Fylogenetické analýzy sekvenačních dat – Lekce VI.

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

12. prosince 2013

Kombinace dat různého charakteru - interpretace nezávislých datových souborů na základě známé fylogeneze (zpravidla molekulární)

I. rekonstrukce ancestrálních stavů znaků **[Mesquite, R]**
II. rekonstrukce ancestrálních tvarů na základě geometricko-morfometrických (GM) dat metodou maximální parsimonie **[tpsTree]**
III. testování fylogenetického signálu v datech, charakteru a rychlosti evoluce znaků - odhady parametrů *lambda*, *kappa* a *delta* pomocí ML metod **[R, BayesTraits] a** pomocí Bayesovské analýzy **[BayesTraits]**

I. Rekonstrukce ancestrálních stavů znaků v programech Mesquite a R - knihovna GEIGER

Pokud máme k dispozici fylogenezi studované skupiny (tj. fylogenetický strom např. na základě molekulárních dat) a tabulku pozorovaných stavů znaků pro přesně stejné jedince, kteří jsou zastoupeni ve fylogenezi, můžeme si pomocí různých metod a programů vypočítat/odhadnout pravděpodobné stavy znaků ve vnitřních uzlech fylogenetického stromu, tj. odhadnout, jak pravděpodobně vypadal předek vybrané skupiny recentních jedinců. Některé programy/skripty umožňují i praktickou vizualizaci změn hodnot v rámci fylogeneze pomocí zabarvení jednotlivých větví stromu.

Než začneme s rekonstrukcí ancestrálních stavů znaků, je důležité si uvědomit:

- jaký **fylogenetický strom** mámé k dispozici
  - plně **rozlišený**
  - **neúplně rozlišený** (tj strom s polytomiemi, neboli s nulovými hodnotami délek větví) – ty mohou být pro některé metody rekonstrukce závažným problémem
- **jakého typu jsou data** pozorovaných znaků
  - kvalitativní – binární (0, 1) nebo vícestavová, tj. kategorická (A, B, C, D,.. apod)
  - Kvantitativní – **diskrétní** (tj. nespojitá, např.: 1, 3, 5, 8, apod) nebo **kontinuální** (tj. spojitá, např.: 1.23, 1.24., 3.28, apod)
- Jakou metodu chceme/můžeme použít pro rekonstrukci daného typu znaků
  - maximální parsimonie – pro data kategorická (tj. neuspořádaná kvalitativní) i spojitá kvantitativní
  - Maximum likelihood – v R kalitativní i kvantitativní, v Mesquite pouze kategorická data

Podle typu stromu a typu dat, které máme k dispozici si vybereme metodu pomocí které budeme data analyzovat. Pokud je možné data analyzovat více způsoby, vyzkoušíme více způsobů a výsledky porovnáme. Podobnost odhadů na základě různých metod může sloužit jako míra věrohodnoti těchto odhadů.

I.A Rekonstrukce diskrétních a kontinuálních proměnných pomocí metody maximální parsimonie v programu **Mesquite**

- programy:
  - Mesquite: http://mesquiteproject.org/mesquite/mesquite.html - Instalace a spuštění – spuštění instalačního souboru „MesquiteInstaller.exe“ a následné poklepání na vytvořený soubor „Mesquite“
- Zdrojová data:
  - Curcuma_1Cx.txt (příp. Curcuma_2C.txt) – soubor s hodnotami monoploidní velikosti genomu (1Cx) (příp. s absolutními velikostmi genomu, tj. 2C) po logaritmické transformaci pro 61 jedinců z rodu *Curcuma*
  - best_tree_figtree.tre (fylogeneze pro 61 jedinců uvedených v souboru „Curcuma_1Cx.txt“; strom je sestrojen na základě molekulárních dat (ITS) metodou Maximum likelihood)

- Načtení fylogenetického stromu na základě kterého se budou rekonstruovat znaky:
  - Pomocí **File →Open File** otevřeme příkladový soubor “best_tree_figtree.tre”
  - Ve vzniklém „projektu“ pojmenovaném podle právě otevřeného souboru si můžeme prohlédnout, co bylo ze souboru využito za informace – v levém panelu, v záložce „Untitled Block of Taxa“ nalezneme seznam jedinců ve stromu seřazených podle bloku „TRANSLATE“. V záložce „Trees from „best_tree_figtree.tre““ klikneme na piktogram stromu, čímž se zobrazí náš strom. Informace o projektu se zobrazují v samostatných záložkách, které se řadí na lištu pod hlavním menu, mezi těmi můžeme překlikávat nebo je případně vypínat křížkem pravo nahoře.

- Načtení pozorovaných znaků pro dané jedince ve stromu
  - Vytvoříme prázdnou matici pro import hodnot znaků pomocí **Characters →New Empty Matrix**. Program se nás zeptá, kolik znaků budeme sledovat a jaký character dat budeme chtít do matice vložit – pokud použijete data z příkladu, poklepejte na “**Continuous Data”.** Pokud byste chtěli analyzovat data nespojitá, zadáte obdobně **“Standard Categorical Data”** či **“Meristic data”**. Nově vytvořená prázdná matice se otevře v nové záložce, případně jí lze otevřít i z levého panelu.
  - Hlavičky jednotlivých sloupců matice můžeme pojmenovat podle znaků, které budeme vkládat, poklepáním na prázdné místo pod názvy sloupců „1“, „2“ atd. Podle těchto názvů se budeme lépe orientovat v dalším postupu analýzy.
  - Do sloupce k jednotlivým jedincům vložíme hodnoty znaků, např. pomocí Ctr+C zkopírujeme data z excelové tabulky, označíme celý sloupec v Mesquitovské matici poklepáním na záhlaví sloupce a pomocí Ctr+V vložíme. Pokud jsou hodnoty ve tvaru desetinných čísel, je třeba jako oddělovač použít desetinnou tečku.

- Vlastní rekonstrukce ancestrálních stavů znaků
  - Vrátíme se do záložky s načteným fylogenetickým stromem (např. poklepáním na odpovídající záložku na horní liště)

- V hlavním menu zvolíme **Analysis → Trace Character History**. V okně „Source of characters to reconstruct“, které se záhy objeví, vybereme z nabídky „ Stored Characters“ a odklepneme „ok“. V následujícím okně vybereme metodu pro rekonstrukci ancesrálních hodnot, tj. „Parsimony ancestral States“ a odsouhlasíme „ok“. Tuto metodu lze použít jak pro kategorická, tak pro kontinuální data. Metodu Maximum likelihood lze použít v tomto programu jen pro data kategorická (pro kontinuální lze zvolit jiný program, viz níže). Pokud máme vytvořeno více matic (např. jednu pro data kategorická a jinou pro data kontinuální), další okno nás nechá vybrat, kterou matici chceme použít jako zdroj dat.
- Pokud analýza proběhne správně, zobrazí se záložka s vyobrazeným stromem, jehož větve jsou vybarveny dle odvozených hodnot znaku. Barevné schema je definováno v legendě v levém dolním rohu.
- Konkrétní hodnoty znaku na požadované větvi či v požadovaném nodu lze odečítat ve spodní části legendy, pokud na danou větev/nod najedeme myší. Hodnoty lze též možné nechat zobrazit přímo na strom pomocí **Trace → Trace Display Mode → Label States.**
- Pokud jsme do matice vložili více sloupců, tj. analyzujeme více znaků zároveň, v horní části legendy se lze mezi pozorovanými znaky přesouvat pomocí šipek vpravo a vlevo.
- Výsledný projekt je dobré uložit.
- Výsledený strom je možné vytisknout do pdf pomocí **File → Print Tree To Fit Page**.

I.B Rekonstrukce diskrétních a kontinuálních proměnných pomocí metody maximum likelihood v programu **R, balík GEIGER.**

- programy:
  - R (http://www.r-project.org/)
- Zdrojová data:
  - Curcuma_GS.csv – soubor s hodnotami monoploidní velikosti genomu (1Cx) a absolutními velikostmi genomu (2C) pro 61 jedinců z rodu *Curcuma,* stejná data jako pro analýzu v programu Mesquite, jen ve formátu *.csv
  - ITS_ML_curcuma.txt - fylogeneze pro 61 jedinců uvedených v *.csv souboru – stejný strom, jako pro analýzu v programu Mesquite, jen jinak pojmenovaný.

- Po spuštění R se přemístíme do vhodné pracovní složky pomocí **File → Change dir…** a pokračujeme psaním příkazů viz níže.

```
library("geiger")        # načtení knihovny

data = read.csv("Curcuma_GS.csv")     # načtení souboru s hondnotami pozorovaných znaků

tree = read.nexus("ITS_ML_curcuma.txt")    # načtení souboru s fylogenezí v nexus formátu

data       # pro kontrolu - vypsání dat ze souboru “Curcuma_GS.csv"

plot(tree)      # pro kontrolu – vykreslení fylogenetického stromu
```

tree$tip.label # kontrola, zdali jsou labely stromu ve stejnem poradi jako vzorky v *.csv souboru

gas_call = getAncStates(data$X2C_value_log10, tree) # výpočet ancestrálních hodnot pro první sloupec proměnných v *.csv souboru (tj. **X2C_value_log10**) za dané topologie stromu.

plot(tree, cex=0.4) # vykreslení stromu s upravenou velikodtí písma labelů (**cex**)

nodelabels(round(10^gas_call, digits=2), adj = c(1.0, -0.6), frame = "none", cex=0.4)

vynesení hodnot uložených v „**gas_call**“ (tj. ancestrálních hodnot) na odpovídající uzly fylogenetického stromu. Vynesené hodnoty jsou v tomto případě **upravené** „gas_call“, tj. jsou použity jako exponent čísla 10 (10^gas_call), neboť pro výpočet ancestrálních hodnot sloužila jako vstupní data zlogaritmovaná data 2C hodnot a pro zobrazení reálných měřítek odvozených hodnot 2C je třeba odhady zpět převést pomocí exponenciální funkce.

- Získaný grafický výstup analýzy lze uložit aktivací grafického okna a pomocí **File → Save** as uložit např. jako **pdf** apod.
- Obdobně lze analyzovat kterýkoli sloupec proměnných ve vstupním *.csv souboru.

I.C Vizualizace evoluce znaků (alternativa k Mesquite + nadstavba pro R-geiger)**.**

- programy:
  - Perl (Active-Perl) – instalace programu viz protokol k 1. lekci.
    - v Perl PackageManager dále nainstalovat tyto package:
      - IO::File
      - XML::DOM
      - XML::Writer
      - Sort::Array
      - Math::Trig
      - Graphics::ColorUtils
      - SVG
  - TreeGradients (http://www.phycoweb.net/software/TreeGradients/index.html)
- zdrojová data (fylogeneze druhů krásivkového rodu *Micrasterias*):
  - Micrasterias_tree.txt–fylogenetický strom ve formátu NEXUS
  - Micrasterias_tree.new– fylogenetický strom ve formátu NEWICK
  - Micrasterias_data.txt –tabulka morfologických dat
  - ace2tg.R – R skript na vytvoření souboru pro TreeExtender
  - TreeExtender103.pl – Perl skript
  - TreeGradients103.pl – Perl skript
  - evoluce_znaku.R – zde použitý R skript

- Rekonstrukce ancestrálních stavů kontinuální proměnné (skript „evoluce_znaku.R“)

```
# načtení knihovny
library("geiger")
library("ape")
library("phytools")

# načtení souboru s hondnotami pozorovaných znaků
my_data<-read.table("Micrasterias_data.txt")
# načtení souboru s fylogenezí v nexus formátu
my_tree<-read.nexus("Micrasterias_tree.txt")
# kontrola shody jmen taxonů na stromě a v tabulce
name.check(my_tree, my_data)
# ANEBO: treedata(my_tree, my_data, sort=FALSE, warnings=TRUE)
# pro kontrolu – vykreslení fylogenetického stromu
plot(my_tree)

# extrakce analyzovaného znaku z nahrané tabulky
data<-my_data$length
# asociace extrahovaného znaku s názvy taxonů
names(data)<-row.names(my_data)

# výpočet ancestrálních hodnot
ancmasses<-ace(data, my_tree, method="ML")

# extrakce ancestrálních hodnot a jejich zaokrouhlení na 2 desetinná místa
ancmass<-format(ancmasses$ace, digits=2)

# vizualizace ancestrálních stavů
plot(my_tree)
nodelabels(ancmass, adj=c(1.1,-0.6), frame="none", cex=0.6)

# VYTVOŘENÍ SOUBORU PRO TREE EXTENDER
source("ace2tg.R")
ace2tg(ancmasses, my_tree, data, file = "output.txt")
```

- Vizualizace evoluce znaku (Perl)
  - generování fylogenetického stromu s ancestrálními hodnotami
  - v příkazovém řádku:

```
perl TreeExtender103.pl -i Micrasterias_tree.new -o tree.xml -p list -f1 output.txt -vtcontinuous -vnlength
```

  - -i = soubor s fylogenetickým stromem ve formátu NEWICK
  - -o = je možné definovat název generovaného souboru
  - -p = definice typu souboru s ancestrálními hodnotami
  - -f1 = soubor ancestrálních hodnot, generovaný v R
  - -vt = typ mapovaného znaku (kontinuální proměnná)

- -vn = definice jména mapovaného znaku

perl TreeGradients103.pl -t tree.xml -o length.svg -vnlength -gtroyg

- -t = název vstupního souboru generovaného skriptem TreeExtender
- -o = definice názvu výstupního souboru
- -vn = jméno mapovaného znaku
- -gt = specifikace barevné škály gradientu
    - Lineární gradienty:
        - royg red – orange – yellow – green
        - rainbowred – orange – yellow – green – blue – violet
        - bly blue – yellow
        - blw blue – white
        - blg blue – green
        - bw black – white (shadesofgray)
        - g[x-y] shadesofgray, fromdarkness x to darkness y
    - Gradienty tří barev:
        - byr blue - yellow - red
        - gyr green - yellow - red
        - gby green - blue - yellow
        - rgb red - green - blue

II. Rekonstrukce ancestrálních tvarů na základě geometricko-morfometrických (GM) dat metodou maximální parsimonie [tpsTree]

Pokud máme k dispozici, namísto klasických kvalitativních či kvantitativních znaků, informace o tvaru nějakého orgánu či celého organismu získané metodami geometrické morfometriky (např. Thin-plate spline, TPS), můžeme i pro tato data odhadovat/rekonstruovat pravděpodobný tvar předků studované skupiny.

Zde se zaměříme na data získaná landmarkovými metodami GM, konkrétně metodou TPS. Datový soubor popisující vybraný tvar pro každého jedince můžeme získat např. zpracováním hrubých dat (obrázky tvarů) v programech z balíku **Tps** (viz http://life.bio.sunysb.edu/morph/ a přednášky J. Neustupy, https://is.cuni.cz/studium/predmety/index.php?do=predmet&kod=MB120P27). Tento soubor, jehož formát je označován jako *.TPS, zpravidla obsahuje koordináty určitého počtu landmarů pro každého jednice v souboru, viz příklad níže (2 jedinci, 3 landmarky, 6 2D koordinát).

```
LM=3

353.00000 878.00000

501.00000 978.00000

480.00000 1038.00000

IMAGE=aeruginosa_71431
```

```
ID=1

LM=3

-0.012776 -0.136711

0.227299 -0.004281

0.228590 0.077969

IMAGE=parviflora_73223

ID=2
```

Neboť tato data jsou poměrně specifická, pro jejich analýzu se většinou používají GM-specifické programy. Pro rekonstrukci ancestrálních tvarů na základě pozorovaných recentních tvarů a známé fylogeneze můžeme použít jeden z programů výše zmíněného balíku **Tps – TpsTree**. Alternativním programem je pak **MorpoJ** (http://www.flywings.org.uk/MorphoJ_page.htm), který umožňuje kromě rekonstrukce ancestrálních tvarů i mapování fylogeneze na tvaroprostor (tj. PCA na základě GM dat), či testování fylogenetického signálu v GM datech (viz někdy příště, nebo odpovídající webové stránky).

- programy:
  - TpsTree: http://life.bio.sunysb.edu/morph/
- zdrojové soubory:
  - Celek_populacni_prumery_ITS.TPS (soubor koordinát 28 landmarků určujících tvar tyčinek 57 jedinců z rodu *Curcuma*)
  - NJ.nex (molekulární fylogeneze na základě sekvencí ITS – NJ strom v nexus formátu, **jedinci musí být v bloku „TRANSLATE“ uspořádáni ve stejném pořadí jako v souboru *.TPS!**)
- Program spustíme poklepáním na ikonku TpsTree
- V sekci „Input“ poklepeme na tlačítko „Data“ a vybereme vstupní soubor ve formátu *.TPS. Dále poklepeme na tlačítko „Additive tree“ a vybereme soubor „NJ.nex“, tj. soubor s fylogenetickým stromem.
- V sekci „Compute“ poklepeme na tlačítko „Consensus“, čímž vypočteme konsensuální tvar ze všech tvarů v souboru *.TPS, jakož i odchylky jednotlivých tvarů od tohoto consensuálního tvaru. Tento krok zahrnuje i superimpozici všech tvarů (tj. odfiltrování vlivů různého umístění, velikosti a úhlu postavení objektů). Výsledek této analýzy lze zobrazit pomocí tlačítka „Consensus“ v sekci „Display“ v pravém horním rohu okna. V záložce „Options“ v nově otevřeném okně zobrazujícím konsensuální tvar lze zaškrtnutím „points“ a „vectors“ zobrazit všechny analyzované landmarky a jejich odchylky od odpovídajícíh „průměrných“ landmarků.
- Dále v sekci „Compute“ poklepeme na tlačítko „Fit“, čímž se přiřadí tvary jednotlivých jedinců (po úpravě superimpozicí) k odpovídajícím terminálním větvím fylogenetického stromu a vypočtou se též teoretické tvary v jakémkoli místě na stromu od terminálních větví až ke kořenu. Pro zobrazení výsledku poklepeme na tlačítko „Tree“ v sekci „Display“ a otevře se nové okno s vyobrazeným fylogenetickým stromem. Při poklepání na **ikonu fotoaparátu**

vlevo nahoře, zobrazí se tzv. deformační mřížka zobrazující tvar objektu v daném nodu fylogeneze a jeho odchylku od konsensuálního tvaru (odchylka je zobrazena jak vektory u jednotlivých landmarků, tak i deformací mřížky). Přetáhnutím červeného kolečka ve fylogenetickém stromu (defaultně je umístěno u kořene stromu) na jakékoli místo fylogeneze zobrazíte odpovídající tvar objektu v jeho deformační mřížce.

- Rekonstruované tvary objektů lze uložit pomocí **File→ Svae plot**, nebo **Print plot**. Pokud je záhodno odstranit deformační mřížku či vektory zobrazující změnu vůči konsensuálnímu tvaru, lze použít záložku "Options" a odškrtnout "Grid" či odškrtnout "Arrow" v nabídce "Options - Reference".

III. Odhady parametrů *lambda*, *kappa* a *delta* – testování fylogenet. signálu v datech, charakteru a rychlosti evoluce – pomocí Maximum likelihood metod [R, BayesTraits]

Pro testování charakteru evoluce kvalitativních i kvantitativních znaků na základě známé fylogeneze se často využívá odhadů třech parametrů a jejich následné porovnávání s jejich limitními hodnotami, které charakterizují určité hraniční charakteristiky vývoje znaku. Testováním signifikance rozdílu mezi odhadnutou hodnotou a limitními hodnotami se pak usuzuje na pravděpodobný charakter evoluce studovaného znaku při dané fylogenezi (viz přednáška a tam zmíněná literatura).

- **Lambda λ** – Parametr lambda se využíva k testování přítomnosti fylogenetického signálu v porovnávaném datasetu na základě „známé" fylogeneze. Nabývá hodnot 0-1, kdy **λ = 0** odpovídá **absenci phylogenetického signálu** v pozorovaných datech, tj. fylogeneze na základě takovýchto dat by byla hvězdicovitá (nerozlišená). Pokud je **λ signifikantně > 0**, značí to, že je v datech pozorovatelný určitý fylogenetický signál, tzn., že pokud bychom rekonstruovali fylogenezi na základě těchto dat, získali bychom topologii podobnou fylogenetickému stromu, který používáme jako výchozí hypotézu.

- **Kappa κ** – Parametr kappa slouží ke škálování délek větví. Na základě odhadu tohoto parametru můžeme usuzovat, zda je vývoj znaku nezávislý na délce větví, tj. jeho evoluce je tzv. punktualistická (skoková), nebo zda-li delší větve odráží výraznější vývoj znaku (tzv. gradualismus). Kappa nabývá hodnot **0-1**. Hodnota **κ = 0** odpovídá evoluci znaku nezávislému na délce větví fylogeneze, κ < 1 odpovídá tendencím ke zkracování delších větví více než kratších a tedy tzv. punktualistickému modu evoluce. Naopak κ = 1 odpovídá tendencím prodlužovat delší větve více než kratší, což naznačuje, že delší větve více přispívají k evoluci daného znaku.

- **Delta δ** – parametr delta škáluje celkovou délku cesty od terminálních větví po kořen stromu a slouží k posouzení rychlosti evoluce znaku v porovnání s podkladovou fylogenezí . Na základě tohoto parametru posuzujeme, zda-li evoluce znaku spíše akcelerovala nebo se zpomalovala v průběhu času. Delta nabývá hodnot od **0 do ∞,** kdy **δ = 0** není definováno. **δ < 1** indikuje, že kratší cesty (časné fáze evoluce) přispívají nepřiměřeně více k evoluci znaku – tj. indikují adaptivní radiaci: rychlé změny na počátku evoluce skupiny a následné zpomalení změn mezi blízce příbuznými druhy. **δ = 1** odpovídá situaci, kdy rychlost evoluce znaku odpovídá délce větví „podkladové" fylogeneze .**δ > 1** naopak indikuje, že

delší cesta od kořenu k terminálním větvím přispívá více k evoluci znaku – tj. evoluce znaku akceleruje v průběhu času a dochází druhově specifické adaptaci.

Jednotlivé parametry můžeme odhadnout pomocí metody **Maximum likelihood** (programy R – knihovny Ape, Caper, Geiger – a BayesTraits) nebo **bayesovskou analýzou** (poze program BayesTraits). Výhodou odhadů pomocí ML je rychlost analýzy (několik sekund). Výhodou bayesovských odhadů je možnost odhadovat parametry i na základě většího množství stromů a tím zahrnout do analýz i jistou míru nejistoty, pokud není jasné, která fylogenetická hypotéza je „nejlepší“.

Abychom mohli správně interpretovat odhadnutou hodnotu daných parametrů, je třeba porovnat hodnoty věrohodnosti odhadu (tj hodnoty **likelihood**) a hodnoty věrohodnosti parametrů v krajních hodnotách (tj. např. porovnat likelihood pro λ = 1 a pro λ = odhad). Hodnoty věrohodnosti se porovnávají většinou pomocí tzv. **Likelihood ratio testu**, který může být součástí skriptu pro odhady parametrů (např. v knihovně Geiger) nebo si ho můžemespočíst pomocí excelové tabulky apod.

III.A Odhad parametrů *lambda* (*kappa* a *delta*) pro spojitá data v programu R - knihovna Geiger

- programy:
  - R (http://www.r-project.org/)
- Zdrojová data:
  - Alpinoideae_log2C.txt (soubor s hodnotami absolutní velikosti genomu, tj.2C, po logaritmické transformaci pro 95 jedinců z různých rodů podčeledi Alpinioideae)
  - ITS_matK_ML_rooted.nex (fylogenetický strom pro stejné jedince, jako ve výše uvedeném souboru; ML strom na základě molekulárních dat v **newick formátu – tj. názvy jedinců shodné s názvy v txt souboru jsou uvedeny přímo v závorkovém formátu stromu**)
- Odhad parametru **lambda λ**

```
library("geiger")        # načtení knihovny

alpdata <- read.table("Alpinoideae_log2C.txt",sep = "\t", header = TRUE)   # načtení souboru s hondnotami pozorovaných znaků

alptree <- read.tree("ITS_matK_ML_rooted.nex")    # načtení fylogenetického stromu

----------Výpočet hodnot likelihood pro modely kdy λ = odhad, λ = 0 a λ = 1----------

lambda_GL_ML<-fitContinuous(phy= alptree, data = (alpdata$X2C_value_log10),data.names = alpdata$Z_FCM_genus_species,model = "lambda")     # odhad parametru lambda

lambda_GL_ML     # vytištění výsledku analýzy, viz níže

$Trait1

$Trait1$lnl                    ← likelihood odhadu lambda
```

```
[1] -10352386072

$Trait1$beta

[1] 1e-08

$Trait1$lambda

[1] 0.9977149
```

← **odhad parametru lambda**

```
$Trait1$aic

......

alptree_L0<-lambdaTree(alptree, lambda = 0)   # simulace stromu, kdy lambda = 0

plot(alptree_L0)#note this is now a star phylogeny   # vytištění stromu při lambda = 0

alptree_L1<-lambdaTree(alptree, lambda = 1)  # simulace stromu, kdy lambda = 1

plot(alptree_L1)  # vytištění stromu při lambda = 1

lambda_GL_L0<-fitContinuous(phy = alptree_L0, data =
(alpdata$X2C_value_log10),data.names = alpdata$Z_FCM_genus_species, model = "BM")

        # výpočet hodnoty likelihood pro model, kdy lambda = 0

lambda_GL_L0      # vytištění výsledku analýzy, viz níže

$Trait1

$Trait1$lnl

[1] 19.6496
```

← **likelihood modelu při lambda = 0**

```
$Trait1$beta

....

lambda_GL_L1<-fitContinuous(phy = alptree_L1, data =
(alpdata$X2C_value_log10),data.names = alpdata$Z_FCM_genus_species, model = "BM")

        # výpočet hodnoty likelihood pro model, kdy lambda = 1

------------Výpočty Likelihood Ratio Testů ------------------------------------------------------------------

LLR0 <- -2*(lambda_GL_L0$Trait1$lnl - lambda_GL_ML$Trait1$lnl)

        # testování hypotézy, zda odhad lambda je signifikantně odlišný of 0

pchisq(LLR0, df=1,lower.tail = FALSE)

        # test significance pro LLR0 – výsledkem je p-hodnota pro zamítnutí nulové hypotézy
```

```
LLR1 <- -2*(lambda_GL_L1$Trait1$lnl - lambda_GL_ML$Trait1$lnl)

pchisq(LLR1, df=1,lower.tail = FALSE)

! Někde výše je ale asi chyba, protože oba LRT vychazi s p=1. Kdo ji najde dostane * :}

-----------------------------------------------------------------------------------------------------------------------
```

Odhad delta a kappa - podobně jako u lambda, viz materiály v digitální formě a

Alternativní skripty pro výpočet parametrů pomocí knihoven „caper“ a „ape“ také v digitální formě

III.B Odhad parametrů *lambda* (*kappa* a *delta*) pro spojitá data v programu BayesTraits

- programy:
  - BayesTraits (http://www.evolution.rdg.ac.uk/BayesTraits.html)
  - Tracer (http://tree.bio.ed.ac.uk/software/tracer/)
- Zdrojová data:
  - Alpinoideae_log2C.txt (soubor s hodnotami absolutní velikosti genomu, tj.2C, po logaritmické transformaci pro 95 jedinců z různých rodů podčeledi Alpinioideae)
  - ITS_matK_ML_rooted.trees(fylogenetický strom pro stejné jedince, jako ve výše uvedeném souboru; ML strom na základě molekulárních dat v **nexus formátu – názvy jedinců v bloku „TRANSLATE“ ve stejném pořadí jako v txt souboru**)
- Program spustíme pomocí příkazové řádky, nejjednoduššeji z pracovní složky v TotalCommanderu zadáním příkazu „BayesTraits.exe“ následovaným názvem souboru s fylogenezí (měl by mít příponu *,trees)a názvem souboru s pozorovanými daty (měl by mít příponu *.txt). Např.

```
BayesTraits.exe ITS_matK_rooted.trees Alpinoideae_log2C.txt
```

- Otevře se okno s další nabídkou – protože máme spojitá data, vybíráme mezi 4 a 5. Jedná se o různé modely evoluce – A je jednodušší než B – a v ideálním případě by se měly otestovat oba, a následně pomocí LRT testu rozhodnout, jestli B je signifikantně lepší než A. Poté vybrat vhodnější model pro data a zbytek analýz a odhadů provádět za předpokladu tohoto modelu. Pro zjednodušení zvolím model A, tj. zvolíme možnost č.4.
- V další nabídce zvolíme metodu odhadu parametru. Pomocí ML jsme odhadovali parametry již výše, ale není na škodu si odhady zkusit i v tomto programu a porovnat s odhady na základě skriptu výše. Zejména jsou-li hodnoty odhadů jakkoli podezřelé. Nyní ale zvolíme metodu **MCMC**, tj bayesovskou analýzu, tj. č. 2.

- V další nabídce si prohlédneme defaultní nastavení pro MCMC analýzu (např. počet iterací, nastavení burn-in fáze apod..). Vše můžeme zatím nechat v původním nastavení, jen změníme nastavení pro lambda, neboť tento parametr chceme odhadnout. Napíšeme tedy příkaz

  `Lambda`

  a následně

  `info`
- Zobrazí se aktualizované nastavení analýzy, kdy u lambda je místo „Not in use“ nastaveno „Estimate“. Pokud se odhaduje jeden parametr, ostatní by měli zůstat v nečinnosti (viz manuál k BayesTraits).
- Analýzu spustíme pomocí příkazu

  `Run`
- V pracovní složce se záhy objeví log file do kterého se zaznamenávají výsledky analýzy. Po skončení analýzy tento soubor otevřeme, smažeme prvních 25 řádků, ponecháme hlavičky odhadovaných parametrů (tj `Iteration, Tree No, Lh, Hmean, Alpha Trait 1, Trait 1 Var, Lambda, Acceptance`)a soubor uložíme pod jiným jménem – např. „MCMC_lambda_estimate_tracer.txt“
- Upravený log file můžeme otevřít v programu Tracer, kde si můžeme zobrazit
  - rozložení odhadů pro lambda
  - hodnotu likelihood daného modelu,
  - 95% interval spolehlivosti odhadovaných parametrů
  - Apod..
- Analýzu v BayesTraits zopakujeme pro **lambda = 0** a **lambda=1** (místo příkazu lambda v nastavení analýzy použijeme příkaz **`lambda 0`**, příp. **`lambda 1`**), jinak je postup obdobný. Cílem je získat hodnotu likelihood pro oba alternativní modely.
- Následně otestujeme, který model je pro naše data nejvhodnější a to pomocí Likelihood Ratio testu, dle rovnice

  `LR = -2ln*(likelihood H0/likelihood H1)`
- Likelihood ratio statistika má přibližně X^2 rozložení s počtem stupňů volnosti rovným rozdílu v počtu parametrů které lze nastavit v pro jednotlivé modely (H0 - nenastavujeme nic, H1 - nastavujeme lambda, tj. rozdíl je 1). Proto můžepoužít hodnotu LR a počet stupňů volnosti jako vstupní data pro výpočet statistiky X^2 a její hodnotu **p,** na základě které se rozhodneme, zda zamítnout či přijmout nulovou hypotézu. Praktický nástroj pro výpočet této statistiky je dostupný na webu:

  http://www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html

  nebo můžeme statistiku vypočítat v R

  `pchisq(LLR1, df=1,lower.tail = FALSE)`

Další praktické rady pro analýzy v BayesTraits nebo pro odhady parametrů lamdda, kappa a delta naleznete např. na stránkách

http://www.anthrotree.info/wiki/pages/s169r4g/5.3.html