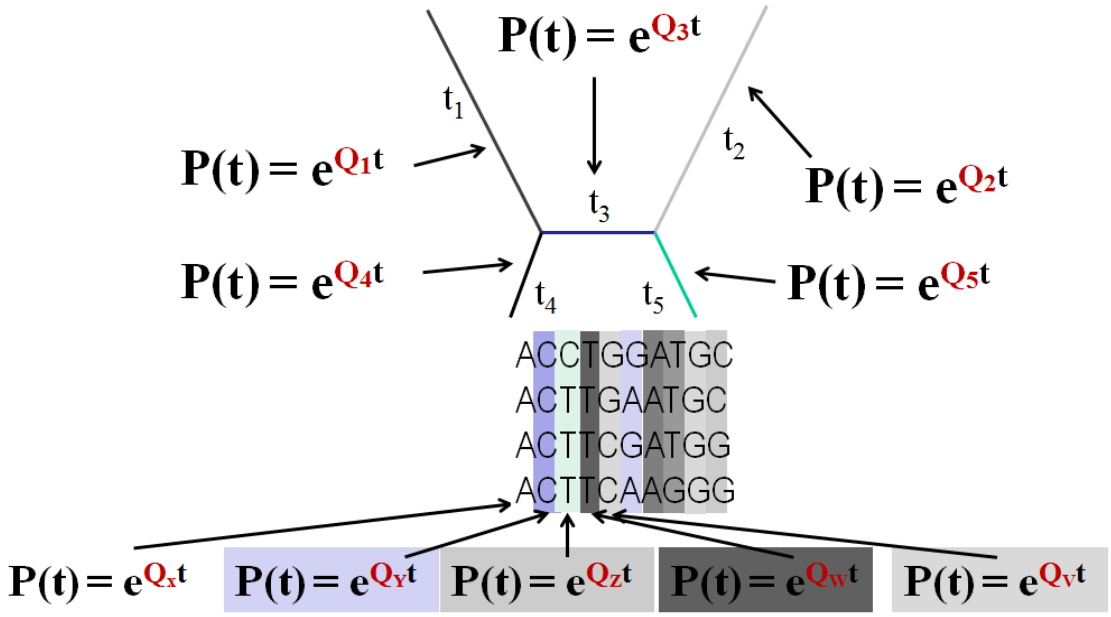
MOLEKULÁRNÍ TAXONOMIE – 9

Zdokonalování substitučního modelu

V předchozích přednáškách jsme si představili metodu maximum likelihood, která počítá s pravděpodobnostmi substitucí na větvích a topologiích. Díky pravděpodobnostním počtům tato metoda má snahu vyhnout se problému inkonzistence, kterým trpí metoda maximální parsimonie v případě, že se v mezi sekvencemi vyskytují takové, jejichž substituční rychlost vůči ostatním přesahuje určitou míru. Maximum likelihood je metodou konzistentní, pokud ovšem substituční model použitý při výpočtu pravděpodobností dokonale vystihuje substituční proces, kterými sekvence prošly. Pokud tomu tak není, tak Felsensteinova zóna stále existuje, ale je menší než v případě maximální parsimonie. I ten nejsložitější model, který jsme si dosud představili (GTR+Γ) nevystihuje veškeré nuance substitučního procesu. Co dále je ještě možné v modelu změnit, uvolnit, aby lépe pasoval na substituční proces, ukazuje obrázek níže.



Je rozumné předpokládat, že matice substitučních rychlostí neplatí univerzálně pro celou fylogenezi. Některé části stromu nebo dokonce každá větev by si jistě zasluhovaly vlastní pro ně specifické substituční matice Q, ze kterých by vycházely matice pravděpodobností záměn P(t) na míru šité jednotlivým větvím. V takovém případě mluvíme o heterotachyi. Podobně je rozumné předpokládat, že jednotlivé sloupce alignmentu neprochází substitucemi podle univerzální matice Q a že by bylo rozumné přidělit každému sloupci jeho vlastní matici Q.
V tomto případě mluvíme o „site-heterogenous“ substitučních modelech. Pokud bychom takto náš model uvolnili, tak by jistě velmi dobře „fitoval“ na substituční proces, ale dostali bychom se do nového problému – přeparametrizování. Pokud bychom chtěli tímto supervolným modelem modelovat substituční proces na alignmentu 10 sekvencí (strom 10 druhů obsahuje 17 větví) a dlouhých 1000 aminokyselin. Tak by se v substitučních maticích Q, z nichž každá obsahuje 190

členů, vyskytovalo celkem 190x17x1000, tj. více než 3 milióny parametrů. Již dříve jsem upozorňoval na to, že hodnoty parametrů, které můžeme získat, jsou vždy jen odhadem skutečné hodnoty, který je zatížen chybou. Pokud se v analýze sejde tak velké množství parametrů zatížených chybou, začne se metoda chovat nedobře. Naší snahou je proplout mezi dvěma zmíněnými nástrahami – Skyllou (model, který nevystihuje substituční proces) a Charybdou (přeparametrizování) a vyvíjet chytré modely, které relativně dobře vystihují substituční děje s málo parametry. Níže uvádím několik typů modelů, které se snaží některé ze zmíněných vlastností substitučního procesu uvažovat. První čtyři spadají do kategorie site-heterogenních, poslední (covarion) modeluje heterotachyi.

CAT model (Lartillot a Philippe 2004[1]) uvolňuje předpoklad, že pro všechny pozice platí jedna matice substitučních rychlostí. Nedovoluje sice, aby každá pozice měla svoji vlastní Q, ale rozdělí si pozice do několika kategorií. Počet těchto kategorií je proměnná, kterou si model také optimalizuje. Každá kategorie substituuje podle vlastní Q matice. Analýzy na reálných datech ukázaly, že na modelování subtitučního procesu jednoho proteinu je třeba 10-30 kategorií pozic. Model CAT v praxi funguje dobře, ale je výpočtetně náročný, byl implementován pouze do Bayéské metody a MCMC někdy trpí neschopností konvergovat do rovnovážného stavu.

LG4X model – tento model je implementován do maximum likelihood metody v programu RAxML a funguje podobně jako CAT s tím rozdílem, že rozděluje pozice jen do 4 kategorií s různými Q maticemi. Je tedy jednodušší.

C10, C20, C60 – tyto modely jsou implementovány do maximum likelihood metody v programu IQtree. Rozdělují pozice do 10, 20 a 60 kategorií, pro ty uvažují různé (empiricky odvozené) rovnovážné frekvence aminokyselin (nikoliv rychlostní matice). Tím snižují množství parametrů oproti modelu CAT.

PMSF – obdoba C modelů, ale rovnovážné frekvence si určuje sám z dat.

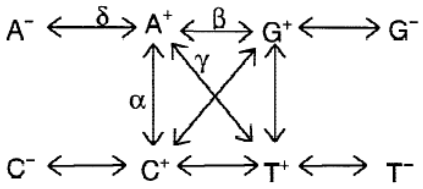
Model Covarion umožňuje modelovat proměnlivost substitučních rychlostí napříč stromem. Ve své nejjednodušší variantě (Penny a kol. 2001[2]) předpokládá, že každý nukleotid substituuje s rychlostí $\underline{\delta}$ na svého dvojníka, který se liší jedině v tom, že není schopen dalších substitucí. Rychlosti substituce takového nukleotidu nebo aminokyseliny na jiné jsou 0. Jediná možná substituce je zpětná substituce na svého dvojníka, který je schopný měnit se na jiná rezidua. Ke zpětné substituci dochází s rychlostí $\underline{\kappa\delta}$. Tento model potřebuje tedy jen dva parametry navíc. Substituční proces podle tohoto modelu je znázorněn níže.

[1] http://www.ncbi.nlm.nih.gov/pubmed/15014145
[2] http://www.ncbi.nlm.nih.gov/pubmed/11677631

A

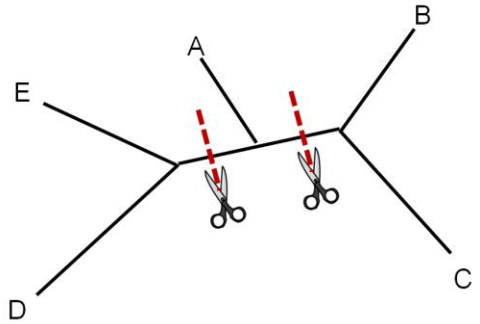
$$K' = \begin{array}{c} \\ A^+ \\ G^+ \\ C^+ \\ T^+ \\ A^- \\ G^- \\ C^- \\ T^- \end{array} \begin{bmatrix} * & \alpha & \beta & \gamma & \delta & 0 & 0 & 0 \\ \alpha & * & \gamma & \beta & 0 & \delta & 0 & 0 \\ \beta & \gamma & * & \alpha & 0 & 0 & \delta & 0 \\ \gamma & \beta & \alpha & * & 0 & 0 & 0 & \delta \\ k\delta & 0 & 0 & 0 & * & 0 & 0 & 0 \\ 0 & k\delta & 0 & 0 & 0 & * & 0 & 0 \\ 0 & 0 & k\delta & 0 & 0 & 0 & * & 0 \\ 0 & 0 & 0 & k\delta & 0 & 0 & 0 & * \end{bmatrix} \begin{array}{c} \\ A^+ \\ G^+ \\ C^+ \\ T^+ \\ A^- \\ G^- \\ C^- \\ T^- \end{array}$$

B

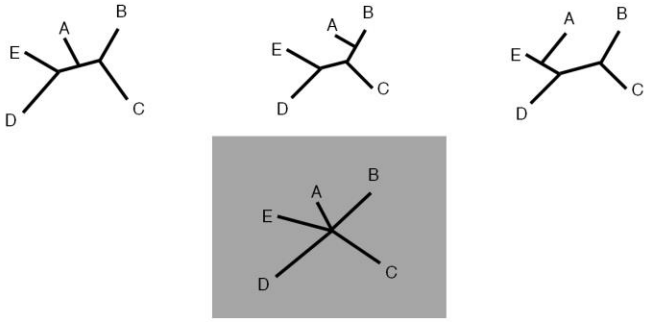


Konsenzus stromů

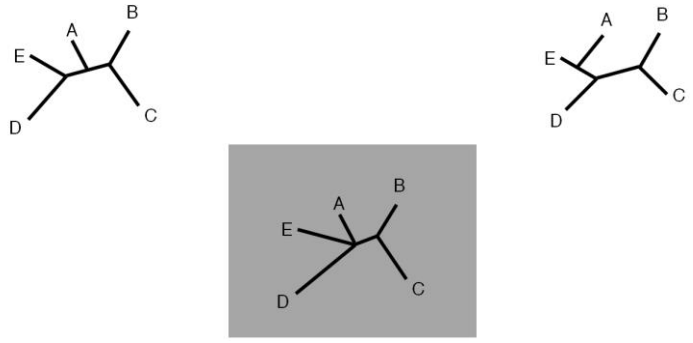
Někdy se ocitneme v situaci, že je potřeba kombinovat sadu topologií do jedné, která by shrnovala informaci obsaženu v této sadě topologií. Říkáme, že vytváříme konsenzuální strom. Pokud topologie obsahují stejnou nebo podobnou sadu OTU (operačně taxonomických jednotek), tak je lze smysluplně kombinovat. Lze postupovat různými způsoby. Při sestavování **striktně konsezuálního stromu** dbáme na to, aby výsledná topologie obsahovala jen ty "bipartitions" neboli "splits", které se vyskytují ve všech topologiích, které kombinujeme. Anglický termín **"bipartition" nebo "split"** označuje sady OTU, na které lze danou topologii rozdělit jedním řezem. Například topologie na obrázku níže se dá rozdělit na následující sady OTU: DE|ABC, DEA|BC. Nelze ji rodělit třeba na DAB|EC. Samozřejmě, že ji lze rozdělit také na část obsahující jednu OTU a zbytek (A|CEBD), ale to jde u všech možných topologií a nepřináší to žádnou informaci, takže to nebudeme řešit. Je důležité podotknout, že u splitu neřešíme topologii v rámci tohoto splitu, takže všechny splity sestávající z ABC jsou v tomto případě shodné, ať je jejich vnitřní topologie jakákoli.



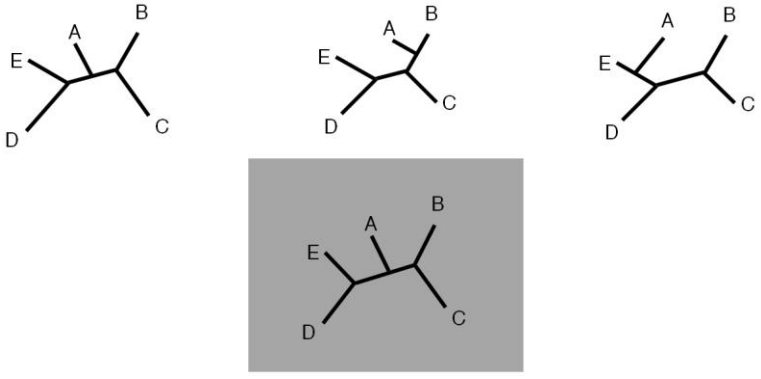
Striktní konsenzus tří níže uvedených topologií je uveden v rámečku pod nimi.



Protože tyto tři topologie neobsahují žádný společný split, je konsezuální topologie nerozlišená hvězdice, říkáme, že obsahuje polytomii neboli multifurkaci. Striktní konsenzus krajních topologií by již byl částečně rozlišený. Krajní topologie totiž obsahují společný split AED|BC. Ten bude proto přítomný v konsenzuální topologii, jak je ukázáno níže.

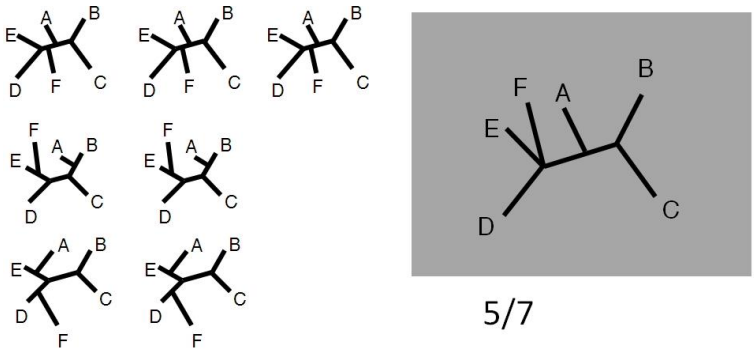


Někdy nechceme být tak přísní a přijmeme do konsenzuálního stromu takové splity, které se vyskytují v nadpoloviční většině stromů v sadě, kterou kombinujeme. Takovému konsenzu se říká **majority rule konsenzus**. V případě naší trojice stromů se jedná o splity AED|BC a ABC|ED. Ty se vyskytují ve ⅔ topologií a musí být tedy zastoupeny v konsenzu uvedeném v šedém rámečky. Splity přítomné v nadpoloviční většině topologií si z principu nemohou vzájemně odporovat, takže vždy lze sestavit takovýto konsenzus.

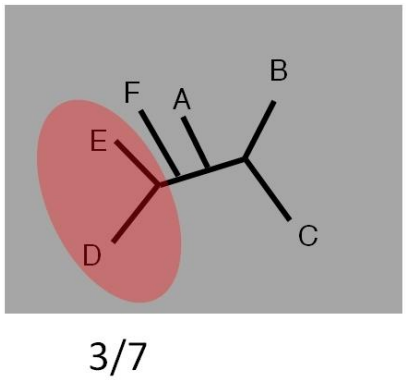


Je možné sestavovat i strom zohledňující splity přítomné ve většině, která však není nadpoloviční. Takovému postupu říkáme **extended majority rule**. V takovém případě je potřeba postupovat opatrněji. Nejprve sestavit majority rule konsenzus a poté rozlišit ty části stromu obsahující polytomie způsobem, který je v sadě kombinovaných topologií nejčastější.

Majority rule sady pěti topologií uvedených níže vypadá jako topologie v rámečku. Zohledňuje to, že splity BC|DEAF a ABC|DEF se vyskují v 5/7 stromů.



Pokud bychom chtěli rozlišit polytomii mezi DEF musíme se ještě podívat, jak je rozlišená tato část topologie v naší sadě. Split ED|FABC se vyskutuje ve 3/7 stromů, kdežto splity DF|EABC a EF|DABC se vyskytují jen ve 2/7 stromů, a proto zvolíme split ED|FABC.



Otázky, který bychom si měli klást

Při fylogenetických analýzách bychom si měli klást následující otázky:

- Podporují naše data (ve většině případů alignment) pevně nebo slabě příbuzenské vztahy na stromu, který jsme získali?

Všechny metody konstruující fylogenezi poskytnou jako výstup fylogenetický strom, a to bez ohledu na to, zda alignment podporuje topologii stromu pevně, tj. mnoho sloupců alignmentu vykazuje vzor znaků souhlasný s topologií, nebo slabě, tj. jen velmi málo sloupců podporuje výslednou topologii.

- Je můj strom skutečně lepší než nějaký jiný?

V určitých situacích je vhodné si ověřit, že zda je výsledný strom statisticky významně lepší než jiný strom. Často se do takové situace dostaneme v případě, když výsledná topologie nepodporuje existenci taxonu, který nás zajímá, protože jeho zástupci nevytváří monofyletickou

skupinu – klád. V takovém případě je třeba ověřit, zda jsou topologie, které existenci taxonu podporují, signifikantně horší či nikoli.

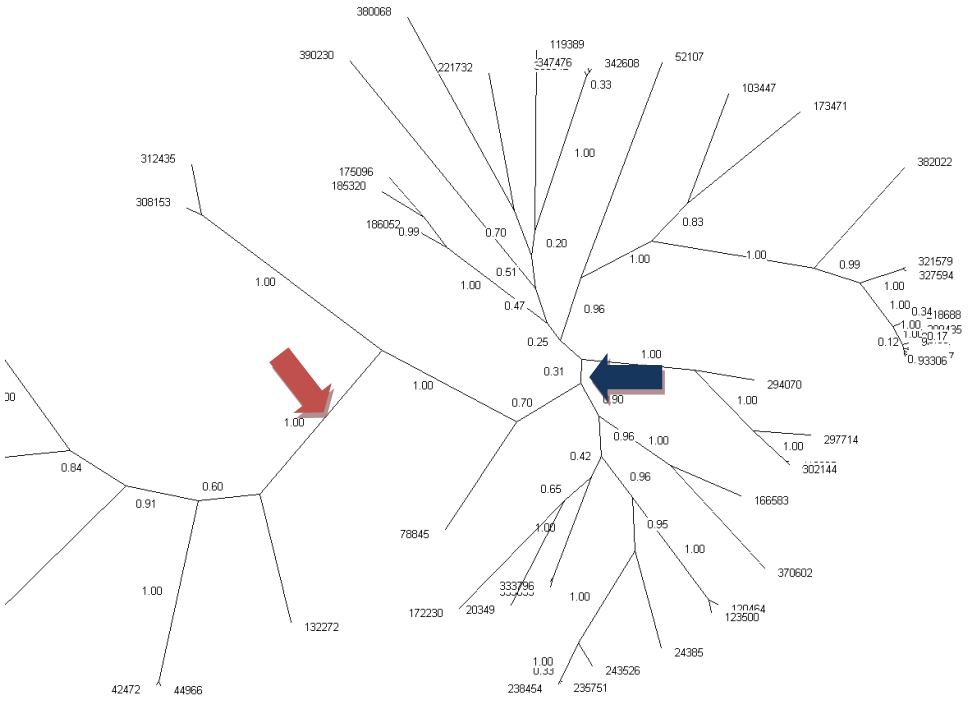
- Je vůbec vhodné vysvětlovat příbuzenské vztahy mezi našimi OTU pomocí stromu?

Všechny metody konstrukce stromů, které jsme si dosud představovali, konstruují dichotomicky se větvící stromy, protože jejich základní předpoklad je, že evoluce takto probíhá. To však nemusí být pravda. Sekvence, které analyzujeme, mohly v minulosti prodělat rekombinaci, tj. různé části genu mají různé předky. V takovém případě, by jejich evoluční minulost zachytil lépe síťový graf. Některé metody rekonstrukci fylogeneze toto umožňují.

Rekonstrukce fylogeneze může být navíc **ztěžována** přítomností **vysoké substituční saturace**, která fylogenetické vztahy maskuje, nebo naopak **malým množstvím fylogenetického signálu** (všechny sekvence téměř stejné). Data mohou navíc obsahovat **zavádějící signál (artefakt)** způsobený různorodým obsahem nukleotidů či aminokyselin nebo způsobený velmi odlišnou délkou větví.

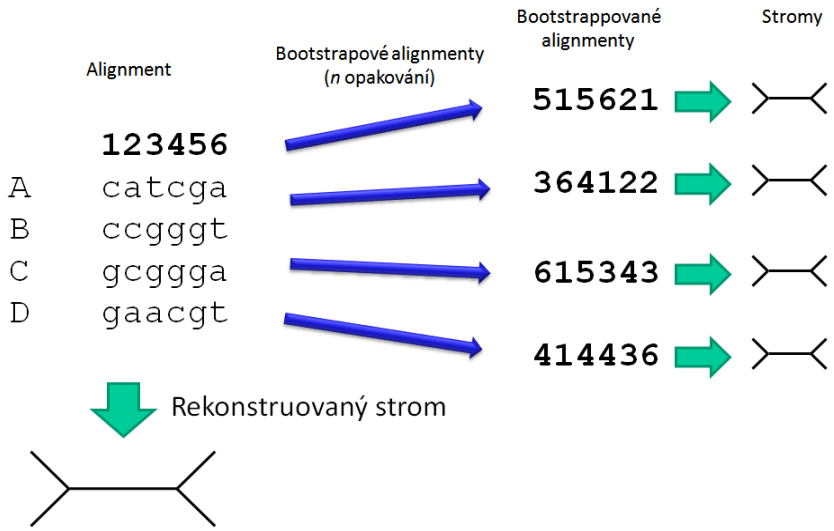
Statistická podpora větvení

Existuje několik způsobů, jak vyčíslit podporu větvení. Výstupem bayéské metody, která používá Marcov Chain Monte Carlo pro odhad posteriorní pravděpodobnosti topologie, jsou posteriorní pravděpodobnoti uzlů. Zopakujme si, že MCMC poté, co dosáhne rovnovážného stavu, navštěvuje opakovaně určitou omezenou skupinu stromů. Frekvence, s jakou strom navštíví, je odhadem jeho posteriorní pravděpodobnosti. Hlavním výstupem bayéské analýzy ovšem není topologie s nejvyšší posteriorní pravděpodobností, i když i tu ve výstupních souborech můžeme nalézt, ale konsenzuální strom vytvořený například metodou majority rule extended ze vzorku všech stromů navštívených v rovnovážném stavu. Tato topologie se vlastně mezi vzorky v rovnovážném stavu nemusí vůbec nacházet, ale je to konsensus vzorku „kvalitních“ topologií. Čísla na každém uzlu této topologie jsou posteriorní pravděpodobnosti „bipartitions“/splitů.



Jejich hodnoty udávají, v jakém procentu topologií v rovnovážném stavu se vyskytuje daný split. Hodnota 1,00 označená v obrázku červenou šipkou znamená, že všechny topologie obsahovaly tento split, tj. že všechny bylo možné rozdělit jedním řezem na část obsahující taxony napravo od šipky část obsahující taxony nalevo od šipky. Hodnota 0,31 na splitu označená modrou šipkou znamená, že tento split se vyskytoval jen na 31% topologií. Přitom vůbec nezáleží na tom, jakou vnitřní topologii měla jedna či druhá část splitu. Zdůrazňuji, že přestože se hodnoty posteriorních pravděpodobností (a totéž bude platit o bootstrapech a jackknifech) často píšou na uzly, jsou to hodnoty náležející ke splitům, tedy k vnitřním větvím. Pokud si to budeme uvědomovat, tak nás nezmatou různé způsoby znázornění stromů, jejich různá zakořenění a ohnutí.

Ostatní metody konstrukce stromu nám samy o sobě neposkytují takové hodnoty. Musíme si je dopočítat pomocí "resampling" metod (bootstraping nebo jakknifing). Základní princip těchto metod je vytvořit permutovaná vstupní data (100 – 1000x) z původního souboru dat (alignmentu). Tyto permutované soubory potom analyzovat a zkonstruovat z nich nové stromy. Z těchto stromů pak vytvoříme konsensus, který bude obsahovat na splitech hodnoty ukazující, jak často byl daný split přítomen v souboru stromů vytvořených s permutovaných dat. Získané hodnoty na splitech bychom pak měli přenést na strom vytvořený z původních dat.



Rozdíl mezi bootstrapingem a jackknifingem spočívá v tom, že v případě boostrapingu vytvoříme alignment permutacemi s opakováním, kdežto při jackkniffingu permutacemi bez opakování. To znamená, že při bootstrapingu vytváříme permutované alignmenty o stejné délce, jako měl původní a sloupce se v něm mohou opakovat, kdežto při jackkniffingu vytváříme alignmenty kratší, než byl původní, a sloupce se neopakují. Bootstraping se využívá v molekulární fylogenetice mnohem častěji. Hodnoty bootstrapu i jackkniffu pro tytéž uzly jsou v průměru nižší než posteriorní pravděpodobnosti vypočtené bayéskou metodou. Ani posteriorní pravděpodobnosti ani bootstrapy nemají vlastnosti p-value, tj. bootstrap 95 neznamená, že alternativní strom, který daný split neobsahuje, je možné zavrhnout na hladině pravděpodobnosti 5%. Existuje ovšem metoda (Susko 2010, Mol Biol Evol[3]), jak převádět BP na aBP (adjustedBP), které mají vlastnosti p-value. Simulace ukázala, že aBP jsou vyšší než BP. Bootstrap 80 odpovídá zhruba 95% a 90 odpovídá zhruba 98-99% (viz. tabulka níže).

Edge Lengths	Rate Variation	Data Type	BP								
			10	20	30	40	50	60	70	80	90
Random	Gamma	Nucleotide	25	41	53	64	73	80	87	93	97
Extreme	Gamma	Nucleotide	28	45	58	69	77	85	91	95	98
Extreme	Equal rates	Nucleotide	33	52	66	76	84	90	95	98	99
Extreme	Gamma	Amino acid	27	43	56	67	75	83	89	94	98
Extreme	Equal rates	Amino acid	29	46	60	70	79	86	92	96	99

Provádět standartní bootstraping je výpočetně náročné, protože analyzujete 100-1000x stejně velký dataset. Proto byly vyvinuty různé způsoby, které proceduru zrychlují za cenu nějakého zjednodušení (RELL, aLRT, RAxML rapid bootstrap, UFboot). Toho dosahují tím, že využívají již spočítané likelihoody pozic (site likelihoods) nebo likelihoody splitů, nebo nepočítají strom od začátku, ale pro každý uzel porovnají alternativy odvozené přesmykem pomocí NNI.

Testy topologických hypotéz

[3] http://www.ncbi.nlm.nih.gov/pubmed/20154180

Někdy se dostaneme do situace, že bychom chtěli vědět, zda naše data statisticky signifikantně zavrhují určitou fylogenetickou hypotézu, která nás zajímá. Typickým příkladem je situace, kdy náš strom nepodporuje existenci taxonu (na stromu se jeví taxon jako nemonofyletický). Než existenci tohoto taxonu vážně zpochybníme, měli bychom si ověřit, zda je fylogeneze podporující monofylii taxonu (nulová hypotéza, H_0) signifikantně horší. V rámci metody maximum likelihoodu je toto teoreticky možné a bylo navrženo hned několik různých druhů statistických testů pojmenovaných často podle jejich autorů - Kishino-Hasegawa (KH), Shimodaria-Hasegawa (SH), expected likelihood weight (ELW) a approximatelly unbiased (AU) test a pod. Ty se liší v různých více či méně podstatných detailech, ale jejich princip je podobný. Nyní si ve stručnosti představíme poslední zmíněný, který je v současnosti nejpoužívanější.

Nejprve si spočítáme rozdíl mezi likelihoodem "nejlepšího" stromu a testované (H_0) hypotézy. Tuto statistiku označme $\underline{\delta}$

$$\delta = LnL_1 - lnL_0$$ [4]

$\underline{\delta}$ bude vždy vyšší než nula, ale abychom zjistili, zda je rozdíl statisticky signifikantní, musíme znát rozložení statistiky δ. Bohužel pro nás, rozložení této statistiky nepřipomíná, žádnou používanou funkci, a proto nám nezbude, než si její rozložení nasimulovat následujícím způsobem. Pro obě hypotézy budeme permutovat (s opakováním) likelihoody pozic alignmentu ("site likelihoods"). Tedy něco jako bootstraping, ale přímo s likelihoody. Pro každou permutaci vypočteme statistiku $\underline{\delta_p}$ podle stejného vzorce, přičemž celkový likelihood získáme jako obvykle vynásobením likelihoodů pozic. Permutací provedeme mnoho (desetitisíce). Procento permutací, pro které platí

$$\delta_p >= \delta$$

představuje hodnotu $\underline{p}$ (statistickou významnost), s jakou můžeme H_0 zavrhnout.

Testy substitučních modelů

Jak vyplynulo z úvodů této přednášky, není úplně snadné najít "zlatou střední cestu" mezi příliš jednoduchým (a tedy nereálným) a přeparametrizovaným substitučním modelem. Naštěstí existují statistické postupy, jak takový model vybrat. Jednou z možností je použít indexy vyjadřující vhodnost modelu - AIC (Akaike infomation criterion) nebo BIC (Bayesian information criterion). Výpočet těchto indexů je uveden níže

$AIC_i = -2\ln L_i + 2p_i$
$BIC = -2\ln(L_i) + p_i \ln(n)$

[4] Proč neporovnáváme likelihoody, ale jejich logaritmy? Důvod je ryze praktický. Likelihood je vždy velmi malé desetinné číslo (vzniká násobením mnoha desetinných čísel). S jeho logaritmem, záporné rozumně veliké číslo, se mnohem lépe pracuje.

L_i ……. Likelihood hypotézy
p_i ……. Počet parametrů modelu
n ……. Počet pozic alignmentu

Oba porovnávají substituční modely podle výše likelihoodu, který nám poskytnou pro zvolenou (ideálně tu nejlepší) topologii a penalizují je za množství parametrů, které používají. BIC přihlíží navíc k počtu pozic v alignmentu. V obou případech volíme substituční model s nižším hodnotou indexu.

Další možností, jak porovnávat dvojici modelů je likelihood ratio test (LRT). V tomto případě spočítáme, podobně jako u topologických testů, statistiku $\underline{\delta}$

$$\delta = 2(\mathrm{Ln}L_1 - \mathrm{ln}L_0)$$

kdy L_0 je likelihood jednodužšího modelu (nulová hypotéza) a L_1 likelihood modelu složitějšího. Důležité je, aby L_0 byl obsažen v L_1, tj. aby byl jeho speciálním případem (např. GTR je speciálním případem GTR+Γ pokud $\alpha = \infty$). V takovém případě rozložení statistiky δ odpovídá rozložení χ^2 s počtem stupňů volnosti odpovídajícím rozdílu v počtu parametrů mezi porovnávanými modely. Signifikanci rozdílu pak odečteme ze statistických tabulek.