Boesenbergia_aurantiaca
                ;
        TREE * Strict=
(34,(33,(((28,(26,27,(7,8,(9,10),(11,12)),(17,(16,(15,(13,14)))),((20,(18,19),(21,22)),(23,(24,25))))),(29,30)),(31,32
),(35,36),((37,(38,39)),((2,3,4,5,6),(1,40))))));
        TREE MajRule=
(34,(33,(((28,(27,(7,8,(9,10),(11,12)),(26,(17,(16,(15,(13,14))))),((20,(18,19),(21,22)),(23,(24,25))))),(29,30)),(35,
36),((31,32),((37,(38,39)),((2,3,4,5,6),(1,40)))))));
ENDBLOCK;
```

Treeview a **Figtree** zobrazují i důležité doplňkové informace o stromech jako jsou např. hodnoty **bootstrapových podpor** (např. z MP analýzy) a **posteriorní pravděpodobnosti** (PP) z Bayesovské analýzy.

Načítání stromů do obou programů je pomocí

- *File* → *Open* → vybereme soubor s koncovkou *.tre, *.trees, *.nex, apod.

Pokud jsou součástí souboru i doplňující informace (např. hodnoty bootstrapu) zobrazíme je v Treeview pomocí

- *Tree* → *Show internal edge labels*

Ve FigTree zobrazímé hodnoty bootstrapu v levém sloupci zaškrtnutím okénka u **„node labels“** případně **„branch labels“**. Záložku je třeba ještě robalit a nastavit „Display“ → „Branch times“.
PP hodnoty z Bayesovské analýzy se nastaví obdobně, změnou na „Display“ → „label“.

Přestože program **Splitstree** je primárně určen na konstrukci síti pomocí algoritmu NeighbourNet, je možné ho použít i pro vizualizaci **nezakořeněných!** stromů. Jeho výhodou je snadná manipulace se stromem (**otáčení, barvení větví, labelů apod.**). Po otevření programu otevřeme záložku „source“ (pod hlavním menu) a manuálně vložíme stromy v nexus formátu, které chceme zobrazit (překopírováním obsahu nexus souboru se stromy). Po klepnutí na záložku „Network“ se zobrazí strom. Barevné označení větví pomocí

- *View* → *Nodes and Edges*

Program D**endroscope** je užitečný zejména při **srovnávání dvou a více stromů a pro vizualizaci nesrovnalostí mezi nimi.** Pomocí

- *File* → *Open* – načteme nexus soubor obsahující všechny stromy, které chceme porovnat. Dále pomocí
- *Algorithms* → *Tanglegrams* – vytvoříme spojnice mezi stejnými jedinci na koncových bodech obou stromů (to předpokládá, že v obou stromech je stejný počet jedinců ☺).