Fylogenetické analýzy sekvenačních dat – Lekce III.

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

7. listopadu 2013

Analýza sekvenačních dat na „mělké“ taxonomické úrovni, rekonstrukce „Species tree“

I. Analýza haplotypů pomocí parsimonických haplotyových sítí [TCS]
II. sítě NNet a detekce rodičovství hybridů a polyploidů [Splitstree]
III. rekonstrukce species tree při incomplete lineage sorting [*BEAST]
IV. Vizualizace gene trees ve species trees [PrimeTv]
V. Odhad hranice druhu [BEAST, R, Genie]

I. Analýza haplotypů pomocí parsimonických haplotyových sítí v programu TCS

- programy:
  - TCS: http://darwin.uvigo.es/software/tcs.html
- zdrojové alignmenty (fylogeneze druhů rodu *Loricaria*):
  - loricaria_upr_TCS.phy
- Konstrukce sítě:
  - Spuštění programu kliknutím na soubor TCS1.21.jar (ve složce TCS1.21)
  - Nastavení analýzy kliknutím na „Start New TCS Analysis“
    - **File → Select NEXUS/PHYLIP Sequence file** – vyberte požadovaný alignment ve formátu Nexus nebo Phylip (sequential format)
    - Nastavení „connection limit“ - zaškrtnutím „**Calculate 95% Connection Limit**”
    - Nastavení „connection limit“ je třeba optimalizovat dle povahy analyzovaných dat. Pro málo divergované sekvence můžeme použít procentuální nastavení podobnosti sekvencí pomocí „**Calculate 90% (-99%) Connection Limit**” ale pro divergovanější data je někdy nutné nastavit „**Fix Connection Limit at**“ a doplnit maximální počet kroků při kterých se ještě dva jedinci spojí. Pokud je limit příliš nízký, vytvoří se sítí několik.
    - V rozbalovacím menu níže je možné nastavit, jak má být naloženo s mezerami v alignmentu (Gap) – vyzkoušejte postupně oboje nastavení a porovnejte výsledek.
    - Spusťte analýzu pomocí tlačítka „**RUN**“
    - Pokud data obsahují pozice kódované IUPAC kódem, objeví se upozornění . potvrďte „ok“
  - Po proběhnutí analýzy se objeví jedna nebo více sítí, které lze rozprostřít do roviny označením jednoho z haplotypů a kliknutím na „**Tree→Tree down**” (případně Tree up, left, right – dle potřeby)
    - Během analýzy program seskupí stejné haplotypy do jedné skupiny

        - Frakvence haplotypů je pak zobrazena jako poměrná velikost jednotlivých oválů (čtverců) ve výsledném grafu
        - Poklepáním na jednotlivé haplotypy získáte informace o jejich rozšíření mezi analyzovanými jedinci
    - Pokud se objevilo více sítí, které se překrývají, rozprostřete je pomocí myši. Přenastavením Connection limit na více kroků lze sítě propojit.
    - Pro výslednou editaci grafu je výhodné graf exportovat ve formátu „postscript“ pomocí **File → Save Network as PostScript“** a poté načíst v některém z grafických programů (např. Adobe Illustrator apod.)

II. Analýza sekvencí v programu SplitsTree za účelem detekce hybridů/polyploidů

Sestrojením sítě algoritmem *neighbour-net* lze vizualizovat konfliktní signály v datasetu. Díky tomu ji lze využít např. pro detekci hybridních jedinců, allopolyploidů nebo rekombinantních sekvencí. Pozor, využívají se distanční metody (v základním nastavení program navíc používá pouze nekorigované p-distance) – jedná se o užitečného pomocníka, ale v analýze sekvenčních dat tento program není plnohodnotnou náhradou parsimonických nebo Bayesovských přístupů tvorby fylogenetických stromů

- programy:
    - Splitstree: (http://www.splitstree.org/
- zdrojové alignmenty (fylogeneze rodů v rámci čeledi Zingiberaceae):
    - GAPDH_Alpinioids.nex
- Konstrukce sítě
    - vstupní formát je NEXUS. *File → Open* (soubor s příponou *.nex), analýza proběhne automaticky s defaultním nastavením (pro orientační analýzu není třeba přenastavovat).
    - z výsledné sítě je možné některé sekvence vyřazovat (pravý klik myší – *Exclude selected taxa*) – zkoumáme jak se mění topologie sítě. Vyřazené sekvence je možné opět vrátit přes *Data → Restore All Taxa*
    - Obrázky je možné snadno exportovat, např. do pdf: (*File → Export Image*).

Hybridní (rekombinantní) sekvence/jedinci se obvykle objevují jako spojnice mezi svými nejbližšími příbuznými, tj v jednom z rohů kosočtverce (viz obrázek 1. níže).

Obrázek 1. Analýza NeighbourNet zobrazující hybridní původ některých jedinců (červeně allopolyploidní jedinci, zeleně pravděpodobní hybridi na diploidní úrovni).

III. Rekonstrukce species tree při incomplete lineage sorting v programu *BEAST

Pro rekonstrukci fylogenetického stromu reprezentující „skutečné“ vztahy mezi analyzovanými druhy (tj. species tree), namísto genealogických stromů reprezentujících pouze fylogenezi dané části genomu, jednoho lokusu (tj. gene trees), se využívá nejčasteji analýz využívajících Bayesovskou analýzu a koalescenční modely speciace. Programy, které tyto analýzy používají jsou např. *BEAST (součást program BEAST od verze 1.6.1.) nebo BEST (součást program MrBayes od verze 3.2). Oba programy využívají podobná vstupní data – dva a více alignmentů reprezentujícíc jednotlivé genové genealogie (tj. “gene trees”) v Nexus formátu. **Každý analyzovaný druh by měl být reprezentován alespoň dvěma jedinci, aby mohly být odhadnuty parametry pro koalescenční model**. Oba programy mají take podobné výchozí předpoklady (Incomplete lineage sorting/deep coalescence je hlavní zdroj rozdílů mezi topologiemi gene trees a species tree).

Výhodou programu *BEAST je např. grafické uživatelské rozhraní a možnost nastavení a sledování vice parametrů analýzy.

- programy:
    - BEAST: http://beast.bio.ed.ac.uk/Main_Page
        - Zahrnuje sérii pomocných programů, které neinstalujeme, ale rovnou spouštíme pomocí souborů:
        - BEAUti v1.7.5.exe
        - BEAST v1.7.5.exe
        - TreeAnnotator v1.7.5.exe
        - LogCombiner v1.7.5.exe
    - Tracer: http://tree.bio.ed.ac.uk/software/tracer/
        - Spouštíme soubor Tracer v1.5.exe
    - FigTree: http://tree.bio.ed.ac.uk/software/figtree/
        - Spouštíme soubor FigTree v1.3.1.exe

- zdrojové alignmenty (fylogeneze rodů z čeledi Zingiberaceae):
    - GAPDH_Alpinioids.nex (nDNA)
    - CHS_Alpinioideae.nex (nDNA)
    - ITS_Alpinioids.nex (nrDNA)
    - matK_trnLF_Alpinioids.nex (cpDNA)

- Rekonstrukce species tree:
    - Otevřeme program BEAUTI (**BEAUti v1.7.5.exe**) a načteme všechny vstupní soubory pomocí **File →Import Data**. Myší označíme všechny soubory a v záhlaví ostupně klikáme na tlačítka "Unlink Subst. Models", "Unlink Clock Models" a "Unlink Trees", čímž zajistíme, že každý dataset bude analyzován na základě vlastního modelu.
    - V levém horním rohu zaškrtneme políčko "**Use Species Tree ancestral reconstruction (*BEAST) Heled & Drummond 2010**"
    - Vyskočí okno s nabídkou, jakým způsobem chceme zadefinovat druhy (seskupit jedince z každého druhu do jednoho OTU). Můžeme druhy zadefinovat manuálně pomocí „**Create a new trait**" a klepnutím na „ok".
    - V následujícím okně ke každému jedinci přiřadíme jméno druhu (jméno druhu se tedy opakuje, pokud analyzujeme dva a více jedninců z daného druhu).
    - V záložce „Sites" zadáme každému analyzovanému lokusu vlastní parametry pro model evoluce DNA (které jsme si predtím pro každý lokus odhadli např. pomocí jModeltest).
        - Např. „Substitution model" – GTR, „Base frequencies" – Estimated, „Site Heterogeneity Model" – Gamma; ostatní necháme defaultně.
        - Provedeme pro všechny lokusy – překlikáváme v levém sloupci „Substitution Model"
    - V záložce „Clocks" nastavíme model pro molekulární hodiny pro každý lokus. Při pilotních analýzách můžeme nechat defaultní nastavení „Strict clock" pro všechny lokusy. Později můžeme zkusit i jiný model molekulárních hodin, např. „Lognormal relaxed clock" a porovnat výsledky. Okénka ve sloupci „Estimate" by měla být zaškrtnuta pro všechny lokusy kromě prvního, neboť mutační rate je pro první lokus

zafixován (rate = 1.0), zatímco mutační rate ostatních lokusů bude odhadnut v závislosti na prvním lokusu. Ostatní nastavení může zůstat defaultně.

- V záložce „Trees“ nastavujeme apriorní parametry pro odhad species tree a gene trees. Nastavení v oddílu „Species Tree prior used to start all gene tree models“ mohou zůstat defaultní. V oddílu „Tree Model“ je třeba pro každý lokus upřesnit v rolovacím menu „**Ploidy type**“, zda se jedná o „autosomal nuclear“ (v případě jaderných genů) lokus, či „mitochondrial“ (v případě plastidových genů). Ve stejném oddílu můžeme nastavit způsob konstrukce počátečního stromu na „UPGMA starting tree“.
- V záložce „Priors“ zkontrolujeme, zda jsou všechna nastavení v pořádku (černým písmem). Pokud jsou některá nastavení červeně, je potřeba je rozkliknout a odsouhlasit znovu defaultní nastavení. Jedná se většinou o situaci, kdy chceme odhadnout nějaký z parametrů a máme velmi široké až nekonečné hranice (interval), ve kterých parametr hledáme – nenastavili jsme vlastně žádnou apriorní informaci. Protože ale „prior“ neznáme, musíme nechat prohledávat celý prostor možností.
- V záložce „MCMC“ nastavíme parametry Bayesovské analýzy, tj.:
    - počet generací, po který se bude prohledávat prostor možných stromů – záleží zejména na počtu analyzovaných vzorků a míře nesourodosti analyzovaných lokusů. Konečný počet generací je tedy potřeba optimalizovat za pomoci kontroly průběhu analýzy v programu Tracer (viz níže). Na poprvé můžeme nechat defaultní nastavení, pokud analyzujeme např. dataset do 30 jedinc. Jinak navyšujeme.
    - Do okénka „File name stem“ zadáme název naší analýzy, nejlépe s datem, kdy analýzu provádíme. Při každé další analýze je dobré jméno analýzy specifikovat, aby se nepletly výstupy jednotlivých analýz (log file apod.).
- V pravém dolním rohu klikneme na tlačítko „Generate BEAST File“
- Objeví se ještě jednou okno vyžadující odsouhlasení nastavení „priors“, odsouhlasíme „continue“ a uložíme *.xml soubor do složky, ve které chcete mít umístěny výsledky analýzy.

- Otevřeme program BEAST (**BEAST v1.7.5.exe**) a pomocí „choose file“ vybereme dříve vytvořený *.xml file a spustíme pomocí „Run“.
    - Průběh analýzy a její předpokládané ukončení můžeme sledovat v nově otevřeném okně.
- Pro kontrolu průběhu analýzy můžeme (a je žádoucí) použít též program Tracer. Otevřeme soubor **Tracer v1.5.exe** a pomocí **File →Import Trace file** načteme *.log file z probíhající nebo již ukončené analýzy programem BEAST. Po načtení souboru kontrolujeme zejména sloupec hodnot “ESS” v levé části okna. Pokud některé z čísel není zobrazeno černě (tj. má hodnotu menší než 200), indikuje to, že analýza ještě neproběhla správně (MCMC řetězce nekonvergují). Pokud jsou ESS hodnoty menší než 200 I po skončení analýzy, je potřeba analýzu spustit znovu s nastavením vyššího počtu generací.

- Po skončení analýzy v programu BEAST a kladném zhodnocení jejího průběhu v programu Tracer je potřeba zkonstruovat konsensuální species tree ze souboru stromů vytvořených během analýzy pomocí programu TreeAnnotator.
- Otevřeme soubor **TreeAnnotator v1.7.5.exe** a do kolonky „Burnin“ vepíšeme, kolik stromů má být odstraněno jako „burnin“ fáze. Obvykle počítáme ca 25% všech zaznamenaných stromů (např. při počtu 10 000 000 generací a zaznamenání každého 1000 stromu vychází burnin fáze na 2500 stromů). Posterior probability limit nastavíme např. na 0.9 a výše. Do kolonky „Input Tree File“ vybereme file se všemi zaznamenanými species tree - najdeme ho mezi ostatními výstupními soubory z analýzy BEAST a má koncovku ***.species.tree**. Po kliknutí na „Choose file“ u kolonky „Output file“ se dostaneme do složky s výstupními fily z analýzy a vytvoříme zde nový soubor pro výstupní strom, např. s koncovkou ***.species_cons.tree**. Spustíme analýzu pomocí „Run“, čímž se vytvoří soubor s konsensuálním stromem.
- Konsensuální species tree se zaznamenanými Posterior probability (statistickými podporami pro jednotlivé větve) si můžeme prohlédnout např. v programu **FigTree.** Hodnoty PP zobrazíme **Display → posterior** v záložce „**Node labels**“ v levé části okna.

IV. Vizualizace gene trees ve species trees pomocí programu PrimeTv

Když jsme získali hypotetickou topologii species tree, rádi bychom se podívali na to, v jakých aspektech se liší jednotlivé gene trees od navrženého species tree. Na základě pozorovaných rozdílů se můžeme lépe rozhodnout, zda za odlišnou topologií gene tree stojí spíše tzv. Incomplete lineage sorting (ancestrální polymorfismus), nebo jiné procesy. Jedním z programů, který vykresluje topologii gene tree do species tree je PrimeTV a to na základě maximálně parsimonického přístupu. Program lze stáhnou z http://prime.sbc.su.se/primetv/ a instalovat, ale daleko příjemnější je webová verze programu na http://prime.sbc.su.se/cgi-bin/primetv.cgi.

- programy:
  - PrimeTV: http://prime.sbc.su.se/cgi-bin/primetv.cgi
- zdrojové soubory:
  - sp-tree-newick-prejmenovany.tre (Species tree druhů z rodu Curcuma)
  - cpDNA-kodovana-newick-prejmenovana.tre (jeden z gene trees – cpDNA)

Před vlastní analýzou je třeba zkontrolovat pojmenování koncových OTU v obou stromech – OTU ve species tree můžeme ponechat jako jednoduché názvy druhů (např. Alismatifolia, bhatii, apod), OTU v gene tree patřící k danému druhu by měly nést stejné jméno, jako má druh ve species tree, ale neboť těchto OTU bude více, je potřeba je indexovat, nejlépe jednoduše pomocí **01_**Alismatifolia, **02_**Alismatifolia apod. Index pro jedince v gene tree bude tedy před „_“ a jméno za „_“ bude odpovídat jménu druhu ve species tree.

- Rekoncilace genealogických stromů do species tree:
  - Na stránkách webového rozhraní programu vložíme do okna „Species Tree“ strom ze souboru „sp-tree-newick-prejmenovany.tre“ a do okna „Gene Tree“ strom ze

souboru „cpDNA-kodovana-newick-prejmenovana.tre“ a klikneme na „Automatic reconcilation“
    - V následujícím okně zkontrolujeme, zda se jméne jedinců v gene tree správně přiřadila jménům druhů ve species tree a klikneme na „Reconcile and view“.
    - V posledním okně si prohlédneme výsledek analýzy. Graf můžeme uložit v několika formátech, např. ve formátu **PostScript** (*.ps), kteyrý můžeme dále otevřít a editovat v programech pro práci s grafikou (jako např. Adobe Illustrator).
    - Postup zopakujeme pro všechny gene trees, které máme k dispozici pro daný species tree.
    - V grafickém programu lze následně zobrazcit všechny gene trees v jediném species tree dohromady.

Upozornění - tato analýza je spíše pomocnou vizualizací složitých vztahů v rámci species tree a je třeba si uvědomit, že např. délky větví jednotlivých gene trees jsou v této analýze ingnorovány a jejich délka je sekundárně odvozena, aby vyhovovala nejparsimoničtějšímu řešení. Je tedy dobré např. porovnat, zdali délky větví jednotlivých gene trees odvozené na základě koalescenční teorie nejsou v příkrém rozporu s délkami větví napržených touto analýzou.

V. Odhad hranice druhů pomocí programů BEAST a Genie

General mixedYule-coalescent model - GMYC

- programy:
    - BEAST (http://tree.bio.ed.ac.uk/software/)
    - FigTree (http://tree.bio.ed.ac.uk/software/)
    - R (http://www.r-project.org/)
    - Genie (http://evolve.zoo.ox.ac.uk/evolve/Genie.html)
- zdrojové alignmenty (kryptická diverzita zelené řasy *Klebsormidium*):
    - Kleb.nex (alignment obsahuje i shodné sekvence)
- vytvoření time-calibrated fylogenetického stromu v programu BEAST
    - BEAUti – načtení nexus formátu se sekvencemi (tlačítko +)
        - Data Partitions–pokud alignment obsahuje různé partitions (je nutné je definovat na konci nexus souboru příkazem beginassumptions; následovanými jednotlivými partitions: charset XX = 1-100, …), pak zvolit „unlingsubstitutionmodels“ (pro každou partition zvlášť)
        - SiteModels – nastavení substitučních modelů
        - ClockModels – nastavení mutačních rychlostí, zadat „estimate“, vybrat Relaxedclock: Uncorrelatedlognormal model
        - Trees – vybrat tree prior: Speciation: YuleProcess, Startingtree: UPGMA generated
        - Priors – vybrat uniform prior pro červeně označené řádky
        - MCMC – 10 mil. generací, Echo state – 10000, Log parameter – 1000

        - Generate BEAST file – OK...
    - BEAST
        - ChooseFile – vybrat vygenerovaný xml soubor, Run. Je možné pomocí BEAGLE zařadit více procesorů
    - Tracer (zobrazení hustoty pravděpodobností v MCMC)
        - tlačítkem + otevřít .logfile MCMC analýzy (estimates ukáže rozdělení dat)
        - podle grafu na kartě „Trace“ nastavit burn-in
    - TreeAnnotator (analyzuje stromy vytvořené MCMC analýzou)
        - nastavit burnin (100x menší číslo než je uvedeno v Traceru), PP limit = 0.5
        - zvolittreesfile z výstupu MCMC (.trees), zvolit jméno output file
    - FigTree
        - nakreslit výsledný stromeček

- GMYC model
  - R (je potřeba mít nainstalovanou package „ape“)
    - počítá se pomocí package „splits“ – viz návod zde: http://barralab.bio.ic.ac.uk/downloads.html
    - použít R script „GMYC.R“:

```
install.packages("splits",repos="http://R-Forge.R-project.org")

library(splits)

my_tree<-read.nexus('klebs_output.tre')

plot(my_tree)

test<-gmyc(my_tree, method = "single", interval = c(0, 5), quiet =
FALSE)

summary(test)

spec.list(test)

plot(test)
```

    - výsledkem analýzy je:
      - test modelu (LR = likelihood ratio test), neboli signifikantního rozdílu ve větvení stromu na hranici druhů
      - počet klastrů (tj. druhů)
      - konfidenční interval počtu klastrů
      - seznam všech sekvencí a jejich zařazení do klastrů
      - grafické výstupy (strom s vyznačenými klastry, lineage-through-time plot, likelihoodsurface)
- pro fajnšmekry: vytvoření vlastíhografulineage-through-time plot
  - Genie
    - optimálně vytvořit v PATH cestu k souboru GENIE.exe, spustit soubor ve složce s uloženými daty a dále jednoduchými příkazy vytvořit ltt plot:
    - loadXXX.tre (NEXUS tree vygenerovaný v BEASTu pomocí TreeAnnotator)
    - log soubor.log (vytvoření logu)
    - lttXXX.tre (vytvoření křivky lineage-through-time)
    - log (zavření logu)
  - log obsahuje data pro X, Y scatter plot, ten lze nakreslit v excelu, R, Statistice, …