Analýza sekvencí a příprava dat pro analýzy – Lekce I.

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

17. října 2013

Analýza sekvenčních dat za účelem připravit data k následné fylogenetické analýze zahrnuje například následující kroky:

I. prohlédnutí sekvencí, zjištění kvality (FinchTV)
II. kontrola totožnosti sekvence (BLAST)
III. vytvoření contigu (SeqMan), uložení ve formátu FASTA
IV. vytvoření alignmentu (MAFFT nebo Clustal)
V. editace alignmentu, práce s datasetem (BioEdit, MEGA)
VI. práce s FASTA formátem (FaBox)
VII. konvertování formátů CLUSTAL, FASTA, NEXUS sequential (FORCON)
VIII. automatické kódování indelů (SeqState)
IX. analýza sekvencí ve Splitstree (neighbour-network), detekce rekombinantů
X. testování modelů evoluce DNA (jModeltest, PartitionFinder, Python)
XI. testování fylogenetické struktury v datech (TreePuzle, PAUP, Excel)
XII. testování substituční saturace sekvencí (MEGA, HyPhy, Perl, Sitestripper, PAUP)

I. **Sekvence** (soubory *.ab1) **si lze prohlédnout** např. v programu **FinchTV** (http://www.geospiza.com/Products/finchtv.shtml). Sekvence by měla být čitelná od cca 30-50 bází do cca 550-650 bází (někdy i mnohem dále). Pokud nevidíme jednoznačné peaky pro jednotlivé báze, je něco v nepořádku. Pouze kvalitní sekvence můžeme dále analyzovat. V programu FinchTV můžeme sekvence editovat, tj. měnit „*base calling*“, mazat i vkládat báze. Program dále umožňuje vyhledávat v sekvencích, export do formátu FASTA, tisk chromatogramu a také prohlížení primárních dat (*raw data*). V případě nejasných sekvencích je vždy nutné se na ně podívat (*View* → *Raw Data*).

Troubleshooting (viz také např. http://www.nucleics.com/DNA_sequencing_support/DNA-sequencing-troubleshooting.html). Pro odhalení problému je nutné si uvědomit, jaký PCR produkt jsme sekvenovali (cpDNA, nrDNA, nDNA, PCR po klonování) a zjistit, zda je problém v PCR, sekvenační reakci, nebo v podstatě amplifikovaného úseku (např. „*multiple-copy*“ vs. „*low-copy*“ úseky).

Sekvence je dlouhá jen několik (desítek) bazí.

- byly osekvenovány zřejmě jen primery, chyba pravděpodobně v sekvenační reakci (nízká koncentrace vstupního PCR produktu, špatné primery apod.). Pokud byl sekvenován PCR produkt po klonování, je pravděpodobné, že klonovací vektor neobsahuje požadovaný insert a byla osekvenována jen část vektoru (špatný výběr kolonie, chyba v ligaci apod.).

Prvních několik desítek bazí (20-100) sekvence je nekvalitních (chaoticky se překrývající píky), zbytek sekvence je dobře čitelný.

- kontaminace sekvenační reakce kratšími produkty, může být i problém na sekvenátoru. Pokud problém přetrvává i po zopakování reakce, je dobré zamyslet se nad optimalizací PCR, případně designem nových primerů.

Sekvence je částečně dobře čitelná a částečně se jednotlivé píky překrývají a tvoří tzv. dvojpíky; případně je celá sekvence dobře čitelná, jen místy se objevují dvojpíky.

- pravděpodobně intra-individuální variabilita amplifikovaného úseku. Pokud byl sekvenován úsek nDNA nebo nrDNA, lze tuto variabilitu vysvětlit jako alelickou variabilitu nebo jako přítomnost více paralogů daného úseku v daném genomu. Řešením tohoto problému je např. klonování daného PCR produktu, nebo design primerů specifických pro konkrétní paralog. Pokud byl sekvenován úsek cpDNA, jedná se nejspíše o kontaminaci a je třeba analýzu opakovat, případně optimalizovat PCR.

Sekvence je nejprve dobře čitelná a v určitém místě se najednou stává nečitelnou, píky se překrývají

- může to být opět intra-individuální variabilita (byly amplifikovány úseky lišící se inzercí/delecí = indel). Pokud se posun stane za úsekem tvořeným větším množstvím stejných bází (např. poly-T úsek), jedná se jen o chybu vzniklou sklouznutím polymerázy. Důležité je zkontrolovat (i) forward vs. reverzní sekvenci (pokud je pouze jeden indel resp. sklouzla polymeráza, mělo by jít celou sekvenci snadno rekonstruovat) a (ii) zda se sekvenční motiv opakuje i v překrývajících se pících, jen je jedna varianta o 1-x (podle délky indelu) posunutá.

Blízko počátku je několik bazí překryto extrémně vysokým a širokým píkem (tzv. blob)

- sekvenační reakce byla nedostatečně nebo špatně přečištěná, velmi často však lze sekvenci pod „blobem" rekonstruovat

II. **Identitu sekvencí** je třeba zkontrolovat pomocí on-line aplikace **BLAST** (Basic Local Alignment Search Tool, http://blast.ncbi.nlm.nih.gov/Blast.cgi), zejména pokud se jedná o nově testovaný marker, či organismus. Ve FinchTV (nebo jiném prohlížeči sekvencí) zkopírujeme čitelnou část sekvence. Otevřeme webové stránky BLASTu a vybereme program „*nucleotide blast*“. Do volného okénka v panelu „*Enter query sequence*“ vložíme zkopírovanou část sekvence ve fasta formátu (viz níže). Prohledáváme většinou pouze databázi „*Others*“ (zaškrtneme v panelu „*Choose search set*“). V panelu „*Program selection*” poté zadáme citlivost hledání na „*Somewhat similar sequences (blastn)*“ a spustíme hledání (tlačítkem „*BLAST*“). Výsledkem hledání je seznam nejpodobnějších sekvencí z prohledávané databáze včetně grafického znázornění, jak nalezené sekvence pasují na naši sekvenci. Pokud nalezené sekvence pocházejí ze stejného nebo blízce příbuzného druhu a odpovídají úseku, který měl být amplifikován, můžeme postoupit k dalšímu kroku.

Troubleshooting

Výsledkem hledání je sekvence z velmi vzdáleného organismu (např. bakterie, člověk, houba apod.) a/nebo nalezený úsek neodpovídá úseku, který měl být amplifikován.

- problém v PCR (kontaminace, málo specifické či velmi degenerované primery preferenčně amplifikující např. symbionta studovaného organismu apod.).

- pro hledanou sekvenci nebyla v databázi nalezena žádná podobná sekvence. Může se stát zejména tehdy, pokud je sekvenován méně obvyklý a vysoce variabilní úsek nDNA (např. intron nějakého genu) u nemodelového organismu. Většinou však lze v sekvenci nalézt alespoň část úseku, která odpovídá exonu genu, nebo konzervativnímu úseku, kde byly umístěny primery.

- pokud byla sekvence získána po klonování, je možné, že málo specifické primery amplifikují preferenčně úsek bakteriálního genomu. Pokud při amplifikaci DNA z kolonií používáme univerzální primery (např. M13F a M13R), může se stát, že amplifikujeme jen bakteriální vektor. K těmto situacím dochází zejména v případech, kdy byl špatně zaligován požadovaný PCR produkt.

III. Pro získání sekvence v celé délce amplifikovaného produktu, je potřeba spojit forward a reverse sekvence (vytvořit tzv. **contig**). V případě méně kvalitních sekvencí lze editovat evidentně špatný „*base calling*“ (překlad peaků na jednotlivé baze). To je možné udělat např. v programu **SeqMan** (součást balíku DNASTAR Lasergene Core Suite, http://www.dnastar.com/t-sub-products-lasergene-seqmanpro.aspx).

- *File* → *New* → *Add Sequences*
- vložit forward i reverse sekvenci od jednoho vzorku (dvojklikem nebo označit + „*Add*“)
- *Assemble*, potom dvojklik na *Contig 1* (pokud jsou sekvence v pořádku a dostatečně se překrývají, měl by se vytvořit pouze jeden contig)
- zkontrolovat, zda je forward sekvence zleva doprava, pokud ne, tak *Contig* → *Complement Contig*
- rozbalit všechny sekvence kliknutím na trojúhelníček vlevo od názvu
- překontrolovat „base calling“ v řádku *Translate* (consensus sekvence), případné jasné chyby zeditovat (přepsat)
- označit myší a zkopírovat (Ctrl+C) horní řádek (*Translate*), případně jeho část
- vložit do textového souboru (např. do poznámkového bloku), název sekvence uvést ve formátu FASTA
- toto provedeme pro všechny sekvence a textový soubor uložíme s koncovkou ∗.fas

FASTA formát – jednoduchý a asi nejběžnější formát při práci se sekvencemi (v názvech sekvencí nepoužíváme mezery, ale podtržítka „_“ nebo svislé čáry“|“). Příklad:

```
>jmeno_sekvence_1|lokus_1

TGCGCTTAATTCCGGTGCCAAAGTCTTCATGGTCGATGCATCCTTAG

> jmeno_sekvence_2|lokus_2

AATGCTCTTAATTCCGGTGCCATGGTCGATGCTTCCTTAGAACATCATT

> jmeno_sekvence_3|lokus_3
```

TTCCGGTGCCATGGTCGATGCTTCCTTAGAACATCATT

Alternativně lze contig samozřejmě vytvářet i v jiných programech (např. Geneious, http://www.geneious.com nebo SeqAssem, http://www.sequentix.de/software_seqassem.php) nebo s využitím webové verze programu CAP3 (http://pbil.univ-lyon1.fr/cap3.php).

Troubleshooting

Program nedokáže spojit požadované sekvence.

- snížená kvalita jedné ze sekvencí nebo obou sekvencí - program nenajde překryv.
- forward a reverse sekvence se překrývají nedostatečně - běžné v případě, kdy je sekvenační reakce předčasně ukončena kvůli úseku mnoha stejných nukleotidů (tzv. poly-T,-A,-C,-G úseky) u forward i reverse primeru. Takové sekvence lze spojit manuálně – pokud jsou forward i reverse sekvence v pořádku až k poly-úseku uprostřed, jediné co nevíme, je počet opakování v tomto úseku (nahradíme ho několika „N“).
- forward a reverse sekvence se nepřekrývají vůbec – je amplifikován příliš dlouhý úsek, nebo je přítomno více poly úseků. Pro sekvenaci PCR produktu je třeba použít i tzv. vnitřní primery (mohou být dostupné, nebo je nutné je nově designovat na základě známých sekvencí). Poté se spojuje do contigu více než dvě, tj. několik forward a reverse sekvencí.

IV. Dále je potřeba vytvořit tzv. **alignment**, tj. seřadit pod sebe homologní báze (*positional homology*). Existuje mnoho algoritmů, jak alignment vytvořit – mezi nejznámější patří Clustal (progresivní metoda), PRRN (iterativní metoda) a MAFFT (kombinace progresivní a iterativní metody). S algoritmy lze experimentovat buď lokálně (instalací příslušných programů) nebo využít webové aplikace:

MAFFT (webová aplikace: http://mafft.cbrc.jp/alignment/server)

- do okna *„Input“* načteme nebo vložíme soubor obsahující sekvence ve fasta formátu.
- v panelu *„Advanced settings“* můžeme nastavit konkrétnější strategii a/nebo parametry pro tvorbu alignmentu (např. pokud analyzujeme blízce příbuzné sekvence, tj. třeba v rámci druhu nebo z blízce příbuzných druhů, je vhodné v sekci *„Parameters“* nastavit *„Scoring matrix for nucleotide sequences: 1PAM/k=2“*).
- tlačítkem *„Submit“* spustíme analýzu
- výstup analýzy je v MAFFT formátu, odkazem nahoře změníme na FASTA format, tento alignment zkopírujeme a uložíme ho do textového souboru s příponou *.fas (FASTA formát s gapy)
- v levé části okna s výstupem jsou grafy s červenými (případně i modrými) přímkami. Za přítomnosti červené i modré přímky graf naznačuje, že ve vstupních sekvencích jsou části sekvencí v *reverse-complement* orientaci oproti ostatním.
- program lze také stáhnout a používat lokálně bez přístupu k internetu (http://mafft.cbrc.jp/alignment/software/windows_without_cygwin.html)

ClustalX (webová aplikace: http://www.ebi.ac.uk/Tools/msa/clustalw2)

- STEP 1 – vložení sekvence ve FASTA formátu, zvolení DNA jako vstupních dat
- STEP 2 – nastavení parametrů párového alignmentu (pokud máme pouze dvě sekvence)
- STEP 3 – nastavení parametrů mnohonásobného alignmentu včetně výstupního formátu (např. NEXUS nebo FASTA)

- STEP 4 – spuštění analýzy

Program lze také stáhnout a používat lokálně (http://www.clustal.org/download/current)

- *File → Load Sequences*
- *Alignment → Do Complete Alignment*
- vytvořený alignment je automaticky uložen do souboru s koncovkou ∗.aln (formát CLUSTAL), případné další formáty je třeba zvolit pod *Alignment → Output Format Options*

Pozn. 1: Alignmenty vytvořené pomocí Clustal algoritmu však mohou být horší a mohou vyžadovat rozsáhlejší manuální úpravy (zejména pokud jsou sekvence více rozdílné nebo je v alignmentu větší množství indelů).

Pozn. 2: Tentýž algoritmus (zvaný ClustalW, tj. program bez grafického rozhraní) je obsažen i přímo v programu BioEdit (viz bod V). V tomto programu si označíme všechny sekvence (Ctrl+A), zmáčkneme *Accessory Application → ClustalW Multiple alignment → Run ClustalW*. Výsledný alignment se otevře rovnou v novém BioEditovém okně.

Mezi další servery, které umožňují dělat alignment patří např. GenomeNet (http://www.genome.jp/tools/clustalw – lze přepínat mezi Clustal, MAFFT a PRRN).

Troubleshooting

Některá sekvence je extrémně odlišná, není možné ji alignovat s ostatními.

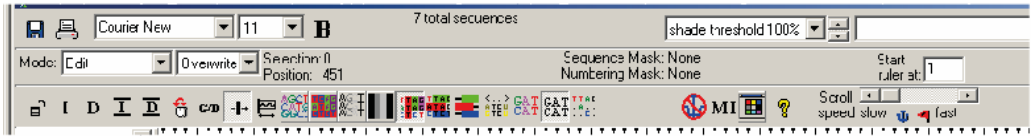
- jedná se o *reverse complement* (prohození forward a reverse primeru) – možné automaticky změnit např. v programu BioEdit (označit příslušnou sekvenci klepnutím na její název a pak příkaz Ctrl+Shift+R). K jednoduché záměně sekvence na *reverse complement* sekvenci lze také použít program RC (http://www.famd.me.uk/AGL/RC.zip)

- jedná se o sekvenci jiného úseku / jiného organismu (viz bod II)

V. Automaticky vytvořený **alignment** bychom měli přinejmenším zkontrolovat a případně **manuálně upravit** buď přímo v textovém editoru (např. v poznámkovém bloku) nebo využít výhod programu **BioEdit** (http://www.mbio.ncsu.edu/BioEdit/bioedit.html)

Jako vstupní formát pro BioEdit lze použít formát CLUSTAL nebo alignované sekvence ve FASTA formátu.

- *File → Open*
- vybrat soubor s koncovkou ∗. aln nebo ∗.fas (např. výstup z analýzy MAFFT)
- v rozbalovacím okně „*Mode*“ nastavíme režim „*Edit*“, který umožňuje manipulaci se sekvencemi (přepisování, mazání bazí – lze přepínat mezi mody *Insert* a *Overwrite* v druhém rozbalovacím okně)
- pro zvýraznění rozdílů v sekvencích lze využít široké škály možností podbarvení (např. dle míry variability sekvencí apod.) pomocí ikon v hlavní nabídkové liště a rozbalovacího okna „*Shade treshold*“ nahoře



- editace alignmentu je intuitivní – přepisování bazí, kopírovnání úseků, vkládání gapů (pomocí pravého tlačítka)
- alignment můžeme uložit jako editovaný FASTA soubor ∗.fas, nebo vybrat jiný formát z nabídky dle potřeby
- pomocí *File* → *Export* → *Sequence alignment* můžeme sekvence uložit i jako soubor NEXUS (s koncovkou ∗.nex, pouze typ *interleave*), který lze přímo použít pro analýzu v programu PAUP (viz bod X), MrBayes (http://mrbayes.csit.fsu.edu) nebo SplitsTree (viz bod IX). Pozor, BioEdit trpí řadou drobných chyb, např. ukládání NEXUS souborů může být problém pokud je název souboru či jeho cesty příliš složitý – v takovém případě je třeba soubor nazvat jednodušeji.
- užitečné zkratky: Ctrl+A (označení všech sekvencí), Ctrl+Del (smazání označené sekvence), Ctrl+Shift+R (vytvoření *reverse complement*)

Může se také hodit...

- pokud je editována sekvence obsahující **kódující i nekódující úseky genů**, může být užitečné takové úseky odlišit, např. pomocí velkých (pro kódující) a malých písmen (pro nekódující DNA). Označte myší požadovaný úsek a v hlavním menu vyberte *Sequences* → *Manipulations* → *UPPERCASE/lowercase*
- pokud jsou editovány pouze **kódující sekvence** genu (např. exony) ve správném čtecím rámci, lze využít překladač z DNA **do sekvencí aminokyselin** (Ctr+T) a odhadnout tak (ne)přítomnost pseudogenů.

Alternativně lze použít jednoduchý editor **PhyDE** (http://www.phyde.de), kde je třeba mít alignment ve formátu FASTA nebo NEXUS sequential.

- *File* → *Open*
- vybrat soubor s koncovkou ∗.fas
- otevře se okno s alignmentem a barevně označenými bázemi
- případné přesuny bází nebo jejich bloků můžeme udělat pomocí myši – označíme při současném stisku levého tlačítka, klikneme pravým, přeneseme na jiné místo, opět klikneme pravým. Módy editace jsou *Align* (lze pouze posouvat gapy) nebo *Edit* (pozor! zde lze sekvence i editovat).
- *File* → *Export as…* → *NEXUS* (nebo FASTA, dle libosti…)

VI. **FaBox – manipulace se sekvencemi ve formátu FASTA** (http://users-birc.au.dk/biopv/php/fabox/)

- pomocí tohoto webu lze manipulovat se sekvencemi ve FASTA formátu (pracovat s hlavičkami, ořezávat a spojovat alignmenty, formátovat pro další programy)
- lze také extrahovat pouze variabilní pozice alignmentu („*Show variable sites only*“)
- volba „*Create TCS input file from fasta (fasta2tcs)*“ vytvoří správně formátovaný soubor pro analýzu v programu TCS
- volba „*Create MrBayes input file from fasta (fasta2mrbayes)*“ vytvoří NEXUS soubor včetně sekce s příkazy pro program MrBayes (v závěru souboru; začíná „*begin mrbayes;*“ a končí „*end;*“). Smazáním této sekce získáme standardní NEXUS soubor použitelný i do jiných aplikací (např. PAUP ad.).

VII. Různé formáty sekvencí (CLUSTAL, FASTA, NEXUS ad.) můžeme **konvertovat** pomocí různých webových aplikací, jako např.

- **Sequence conversion** (http://sequenceconversion.bugaco.com/converter/biology/sequences)
- **Format Converter v2.0.5** (http://hcv.lanl.gov/content/sequence/FORMAT_CONVERSION/form.html)

Alternativou k webovým aplikacím je programu **FORCON** (http://bioinformatics.psb.ugent.be/webtools/ForCon), který však funguje pouze pod WinXP.

- *Input format*: CLUSTAL, *Output format*: FASTA, *OK*
- změnit typ souboru na „všechny“, otevřít soubor alignmentu (s koncovkou ∗.aln)
- *Cut-off after how many characters ?* – napsat více než je pozic alignmentu
- uložit s koncovkou ∗.fas
- *Select All*, *OK*

VIII. Pokud alignment obsahuje **mezery (gaps),** tedy inserce/delece (souhrnně indely), které mohou být fylogeneticky informativní, je potřeba je okódovat jako zvláštní znaky. To lze automaticky provést např. v programu **SeqState** (http://bioinfweb.info/Software/SeqState). Podrobnosti ohledně způsobu **kódování indelů** lze nalézt např. v článku Müller (2006): Incorporating information from length-mutational events into phylogenetic analysis. *Molecular Phylogenetics and Evolution*, 38, 667-676.

- *File* → *Load NEXUS file* (nebo *Load FASTA file*)
- vybrat soubor s koncovkou ∗.nex (nesmí být v interleave, ale v sequential formátu, tj. vytvořený např. v programu PhyDE nebo přeformátovaný pomocí webové služby, viz bod VII), resp. ∗.fas
- *IndelCoder* → *simple indel coding* (bližší informace o způsobu kódování viz zmiňovaný článek)
- vygeneruje NEXUS soubor (interleave – k názvu původního souboru přidá _mcic.nex) s původní maticí a přidaným kódováním indelů, který lze přímo použít pro analýzu např. v programu PAUP

IX. **(Pre-)analýza sekvencí v programu SplitsTree** (http://www.splitstree.org/). Sestrojením sítě např. algoritmem *neighbour-net* lze vizualizovat konfliktní signály v datasetu. Díky tomu ji lze využít např. pro detekci hybridních jedinců 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.

- 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*).

X. Testování modelů evoluce DNA. Vhodně zvolený model evoluce DNA je klíčový při výpočtech věrohodností topologií fylogenetických stromů pomocí pravděpodobnostních metod (např. Maximum likelihood nebo Bayesovská analýza).

Před vlastním testováním modelů, které nejlépe vystihují evoluci DNA v našich datech, je dobré zjistit **strukturu studovaného úseku**, tj. jakou část úseku tvoří např. kódující sekvence, nekódující sekvence apod. Hranice těchto úseků odvozujeme proto, že každý z nich se potenciálně mohl vyvíjet jiným způsobem (jiné mutační rychlosti apod.) a mohl by být charakterizován jiným evolučním modelem. Informace o studovaném úseku nalezneme např. pomocí on-line aplikace **BLAST** (viz výše, odst. II.), kdy hledáme nejpodobnější sekvenci k námi analyzované, která je anotována (tedy je zde uvedena struktura DNA úseku, viz. obr. níže).

Pokud jsme potřebné informace zjistili, můžeme **dataset manuálně rozdělit** např. na kódující část a nekódující část, tj. vytvořit z jednoho fasta souboru dva a více jiných fasta souborů (vhodně označených jako „....exon.fas“, „...intron.fas“ apod). Rozdělení původního souboru můžeme provést např. v programu BioEdit odstraněním nežádoucích částí alignmentu. Ze vzniklých souborů vytvoříme formát NEXUS, který použijeme při vlastním testování modelů.

Alternativním způsobem, jak rozdělit původní dataset na jednotlivé úseky, je připojit na konec souboru NEXUS (nebo i do separátního souboru) několik příkazů, které definují rozdělení alignmentu na tzv. **partition**. Způsob definování jednotlivých úseků vždy raději konzultujte s dokumentací k programu, pro který rozdělený dataset potřebujete (např. PAUP, MrBayes, PartitionFInder apod..), mohou se totiz více či méně lišit. Příkladový Nexus blok pro rozdělení kódujících sekvencí podle pozice nukleotidů v tripletu pro účely Baysovské analýzy viz níže.

```
begin mrbayes; # začátek nového nexus bloku (může být např. i "Begin sets;")
outgroup 22;
charset CHS_pos1 = 1-957\3; # definice úseků - kódující sekvence, první nukleotidy v tripletu
charset CHS_pos2 = 2-957\3; # kódující sekvence, druhé nukleotidy v tripletu
charset CHS_pos3 = 3-957\3; # kódující sekvence, třetí nukleotidy v tripletu
partition Position = 3: CHS_pos1, CHS_pos2, CHS_pos3; # rozdělení datasetu
end;
```

Testování modelů pomocí **jModeltest** (http://code.google.com/p/jmodeltest2/) **pro následnou analýzu pomocí ML**. Program jModeltest nepřijímá datasety rozdělené pomocí partitions, takže je nutné analyzovat dataset jako celek, případně analyzovat separátně jednotlivé jeho části (tj. zvlášť exony, zvlášť introny apod...).

- *File → Load DNA alignment* # vyberte soubor ve formátu Phylip nebo Nexus

- *Analysis → Compute likelihood scores →* nasatvení komplexity analýzy (od 3 do 88 modelů) – pokud budeme data analyzovat dále pomocí ML, nastavíme maximální počet modelů (tj. nastavíme „Number of substitution rates“ = 11 a zaskrneme všechna okénka v „Base frequencies“ a „Rate variation“. Ostatní můžeme nechat jako default. Pokud budeme analyzovat data pomocí programu MrBayes, stačí nastavit komplexitu nižší, neboť v programu lze nastavit jen několik specifických modelů. Pro testování modelů pro MrBayes lze využít i specifického programu MrModeltest, který testuje právě jen modely, které lze v této analýze použít.
- *Analysis → Do AIC calculation →* ponecháme defaultní nastavení a zaškrtneme „write *PAUP block”
- Program zobrazí výsledky porovnání jednotlivých modelů a jejich likelihood scores (věrohodnosti jednotlivých modelů). V sekci „AKAIKE INFORMATION CRITERION (AIC)“ nalezneme nejoptimálnější model pro naše data vyhodnocený podle kriteria AIC. Níže zkopírujeme soubor příkazů, kódující použití vybraného modelu pro naše data např. pro analýzu ML, a vložíme je do Nexus souboru s původním alignmentem.

```
--
PAUP* Commands Block:
 If you want to load the selected model and associated estimates in PAUP*,
 attach the next block of commands after the data in your PAUP file:
[!
Likelihood settings from best-fit model (HKY+G) selected by AIC
with jModeltest 0.1.1 on Thu Oct 17 13:26:26 CEST 2013]

BEGIN PAUP;
Lset base=(0.4010 0.2062 0.1059 0.2870) nst=2 tratio=3.3506 rates=gamma shape=0.4330 ncat=4
pinvar=0;
END;
--
```

Testování modelů pomocí **PartitionFinder** (http://www.robertlanfear.com/partitionfinder/) **pro následnou analýzu pomocí ML**. Tato metoda má oproti výše zmíněnému jModeltestu výhodu ve schopnosti analyzovat jediný vstupní dataset, který může být rozdělen na několik úseků pomocí definice partitions. Program otestuje všechny modely na všech partitions a následně rozhodne, které partition si zaslouží specifický model, případně, které partitions je možné spojit apod. To je výhodné zejména v situaci, kdy chceme zjistit, jestli je v daném úseků potřeba rozlišovat rychlost evoluce nejen mezi kódujícími a nekódujícími úseky, ale i např. v rámci kódujících úseků na jednotlivých pozicích v tripletu. Takové rozdělení alignmentu končí s více než třemi partitions a není příliš praktické takové členění dělat manuálně. PartitionFinder je skript napsaný v jazyce Python a pro jeho použití je třeba mít prostředí Python předem nainstalované (viz http://www.python.org/).

- Vstupními soubory, které by měly být uloženy spolu v jedné složce, je alignment sekvencí ve formáty Phylip a soubor „partition_finder.cfg“, ve kterém je uveden název souboru obsahujícího data (*.phy), zadefinováno rozdělení vztupního datasetu na partition a specifikace vlastního testování modelů. Specifikace vlastní analýzy a editaci souboru *.cfg je dobré konzultovat s dokumentací k programu. Defaultní nastavení ale může vypadat např. takto:

```
## ALIGNMENT FILE ##
alignment = test.phy;

## BRANCHLENGTHS: linked | unlinked ##
branchlengths = linked;

## MODELS OF EVOLUTION for PartitionFinder: all | raxml | mrbayes | beast | <list> ##
##              for PartitionFinderProtein: all_protein | <list> ##
models = all;
```

```
# MODEL SELECCTION: AIC | AICc | BIC #
model_selection = BIC;

## DATA BLOCKS: see manual for how to define ##
[data_blocks]
Gene1_pos1 = 1-1062\3;
Gene1_pos2 = 2-1062\3;
Gene1_pos3 = 3-1062\3;

## SCHEMES, search: all | greedy | rcluster | hcluster | user ##
[schemes]
search = greedy;

#user schemes go here if search=user. See manual for how to define.#
```

- Vlastní analýzu spustíme pomocí příkazové řádky, kterou přivoláme ve Windows např. pomocí Start → Run → cmd
- Do příkazové řádky vepíšeme příkaz se strukturou
  python ”<PartitionFinder.py>” “<inputFolderName>”
  kde ”<PartitionFinder.py>” představuje cestu ke skriptu PartitionFinder.py (může být kdekoliv v počítači) a “<inputFolderName>” představuje cestu ke složce se vstupními formáty (take umístěnými kdekoliv). Obě cesty můžeme dostat do příkazové řádky za příkaz “python” pouhým přetažením ze složek kde jsou umístěny.

- Po skončení analýzy se vytvoří v pracovní složce nova složka s názvem “analysis”. V této složce nás bude nejvíce zajímat soubor “best_schemes.txt”, kde nalezneme informace o nejlepší variantě poskládání jednotlivých partitions a do skupin se společným evolučním modelem.

```
Subset | Best Model | Subset Partitions| Subset Sites | Alignment
1 | F81+G | Gene1_pos1, Gene2_pos1, Gene3_pos1 | 1-789\3, 790-1449\3, 1450-2208\3 | .....
2 | F81+G | Gene1_pos2, Gene2_pos2, Gene3_pos2 | 2-789\3, 791-1449\3, 1451-2208\3 | .....
3 | K80+G | Gene1_pos3 | 3-789\3 | .....
4 | TrN+G | Gene2_pos3, Gene3_pos3 | 792-1449\3, 1452-2208\3 | ....

Scheme Description in PartitionFinder format
Scheme_step_5 = (Gene1_pos1, Gene2_pos1, Gene3_pos1) (Gene1_pos2, Gene2_pos2, Gene3_pos2)
(Gene1_pos3) (Gene2_pos3, Gene3_pos3);
```

- Program však negeneruje jednotný blok pro Nexus soubor, který bychom mohli vložit za náš alignment. Je nutné si vytvořit blok vlastní se specifikacemi modelů pro jednotlivé skupiny partitions.

XI. Testování fylogenetické struktury v datech (TreePuzle, PAUP, Excel)

Pomocí těchto analýz zjišťujeme, jaká je míra fylogenetické informace a šumu v datech.

Likelihood mapping

- programy:
  - Tree Puzzle: tree-puzzle-5.2-windows.zip
- zdrojové alignmenty (fylogeneze druhů krásivkového rodu *Micrasterias*):
  - SSU: Micrasterias_SSU.phy
  - psaA: Micrasterias_psa.phy
  - coxIII: Micrasterias_cox.phy
- instalace programu:
  - odzipovat
  - v adresáři „src“ přejmenovat „puzzle-windows-mingw-static.exe“ na „puzzle.exe“
  - nastavení cesty k programu (program lze pak pouštět z jakékoli složky):

    - kliknout pravou myší na „můj počítač“ – vybrat „vlastnosti“
    - kliknout na „upřesnit nastavení systému“
    - na kartě „Upřesnit“ kliknout na „Proměnné prostředí...“¨
    - v okně „Systémové proměnné vybrat „Path“, kliknout na „Upravit...“
    - na začátku řádku „Hodnota proměnné“ přidat cestu k programu Tree puzzle (např. C:\tree-puzzle\src;). Cesta musí končit středníkem!
    - OK
- v adresáři se zdrojovými daty spustit program (napsat do příkazového řádku „puzzle.exe“)
- vybrat alignment (např. „Micrasterias_SSU.phy“)
- nastavit analýzu: „b“ – likelihood mapping, ostatní parametry nechat defaultně nastavené
- nastavit analýzu: „m“ – vybrání správného modelu evoluce. Program pracuje dobře s HKY modelem. Pokud by člověk chtěl nastavit GTR model, pak se musí zadat jednotlivé substituční rychlosti podle výstupu Model testu
- spustit analýzu: „y“
- výstupy programu:
    - outfile: zpráva o průběhu a výsledcích analýzy. Úplně dole je pak procentuální zastoupení resolved a unresolved quartets
    - outdist: ML distance všech dvojic sekvencí
    - outlm.eps: grafický výstup analýzy

g_1 statistika

- programy:
  - PAUP
  - Excel či obdobný tabulkový program
- zdrojové alignmenty (fylogeneze druhů krásivkového rodu *Micrasterias*):
  - SSU: Micrasterias_SSU.nex
  - psaA: Micrasterias_psa.nex
  - coxIII: Micrasterias_cox.nex
- instalace programu PAUP:
  - viz instalační CD
  - nastavit cestu k programu (viz výše)
- v adresáři se zdrojovými daty spustit program PAUP (paup.exe)
- načíst alignment – např. „exe Micrasterias_SSU.nex
- generovat 1000 náhodných stromů
  - generatetrees random ntrees=10000;
  - pscore/scorefile=ACT_exon_scores;
  - end;
- nakreslení histogramu v R:
  - my_data<-read.table("ACT_exon_scores",header = TRUE)
  - tree_lengths<-my_data$Length
  - hist(tree_lengths)
- nakreslení histogramu v Excelu:
  - vygenerované délky stromů importovat do Excelu (oddělovač: tabulátor)
  - nainstalovat doplněk Analytické nástroje: Soubor – Možnosti – Doplňky - Spravovat: Doplňky aplikace Excel – Přejít... – zaškrtnout „Analytické nástroje“ – OK
  - Podle minimální a maximální hodnoty délek stromů vygenerovat sloupeček s hranicemi tříd (např. čísla 100, 120, 140, 160, 180, ...)
  - Vytvoření histogramu: Data – Analýza dat – Histogram – vybrat sloupečky s délkami stromů a hranicemi gtříd. Zaškrtnout: Vytvořit graf
- vypočtení hodnoty g_1 v R
  - nahrát package „e1071“ – Install package...., Load package...
  - skewness(tree_lengths)
- vypočtení hodnoty g_1 v Excelu
  - spočítat průměr a směrodatnou odchylku délek stromů
  - pro každou délku stromů spočítat třetí mocninu jejího rozdílu s průměrnou délkou
  - spočítat sumu všech třetích mocnin
  - g_1 = (suma mocnin)/(1000*(směrodatná odchylka délek)^3)

$$\frac{\sum_{i=1}^{n} (T_i - \bar{T})^3}{n\, s^3}$$

where n is the number of trees of length T and s is the standard deviation of tree lengths. For a perfectly symmetrical tree-length distribution $g_1 = 0$, whereas a left-skewed distribution has a $g_1 < 0$ and a right-skewed distribution has a $g_1 > 0$.

XII. Testování substituční saturace sekvencí

Pomocí těchto analýz určujeme míru šumu v datech způsobenou substituční saturací.

Saturační křivky

- programy:
  - PAUP
  - Excel
- zdrojové alignmenty (fylogeneze třídy Chrysophyceae – zlativky):
  - rbcL – 1. pozice kodónu: chryso_rbcl1.nex
  - rbcL – 2. pozice kodónu: chryso_rbcl2.nex
  - rbcL – 3. pozice kodónu: chryso_rbcl3.nex
- pro každý allignment vytvořit dva Nexus soubory, pro výpočet uncorrected a corrected distancí. Na konec Nexus souborů připojit PAUP bloky:
  - PAUP blok pro výpočet uncorrected distancí:
    - begin paup;
    - dset distance=p;
    - savedist format=onecolumn file=uncorrected_distances undefined=asterisk;
    - end;
  - PAUP blok pro výpočet corrected distancí (hodnoty „dset“ uvedené kurzívou vybrat či zadat podle výstupu z Model testu):
    - begin paup;
    - dset distance=***JC/F81/HKY85/GTR*** rates=***equal/gamma*** shape = ***0.0000*** pinvar = ***0.0000***;
    - savedist format=onecolumn file=corrected_distances undefined=asterisk;
    - end;
- oba alignmenty otevřít a zanalyzovat v programu PAUP. Po každé analýze je nutné PAUP zavřít, protože si pamatuje nastavení modelů!
  - paup ***název_souboru***
  - výstupem analýzy je soubor obsahující distance
- graf v Excelu: distance importovat do Excelu, vytvořit XY graf
- graf v R:
  - uncorrected<-read.table("***uncorrected_distances***", sep= "\t")
  - uncorrected_dist<-subset(uncorrected, select=V3)
  - corrected<-read.table("***corrected_distances***", sep= "\t")
  - corrected_dist<-subset(corrected, select=V3)
  - distances<-data.frame(corrected_dist,uncorrected_dist)
  - plot(distances, xlab="corrected distances", ylab="uncorrected distances")
  - abline(0, 1,col = "red")

Site stripping

- programy:
  - HyPhy
  - MEGA
  - Perl
  - SiteStripper
- zdrojové alignmenty (fylogeneze třídy Chrysophyceae – zlativky):
  - rbcL – 1. pozice kodónu: chryso_rbcl1.fas
  - rbcL – 2. pozice kodónu: chryso_rbcl2.fas
  - rbcL – 3. pozice kodónu: chryso_rbcl3.fas
- instalace programu HyPhy:
  - viz http://hyphy.org/w/index.php/Main_Page
- instalace programu MEGA:
  - viz http://www.megasoftware.net/
- instalace prostředí Perl (Active Perl):
  - viz http://www.activestate.com/activeperl
  - nainstalovat program (v rámci instalace povolit přidání cesty do PATH)
  - spustit Perl Package Manager
  - nainstalovat Bioperl ze stejného manageru – návod viz „GUI Installation“ na stránce:http://www.bioperl.org/wiki/Installing_Bioperl_on_Windows
    - tj., nainstalovat jednotlivé repositories podle doporučených odazů pro Perl 5.10, některé z nich lze najít v „Suggested“
    - nainstalovat Bioperl
  - v Perl Package Manager dále nainstalovat package „Sort-Array“ (stejně jako při instalaci package Bioperl. Pokud se objeví hláška „Authorization denied“, pak nainstalovat druhou nalezenou package stejného jména
- instalace programu SiteStripper:
  - stáhnout si package sitestripper.pl - viz http://www.phycoweb.net/software/SiteStripper/index.html
- tvorba rychlého ML stromu v MEGA:
  - otevřít Fastaalignment v programu MEGA
  - Data – Export alignment – MEGA Format
  - otevřít MEGA alignment
  - Phylogeny – Construct/Test Maximum LikelihoodTree:
    - Test ofPhylogeny – none
    - Substitution Model &RatesamongSites – vybrat podle Model Testu
    - Compute
  - Po vypočtení stromu: File – Export CurrentTree (Newick)
- spočtení substitučních rychlostí v programy HyPhy
  - spustit HyPhy, Standart analysis = SubstitutionRates (SiteRates.bf)
  - vybrat alignment, vybrat optimální substituční model (podle Model Testu), Model Options: „Global“

    - vybrat ML strom vypočtený v programu MEGA
    - po výpočtu program vyzve k uložení substitučních rychlostí
- sitestripping
    - zkopírovat Perl skript „sitestripper.pl“ do adresáře s alignmentem a souborem substitučních rychlostí
    - spustit Perl skript: „perl sitestripper.pl -a -r -f -o“, kde
        - -a = fastaalignment sekvencí
        - -r = soubor se substitučními rychlostmi
        - -f = procento ponechaných bází
        - -o = výsledný alignment
    - např.: „perl sitestripper.pl -a ***chryso_rbcl1.fas*** -r ***rbcl1rates.txt*** -f ***.90*** -o ***rbcl1out.nex***“
    - procento ponechaných bází je ideálně určeno na základě saturačních křivek. To lze zobrazit a spočítat v R takto:
        - corrected_sorted<-sort(corrected_dist$V3, decreasing = FALSE)
        - strip<-which.min(abs(corrected_sorted - ***0.6***))
        - stripping<-strip/(length(corrected_sorted)/100)
        - stripping
        - abline(v=***0.6***, col = "blue")