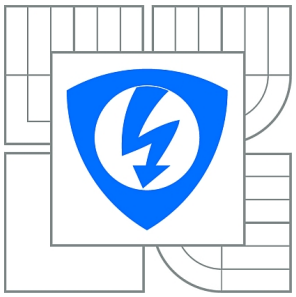
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGINEERING

STATISTICKÉ VYHODNOCENÍ FYLOGENEZE BIOLOGICKÝCH SEKVENCÍ

STATISTIC EVALUATION OF PHYLOGENY OF BIOLOGICAL SEQUENCES

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

AUTOR PRÁCE — Bc. ŠIMON VADJÁK

AUTHOR

VEDOUCÍ PRÁCE — Ing. HELENA ŠKUTKOVÁ

SUPERVISOR

BRNO 2014

VYSOKÉ UČENÍ
TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

Ústav biomedicínského inženýrství

Diplomová práce

magisterský navazující studijní obor
Biomedicínské inženýrství a bioinformatika

Student: Bc. Šimon Vadják
Ročník: 2

ID: 125085
Akademický rok: 2013/14

NÁZEV TÉMATU:

Statistické vyhodnocení fylogeneze biologických sekvencí

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši metodologie odhadu průběhu fylogeneze na základě podobnosti biologických sekvencí (DNA a bílkovin). Zaměřte se na vzniklé nepřesnosti v odhadu, čím jsou způsobeny a na možnosti jejich odstranění. 2) Proveďte srovnání metod pro statistické vyhodnocení správnosti průběhu fylogeneze. 3) Navrhněte a realizujte algoritmy pro testování správnosti fylogenetických stromů na základě resamplingových testů. 4) Vytvořte uživatelské programové rozhraní umožňující ze souboru biologických sekvencí ve FASTA kódu vykreslit fylogenetický strom metodou neighbor-joining s možností změny distančního modelu a substituční matice v programovém prostředí Matlab s Bioinformatickým toolboxem. 5) Program pro vykreslení fylogenetických stromů bude obsahovat jejich statistické vyhodnocení pomocí resampligových testů. 6) Proveďte srovnání realizovaných statistických metod na různých typech biologických sekvencí, posuďte jejich nedostatky a omezení.

DOPORUČENÁ LITERATURA:

[1] HOLMES, Susan. Bootstrapping Phylogenetic Trees: Theory and Methods. Statistical Science. 2003, roč. 18, č. 2, s. 241-255.
[2] NEI, Masatoshi a Sudhir KUMAR. Molecular Evolution and Phylogenetics. New York: Oxford University Press, 2000. ISBN 0-19513585-7.

Termín zadání: 10.2.2014 ***Termín odevzdání:*** 23.5.2014

Vedoucí práce: Ing. Helena Škutková
Konzultanti diplomové práce:

prof. Ing. Ivo Provazník, Ph.D.
předseda oborové rady

UPOZORNĚNÍ:

ABSTRAKT

Diplomová práce poskytuje ucelený přehled resamplingových metod pro testování správnosti topologie fylogenetických stromů odhadujících průběh fylogeneze na základě podobnosti biologických sekvencí. Důraz byl kladen také na možnosti vzniku nepřesností v tomto odhadu a na možnosti jejich odstranění a odhalení. Tyto metody byly realizovány v programovém prostředí Matlab pro Bootstrapping, Jackknifing, OTU jackknifing a PTP test (permutation tail probability). Práce si klade za cíl otestovat jejich použitelnost na různých biologických sekvencích a také posoudit vliv volby vstupních parametrů analýzy na výsledky těchto statistických testů.

KLÍČOVÁ SLOVA

Fylogenetický strom, substituční matice, skórovací matice, distanční matice, evoluční model, taxon, neighbor joining, bootstrapping, jackknifing, PTP test, permutation tail probability, OTU jackknifing, resamplingové testy.

ABSTRACT

The master's thesis provides a comprehensive overview of resampling methods for testing the correctness topology of the phylogenetic trees which estimate the process of phylogeny on the bases of biological sequences similarity. We focused on the possibility of errors creation in this estimate and the possibility of their removal and detection. These methods were implemented in Matlab for Bootstrapping, jackknifing, OTU jackknifing and PTP test (Permutation tail probability). The work aims to test their applicability to various biological sequences and also to assess the impact of the choice of input analysis parameters on the results of these statistical tests.

KEYWORDS

Phylogenetic tree, substitution matrix, scoring matrix, distance matrix, evolutionary model, taxon, neighbor joining, bootstrapping, jackknifing, PTP test, permutation tail probability, OTU jackknifing, resampling tests

BIBLIOGRAFICKÁ CITACE:

VADJÁK, Š. *Statistické vyhodnocení fylogeneze biologických sekvencí*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2014. 70 s. Vedoucí diplomové práce Ing. Helena Škutková.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci, na téma statistické vyhodnocení fylogeneze biologických sekvencí, jsem vypracoval samostatně pod vedením vedoucí diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením tohoto projektu jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

V Brně dne 23. května 2014 .....
podpis autora

PODĚKOVÁNÍ

Děkuji vedoucí diplomové práce Ing. Heleně Škutkové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé diplomové práce.

Velmi rád bych poděkoval mé matce Vlastě Vadjákové za intenzivní podporu během celého mého studia. Zvláštní poděkování patří také mé přítelkyni Bc. Sabině Velasové, za její podporu, trpělivost a pomoc nejen při psaní této práce.

V Brně dne 23. května 2014 .....
podpis autora

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Obecná fylogenetika se snaží najít vývojové vztahy mezi organizmy na základě představy, že všechny organizmy měly svého univerzálního společného předka. Hlavní náplní fylogenetiky je tedy rekonstruovat fylogenetické stromy, popisující čas a způsob větvení jednotlivých organismů. Pokud za tímto účelem využíváme molekulární znaky, kterými jsou sekvence nukleotidů (DNA či RNA) nebo aminokyselin, můžeme pro jejich konstrukci využívat množství algoritmizovaných metod molekulární fylogenetiky. I tak je ovšem konstrukce těchto fylogenetických stromů náročnou disciplínou. Nikoliv snad svou samotnou realizací. Ta je díky rozvoji výpočetní techniky a programů, umožňujících tuto analýzu, snadná. Může však být o to zrádnější. Fylogenetický strom je totiž možné zkonstruovat na základě téměř jakýchkoli molekulárních dat a metod. Pro různá data a metody však můžeme dostávat různé výsledky. Interpretace a vyvozování závěrů z jediného fylogenetického stromu, u něhož si ani neověříme jeho robustnost, je zavádějící a pravděpodobně chybné [3], [4], [6], [7], [9].

Je tedy jasné, že pro správnou analýzu je zapotřebí dobře porozumět jednotlivým krokům konstrukce fylogenetického stromu. V každém z těchto kroků se totiž můžeme dopustit chyb, které by výsledky analýzy znehodnocovaly či zkreslovaly. Nejprve musíme vybrat vhodné vstupní sekvence, jejichž porovnání nám poskytne relevantní informace o evolučních změnách mezi srovnávanými taxony. Sekvence pak musíme vhodně zarovnat a odhadnout evoluční vzdálenost mezi každou dvojicí sekvencí na základě příslušného evolučního modelu. Teprve poté můžeme přistoupit k samotné konstrukci stromu. I zde se však mohou vyskytnout chyby. V této souvislosti nejčastěji hovoříme o artefaktech dlouhých větví. Všechny tyto problémy jsme však většinou schopni řešit. Nevhodné sekvence můžeme nahradit jinými. Pro odhad evoluční vzdálenosti můžeme použít jiné zarovnání i evoluční model. Pokud strom ovlivňují dlouhé větve, můžeme je z analýzy odstranit, nebo můžeme zkusit použít jinou konstrukční metodu. Tím se však dostáváme k nejdůležitější otázce, a sice jak vlastně poznáme, že strom nepopisuje evoluci korektním způsobem. Odpověď najdeme ve statistických metodách, díky kterým jsme schopni stromy testovat.

Práce se věnuje testování topologie stromů konstruovaných metodou Neighbor-Joining. Jedná se o distanční konstrukční metodu u které se pro testování využívají tzv. resamplingové testy. Ty jsou založené na opakovaném výběru a vycházejí z původních zarovnaných sekvencí, které postupně převzorkovávají. Na základě převzorkovaných dat poté konstruují stromy, které mezi sebou srovnávají. Jednotlivé testy se mezi sebou liší buď metodikou převzorkování sekvencí nebo srovnávání stromů. To, jak dobře fylogenetický strom popisuje analyzované znaky nám mohou pomoci odhalit první dvě metody. Jackknifing a Bootstrapping. OTU Jackknifing pak zkoumá vliv jednotlivých větví na topologii celého stromu. Metoda PTP (permutation tail probability) testuje vytvořený strom z hlediska délky evoluční cesty

a udává, zda sekvence obsahují dostatečnou fylogenetickou informaci pro odhad evoluce [9], [12].

Práce tak poskytuje ucelený přehled problematiky konstrukce fylogenetických stromů a zejména přehled statistických resamplingových metod, které mohu topologii těchto stromů testovat. Takovýto ucelený přehled o zmíněných statistických metodách zatím nebyl příliš publikován, a proto si práce klade za cíl otestovat jejich použitelnost na různých biologických sekvencích a také posoudit vliv volby vstupních parametrů analýzy na výsledky resamplingových testů.

1. Fylogenetika

Fylogenetika je věda, zabývající se fylogenezí, tedy vývojem jednotlivých vývojových linií organismů v čase. Fylogeneze vychází z evoluční teorie, podle které se jednotlivé druhy odvětvovaly od společného předka, čímž vnikaly vývojové linie, které se dále větvily. Tohoto společného předka často označujeme anglickou zkratkou LUCA (Last Univarsal Common Ancestor). Je jasné, že vývojově mladší linie budou zahrnovat podobnější druhy, v porovnání s liniemi, které se diferencovaly po delší dobu.

Hlavní náplní fylogenetiky je tedy rekonstruovat pořadí a způsob větvení všech vývojových linií v průběhu evoluce. Tuto rekonstrukci větvení nazýváme kladogeneze. Nezbytným doplňkem při studiu fylogeneze jí musí být anageneze, která studuje vznik a vývoj fenotypových znaků v rámci jednotlivých linií. Fylogeneze tedy zahrnuje celkový proces postupného oddělování jednotlivých vývojových linií (tj. kladogeneze) a hromadění jejich anagenetických změn [3], [4], [5].

V dalších kapitolách této práce se budeme setkávat s množstvím pojmů, jejichž význam nemusí být vždy zcela vysvětlen. Pro větší přehlednost nalezneme jejich vysvětlení v příloze 1 – Základní fylogenetické pojmy.

1.1. Převod a zpracování molekulárních znaků

Molekulární znaky, kterými se budeme v této práci zabývat, jsou sekvence nukleotidů či proteinů. Tato data získáváme sekvenováním DNA, RNA, případně proteinového řetězce. Sekvenování patří mezi přímé molekulární metody. Rozlišujeme několik sekvenačních technik, které se vzájemně liší především principem, ale také cenou a rychlostí. Jmenujme alespoň dvě základní techniky, Sangerova metoda a Maxam-Gilbertovo sekvenování. Existují také nepřímé metody, které jsou finančně a časově méně náročné. Neodhalují však všechny genetické rozdíly. Navíc mohou poskytovat znaky, které nejsou selekčně významné. Řadíme zde například hybridizaci DNA a imunologické metody [7], [8], [9].

Hlavním cílem kompletní analýzy je tedy zjistit pořadí nukleotidů, případně aminokyselin, tak, jak jsou za sebou v řetězci uspořádány. Dle směru syntézy DNA nukleotidy čteme od 5‘ konce ke 3‘ konci. Abychom data mohli lépe zpracovat, jsou v digitální podobě jednotlivé nukleotidy zapsány pomocí začátečních písmen dusíkaté báze, kterou obsahují: A, C, G, T (U), viz tabulka 1. Nukleotidy můžeme dále dělit dle typu báze (puriny, pyrimidiny), dle funkční skupiny (amino -NH2, keto -CO) a dle síly vazby (weak - slabá, strong - silná) [8], [9].

Tabulka 1: IUPAC kódy nukleotidů.

Základní tabulka		**Rozšířená tabulka**			
IUPAC kód	**Název**	**IUPAC kód**	**Dělení dle**	**Skupina**	**Nukleotidy**
A	Adenin	R	typ báze	pu**R**in	Adenin, Guanin
C	Cytosin	Y		p**Y**rimidin	Cytosin, Thymin, Uracil
G	Guanin	W	síla vazby	**W**eak	Adenin, Thymin, Uracil
T	Tymin	S		**S**trong	Cyticin, Guanin
U	Uracil	M	funkční skupina	a**M**ino	Adenin, Cytosin
		K		**K**eto	Guannin, Thymin, Uracil
		B	ne A		Cytosin, Guanin, Thymin, Uracil
		D	ne C		Adenin, Guanin, Thymin, Uracil
		H	ne G		Cytosin, Adenin, Thymin, Uracil
		V	ne T, U		Cytosin, Adenin, Guanin
		N	cokoli		Adenin, Cytosin, Guanin, Thymin, Uracil
		-	mezera		-

Poznámka: IUPAC (International Union of Pure and Applied Chemistry) - Mezinárodní unie pro čistou a užitou chemii je organizace zabývající se mimo jiné chemickou nomenklaturou [10].

Tabulka 2: IUPAC kódy aminokyselin s příslušnými kodony.

	Aminokyselina		**Kodóny**
IUPAC KÓD	**Zkratka**	**Název**	
A	Ala	alanin	GCT,GCC,GCA,GCG
C	Cys	cytosin	TGT,TGC
D	Asp	asparát	GAT,GAC
E	Glu	glutamát	GAA,GAG
F	Phe	fenylalanin	TTT,TTC
G	Gly	glycin	GGA,GGT,GGC,GGG
H	His	Histidin	CAT,CAC
I	Ile	izoleucin	ATC,ATA,ATT
K	Lys	lysin	AAA,AAG
L	Leu	leucin	TTA,TTG,CTT,CTC,CTA,CTG
M	Met	methionin	ATG
N	Asn	asparagin	AAT, AAC
P	Pro	prolin	CCT,CCC,CCA,CCG
Q	Gln	glutamin	CAA,CAG
R	Arg	arginin	CGT,CGC,CGA,CGG,AGA,AGG
S	Ser	serin	TCT,TCC,TCA,TCG,AGT,AGC
T	Thr	threonin	ACT,ACC,ACA,ACG
V	Val	valin	GTT,GTC,GTA,GTG
W	Trp	tryptofan	TGG
Y	Tyr	tyrosin	TAT,TAC
X	Xxx	cokoliv	

Strukturu bílkovin standardně zapisujeme od N-konce k C-konci, tedy od aminové skupiny ke karboxylové. Každá aminokyselina má svou zkratu i kód. Aminokyselina je vždy kódována tripletem nukleotidů. Tabulka 2 jasně ukazuje, že jedna aminokyseliny může být kódována několika různými kodony. Mutace v jednom nukleotidu nemusí nutně znamenat změnu v expresi genu, protože translací vznikne stejný protein jako v genu původním. To činí genetický kód odolnější vůči mutacím [11], [12], [13].

1.2. Molekulární fylogenetika

Díky rozvoji molekulárně-biologických metod a výpočetní techniky jsou v současné době ve fylogenetice ve velkém měřítku využívány molekulární znaky. Jedná se o znaky uložené v sekvencích informačních makromolekul – DNA, RNA, Proteiny. Nejčastěji využíváme pořadí monomerů v řetězci. Dají se však využívat i chemické či fyzikální vlastnosti těchto látek.

Molekulární znaky mají ve srovnání se znaky klasickými množství důležitých výhod. Pro lepší představu vnímejme v následujícím výčtu jako zástupce molekulárních znaků sekvenci DNA. Jako zástupce morfologických znaků si představme např. osteologické znaky, popisující vnější morfologii kostí [3], [6].

Výhody molekulárních znaků oproti morfologickým:

- Molekulárních znaků máme k dispozici podstatně větší množství.

V praxi jsme při získávání znaků limitováni pouze množstvím finančních prostředků a času, které jsme schopni a ochotni do daného výzkumu investovat. Je jasné, že např. mitochondriální DNA nám poskytne omezené množství informací. Pokud tedy nejsou výsledky průkazné, můžeme výzkum doplnit o jiné sekvence DNA, případně analyzovat stejný úsek jinými molekulárně-biologickými technikami.

- Jednotlivé molekulární znaky jsou na sobě nezávislé.

Pokud u některých taxonů chybí například celá morfologická struktura (celá končetina) bývá obtížné určit, zda se jedná o jeden znak, nebo o sérii chybějících znaků (všech kostí dané končetiny). Naproti tomu znaky zakódované v DNA jsou na sobě s daleko větší pravděpodobností nezávislé. Závislost se však v některých případech může objevit i u těchto znaků. Může se jednat o funkční závislost (komplementární nukleotidy v dvou-řetězcových oblastech rRNA), nebo historickou závislost (mutace vznikly v důsledku stejného poškození DNA). Díky tomu, že jsme však schopni jasně vymezit jejich polohu v řetězci DNA, můžeme jejich závislost odhalit pomocí statistických testů a jejich vliv na výsledek analýzy tak minimalizovat.

- Molekulární znaky jsou zpravidla kvalitativní.

Znaky klasické mají většinou charakter kvantitativní, či dokonce pravděpodobnostní. Znaky DNA jsou v praxi zapsány do alfabetických znaků a tím nejsou zatíženy žádnou ztrátou ani zkreslením informace. Díky tomu můžeme jednotlivé znaky snadno vymezit i popsat.

- Molekulární znaky umožňují porovnávat a třídit i vzájemně nepříbuzné organismy.

Pokud bychom měli určit morfologickými znaky, zda je si příbuznější například mech a slon v porovnání s žampionem a slonem, nejsme toho v podstatě schopni. Díky molekulárním znakům však tyto organismy jsme schopni porovnat. Nutno však dodat, že takováto analýza by mohla být zatížena velkými chybami.

- Molekulární znaky mají totožnou váhu

Při určování morfologických znaků často přisuzujeme odlišnostem většího významu větší váhu. Určení správné váhy je však zjevně zatíženo subjektivní chybou lidského úsudku. Molekulární znaky mají zpravidla shodné váhy. Existují i případy, kdy posuzujeme třeba, zda nukleotidy na dané pozici podléhají substitucím. V těchto případech však přisuzujeme váhy na základě jasného algoritmu a výsledek tak není zatížen subjektivní chybou.

- Počet společně sdílených znaků mezi dvěma druhy odráží příbuznost těchto druhů, nikoliv podobnost selekčních tlaků, které na ně v minulosti působily.

Molekulární znaky jsou často selekčně neutrální. Většinou tedy nedochází k chybnému zařazení v důsledku homoplazií, jak tomu bývá u morfologických znaků. Ve výjimečných případech může stejný selekční tlak vést k fixaci stejných mutací u vzájemně nepříbuzných druhů. Tyto případy jsou však ojedinělé a v případě, že k analýze využíváme velké množství znaků, nemůžou ojedinělé fixované mutace výsledek ovlivnit [3], [6].

Nevýhody molekulárních znaků oproti morfologickým:

- Molekulární znaky neposkytují informaci o anagenezi. Většina molekulárních znaků se vůbec neprojeví na fenotypu.

- Přestože se náklady na molekulární analýzu rok od roku snižují, je zisk molekulárních znaků stále v průměru nákladnější, než zisk znaků morfologických.

- Zisk morfologických znaků je v některých případech nevratný proces. Organismus je třeba zničit. Tato nevýhoda je však společná také pro anagenezi [6], [9].

2. Fylogenetické stromy

Výsledkem fylogenetické analýzy bývá nejčastěji nějaký typ fylogenetického stromu. Nikdy nemůžeme s jistotou tvrdit, že fylogeneze daných taxonů probíhala přesně nalezenou cestou. Vždy se jedná pouze o odhad fylogeneze popsaný příslušným fylogenetickým stromem. V této kapitole se budeme věnovat tomu, jak takový strom postupně vzniká, od výběru vhodných dat, které do analýzy budou vstupovat, až po metody samotné konstrukce stromů. Každá podkapitola na svém konci obsahuje shrnutí možných zdrojů nepřesností majících vliv na výsledný odhad fylogeneze [3].

2.1. Výběr vstupních dat fylogenetické analýzy

Základem našeho srovnávání je odhalit v sekvencích homologní části. Tedy ty části, které vznikly odvozením od společného předka. Konkrétně potřebujeme odhalit geny ortologní tak, abychom mohli změnu znaku za jiný hodnotit jako substituci, záměnu znaku za mezeru (ve značení IUPAC za pomlčku) považovat za deleci, případně inzerci. Právě s tímto předpokladem totiž počítají evoluční modely a vlastně i celá fylogeneze [3], [7], [18], [19].

Paralogní geny představují poměrně vysoké riziko záměny s geny ortologními. Tyto geny mají však u taxonů pozměněnou funkci i strukturu a výslednou analýzu by znehodnotili. Podobně negativní vliv mají také geny xenologní, tedy sekvence, které se do organismu dostaly horizontálním přenosem.

Další riziko představují pseudogeny. Tyto genové fragmenty jsou zbytky genů, které během evoluce ztratily svůj význam. Jde tedy o jakési nefunkční geny, které již nemají žáden vliv na fenotyp organismu. Pokud nemáme o použitých sekvencích dostatek apriorních znalostí, nejsme schopni pseudogeny, paralogní ani xenologní geny před analýzou v podstatě odhalit. Projeví se však nízkou podporou statistického hodnocení [3], [7], [12].

Dalším problémem je tzv. konvergentní evoluce, tedy evoluce probíhající v důsledku stejného selekčního nebo mutačního tlaku, například vlivem prostředí nebo stravy. Typickým příkladem jsou kytovci, kteří tvarem těla připomínají ryby, přestože se jedná o savce. Podobné jevy můžeme sledovat i na molekulární úrovni.

Lepší cestou, kterou se můžeme vydat, je vyhledat vstupní sekvence takové, jejichž okrajové části jsou dostatečně konzervované (není problém s jejich identifikací), jsou dostatečně fenotypově významné a specifické pro každý druh organismu a samozřejmě se jedná o geny ortologní. Vhodným bonusem by byl také fakt, kdyby byly geny přiměřeně dlouhé (kvůli výpočetní náročnosti) a v genomu dostatečně zastoupené.

Takové vlastnosti má například ve fylogenetice asi nejznámější gen 18s rRNA. Jedná se o stavební část živočišného ribozomu, má však důležitou úlohu také v rozpoznávání a správném nasměrování mRNA a tRNA. V roce 1988, kdy byl poprvé publikován, byl tento

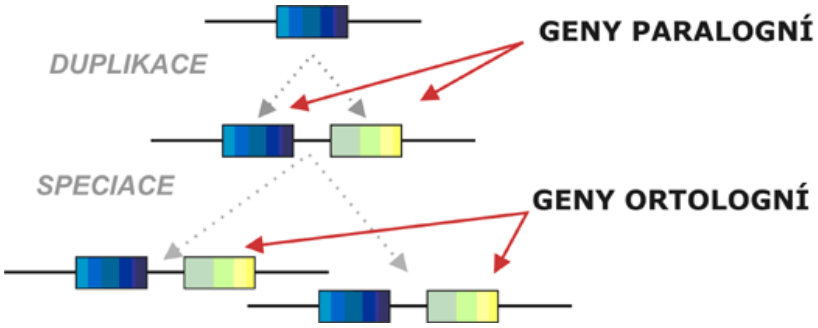
gen označen za ideálního kandidáta pro konstrukci živočišných fylogenetických stromů. Postupem času se našli i u tohoto genu problémy například při identifikaci některých mlžů a korýšů. Navíc bylo zjištěno, že u některých skupin živočichů klesá schopnost rozlišení všech vývojových linií. Obdobou pro fylogenezi bakterií je gen 16s rRNA [20], [21], [22].

Zdroje nepřesností pro odhad fylogeneze

Jedním z problémů může být nedostatečná délka vstupních sekvencí pro kompletní rekonstrukci evolučního rozvoje. Nejde však obecně říci, jaká délka je ideální, protože se pro různé studie bude vždy lišit.

Další problém představuje konvergentní evoluce. Homologie v sekvencích by pak neznačila evoluční příbuznost, ale fakt, že se organismy vyvíjeli vlivem stejného evolučního tlaku.

Největším rizikem je však zařazení pseudogenů, paralogních nebo xenologních genů jako vstupu do fylogenetické analýzy. Případně výběr zcela nehomologních genů. Někdy se však zařazení takovýchto nevhodných vstupů do analýzy můžeme dopustit. Všechny popsané problémy nejsme většinou schopni odhalit před analýzou. I proto existují statistické metody pro hodnocení podpory větvení, které nám mohou takovéto geny pomoci odhalit. Nejlepším řešením pro minimalizaci chybného výběru genů je mít dostatek informací o sekvencích, které hodláme využít [3], [7], [12].



Obrázek 1: Vznik paralogních a ortologních genů.

2.2. Zarovnání sekvencí

Pokud již máme vybrány vhodné sekvence, jejichž porovnání nám poskytne relevantní informace o evoluci jednotlivých taxonů, je našim dalším úkolem sekvence vhodně zarovnat pod sebe. Evoluční modely, o nichž pojednáváme v následující podkapitole, aplikujeme vždy na dvojici zarovnaných sekvencí. Je tedy jasné, že volba metody zarovnání bude mít na výslednou evoluční vzdálenost dvojice sekvencí veliký význam.

Zarovnání sekvencí chápeme v bioinformatice jako seřazení sekvence DNA, RNA či proteinů pod sebe tak, že výsledné skóre mezi sekvencemi je minimální. Skóre spočteme díky tomu, že přiřadíme každé substituci mezi sekvencemi, včetně možné mezery, číselnou hodno-

tu. Výsledné skóre je součtem dílčích substitucí, které jsme v sekvencích byli nuceni udělat [15], [16].

Penalizaci všech možných substitucí nejčastěji popisujeme skórovací maticí. Existuje celá řada již sestavených skórovacích matic, jako je například matice NUC44 pro zarovnání nukleotidových sekvencí, viz tabulka 3. Pro zarovnání aminokyselin používáme nejčastěji matice PAM (Point Accepted Mutation) nebo BLOSUM (BLOcks SUbstitution Matrix). Hodnoty těchto matic vycházejí z heuristicky naměřených dat. U PAM reflektují frekvence záměn jednotlivých aminokyselin. Jsou vztaženy vždy na daný počet mutací. Například PAM 1 bude skórovací matice, která vzešla z aminokyselin, které se lišily z 1%. Matice BLOSUM pak vycházejí z množství reálných zarovnání velkého množství sekvencí. Například BLOSUM 62 je stanoveno ze zarovnání s 62% identitou [15], [16].

Tabulka 3: Ukázka základní skórovací matice NUC44 pro popis substituce nukleotidů.

	A	**C**	**G**	**T**
A	5	-4	-4	-4
C	-4	5	-4	-4
G	-4	-4	5	-4
T	-4	-4	-4	5

Prvotním úkolem při fylogenetické analýze je nalezení homologních částí zkoumaných dat. Za tímto účelem je volba optimálního globálního zarovnání stěžejním předpokladem pro vytvoření vhodného fylogenetického výstupu. Bylo dokonce zjištěno, že výsledek fylogenetické analýzy je více závislý na metodě zarovnání, než na metodě fylogenetické rekonstrukce.

Některé prameny uvádějí vztah mezi sekvenční identitou a homologií, viz tabulka 4. Závěry v tabulce 4 jsou však formulovány dost vágně. Je nutné si uvědomit, že homologie a sekvenční identita jsou dva striktně oddělené údaje, přestože nám o sobě mohou leccos napovědět. Dalším problémem je fakt, že i velmi podobné sekvence, které budou zarovnány pod sebe, mohou, ale nemusí být homologní, viz předchozí kapitola 2.1 [15], [16], [17], [18], [19].

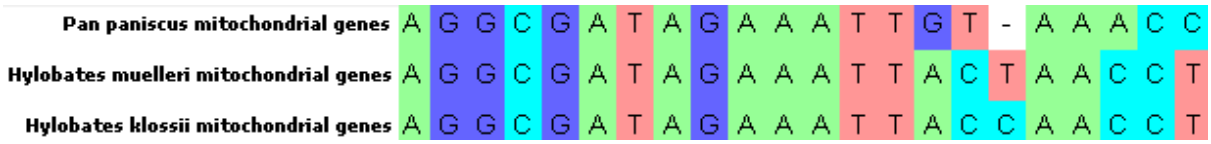
Tabulka 4: Vztah mezi sekvenční identitou a homologií.

Sekvenční identita	**Homologie**
35% - 100%	Pravděpodobně homologní
20% - 35 % „Twilight zone“	Může se jednat o homologii
0% - 20% „Midnight zone“	Nelze určit homologii

Poznámka: Tabulka platí pro dva proteiny o 100 a více aminokyselinách.

Zdroje nepřesností pro odhad fylogeneze

Pokud pomineme, že bychom zarovnávali nevhodné sekvence, viz předchozí kapitola 2.1, může být zdrojem nepřesností použití chybné skórovací matice či nevhodného zarovnávacího algoritmu, který by nesprávně zarovnal pod sebe nehomologní části sekvencí. Abychom tedy mohli výsledky analýzy považovat za věrohodné, je nutné mít o zkoumaných sekvencích co nejvíce informací, vybrat vhodné vstupy fylogenetické analýzy a ty poté vhodně zarovnat.



Obrázek 2: Zarovnaná části genů 18s rRNA tří ruzných organismů při použití skórovací matice NUC44.

2.3. Evoluční modely

Zarovnané sekvence potřebujeme mezi sebou porovnat z hlediska evoluce. Nejjednodušším modelem pro porovnání dvou sekvencí např. DNA je zarovnat sekvence pod sebe a spočítat rozdílná místa, případně podíl rozdílných míst k celkové délce sekvence. Tento podíl nazýváme proporcionální vzdálenost a značíme jej p , nebo také $p\text{-}distance$ (1):

$$p = \frac{počet\ rozdílných\ míst\ (=počet\ bodových\ mutací)}{délka\ sekvence}. \quad (1)$$

Výpočet proporcionální vzdálenosti je ilustrován na následujícím příkladu:

```
5' AGTAGATTAGGT 3'
3' CGCACATAAAGC 5'
   * * *   * * *
```

Znak * značí, že se na tomto místě sekvence liší, přičemž bereme v potaz také mezery vzniklé při zarovnání. Počet rozdílných míst je tedy 6 a celková délka sekvencí je 12. Po dosazení dostáváme:

$$\mathrm{p} = \frac{6}{12} = 0{,}5.$$

Výsledek můžeme interpretovat také tak, že se sekvence liší z 50 % [12].

Díky tomu, že mezi proporcionální a evoluční vzdáleností existuje vztah, můžeme na základě jednoduše pozorovatelných rozdílů mezi sekvencemi ($p\text{-}distance$) odhadovat jejich příbuznost ($d\text{-}distance$). K tomuto účelu slouží evoluční modely. Vycházejí z pravděpodobnostního rozložení řídkých jevů, jako je Poissonovo či Gamma rozložení. Tato rozložení modifikují pomocí známých faktů o bodových mutacích [3], [8], [12].

Přehled modelů nukleotidů, které budou použity i v praktické části práce, nalezneme v tabulce 5. Jednotlivé modely se navzájem liší počtem parametrů, se kterými počítají. V tabulce jsou seřazeny od nejjednodušších po nejsložitější [8], [12].

Tabulka 5: Přehled evolučních modelů sekvencí nukleotidů.

Název modelu	**Odhad evoluční vzdálenosti**
Jukes-Cantor [23]	$d = -\frac{3}{4}\ln\left(1-\frac{4}{3}\mathrm{p}\right)$
Kimura [24]	$d = -\frac{1}{2}\ln(1-2\mathrm{P}-\mathrm{Q})-\frac{1}{4}\ln(1-2\mathrm{Q})$
Tamura [25]	$d = -2\theta(1-\theta)\ln\left[1-\frac{\mathrm{P}}{2\theta(1-\theta)}-\mathrm{Q}\right]-$ $-[1-2\theta(1-\theta)]\frac{1}{2}\ln(1-2\mathrm{Q})$
Tamura-Nei [26]	$d = -\frac{2g_A g_G}{g_R}\ln\left(1-\frac{g_R}{2g_A g_G}\mathrm{P}_1-\frac{1}{2g_R}\mathrm{Q}\right)-$ $-\frac{2g_T g_C}{g_Y}\ln\left(1-\frac{g_Y}{2g_T g_C}\mathrm{P}_2-\frac{1}{2g_Y}\mathrm{Q}\right)-$ $-2\left(g_R g_Y-\frac{g_A g_G g_Y}{g_R}-\frac{g_T g_C g_R}{g_Y}\right)\ln\left(1-\frac{1}{2g_R g_Y}\mathrm{Q}\right)$

Poznámka: d - evoluční vzdálenost, p - proporcionální vzdálenost, P - relativní počet transic, Q - relativní počet transverzí, θ - relativní zastoupení cytosinu a guaninu, P_1 - relativní počet purinových transic, P_2 - relativní počet pyrymidinových transic, $g_Y = g_T + g_C$ a $g_R = g_A + g_G$, kde $g_{A,C,G,T}$ - relativní zastoupení příslušného nukleotidu.

Podobně jako u sekvencí nukleotidů postupujeme u evolučních modelů sekvencí aminokyselin. Základem je výpočet proporcionální vzdálenosti. Tuto vzdálenost poté aproximujeme dle Poissonova či Gamma rozložení, viz tabulka 6. Jednotlivé přístupy navíc můžeme kombinovat, například Jukes-Cantorův model s gamma korekcí [12].

Další možností je výpočet evoluční vzdálenosti na základě kodónového zápisu, který převedeme na zápis aminokyselin například pomocí některé z nejznámějších metod Nei-Gojobori či Li-Wu-Luo [27], [28].

Poslední možností pro odhad evoluční vzdálenosti aminokyselin je využití některé ze substitučních (skórovacích) matic (PAM, BLOSUM) [6].

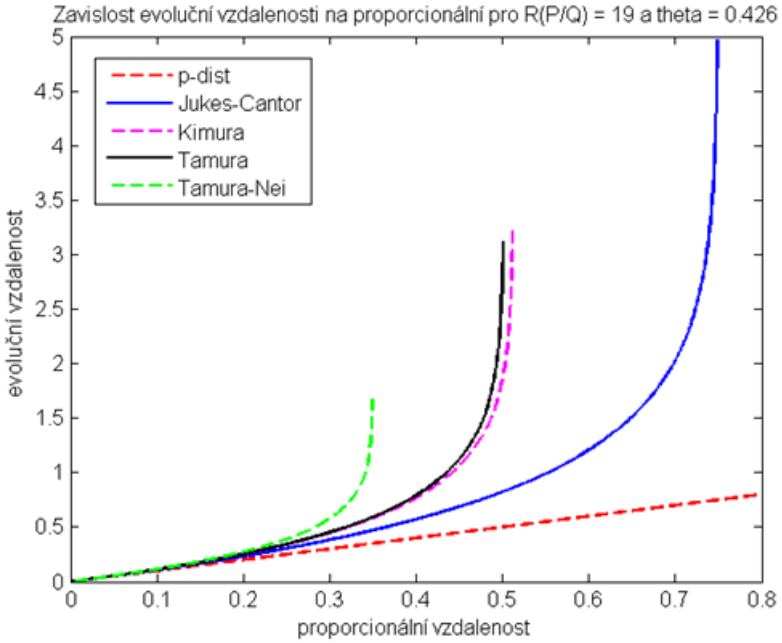
Tabulka 6: Přehled evolučních modelů sekvencí aminokyselin.

Název modelu	**Odhad evoluční vzdálenosti**
Poissonova korekce[12]	$\mathrm{d} = -\ln(1-\mathrm{p})$
Gamma korekce [12]	$\mathrm{d} = a[(1-p)^{\frac{1}{-a}}-1]$
Jukes-Cantor [12], [23]	$\mathrm{d} = -\frac{19}{20}\ln\left(1-\frac{20}{19}\mathrm{p}\right)$

Poznámka: d - evoluční vzdálenost, p - proporcionální vzdálenost, a - parametr řídící evoluční rychlost.

Zdroje nepřesností pro odhad fylogeneze

Je dobré si uvědomit, že ani nejdokonalejší model nikdy nebude absolutně přesný a tedy i použití modelu může být zdrojem chyb fylogenetické analýzy. Výsledkem není jednoznačně daná evoluční vzdálenost srovnávaných taxonů, nýbrž odhad jejich příbuznosti na základě daného modelu. Rozdílné průběhy evolučních modelů ukazuje obrázek 3.

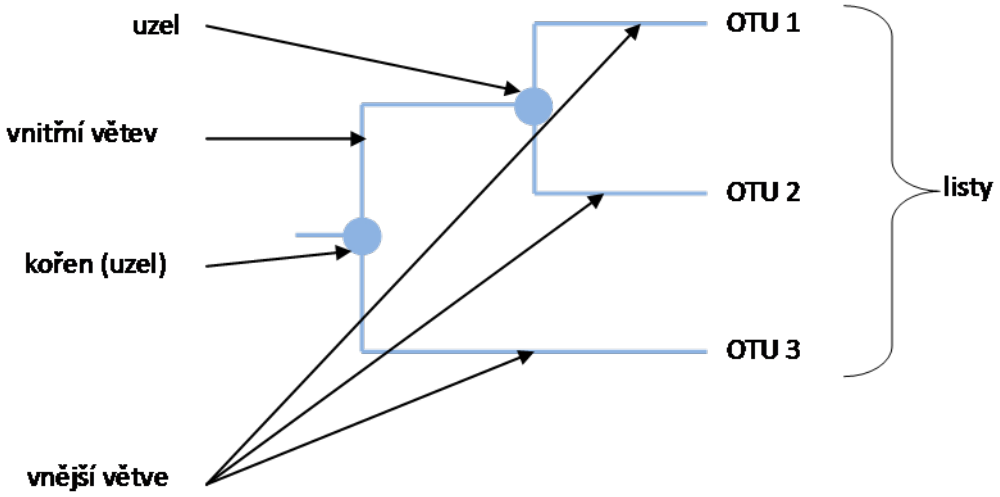


Obrázek 3: Závislost evoluční vzdálenosti na proporcionální pro různé evoluční modely.

2.4. Konstrukce fylogenetických stromů

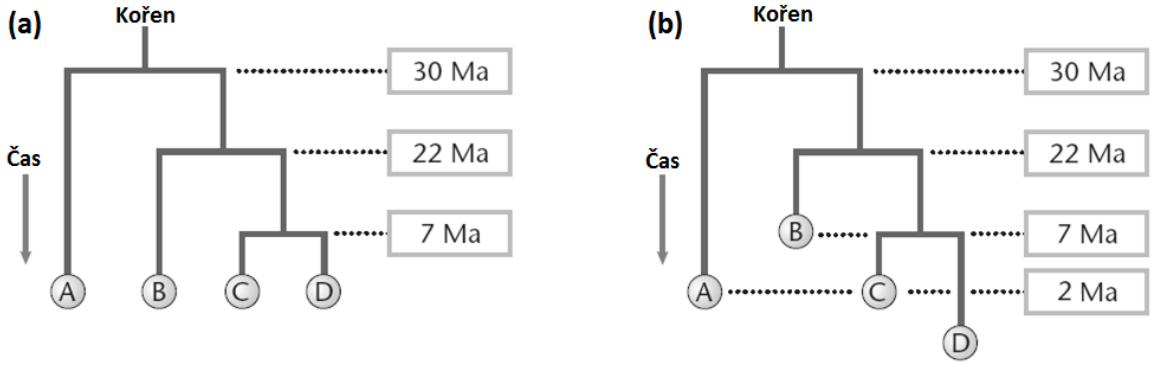
Fylogenetický strom graficky znázorňuje posloupnost událostí, formující množinu druhů. Příbuzenské vztahy zde můžeme popisovat pomocí morfologických nebo genetických dat. V této práci se budeme zabývat jen stromy vytvořenými na základě genetických dat [12].

Základními útvary, které při popisu fylogenetických stromů používáme, vycházejí z připodobnění ke stromům klasickým. Rozeznáváme tedy kořen, listy, větve a uzly stromu, viz obrázek 4.



Obrázek 4: Základní schéma dendrogramu - kladogram.

Kořen stromu může, ale také nemusí, být určen. Z toho vychází dva typy stromů, zakořeněný a nezakořeněný. Kořen v podstatě přestavuje nejbližšího společného předka všech srovnávaných taxonů. Uzel pak představuje nejbližšího společného předka dvou spojovaných taxonů. Taxony představují listy stromu. Větve na obrázku 4 určují jen pořadí divergence jednotlivých taxonů. Délky větví v tomto konkrétním případě nenesou žádnou informaci. Jedná se o tzv. kladogram. Pokud ke stromu přidáme i časovou osu, budou délky větví značit čas, za který od sebe jednotlivé taxony divergovali, viz obrázek 5 vlevo. Pokud se taxony nevyvíjejí se stejnou evoluční rychlostí, budou se lišit i délky větví, viz obrázek 5 vpravo. Časovou jednotkou bývají roky, značené a (z latinského annus). Protože bývají evoluční vzdálenosti často řádově vyšší než jednotky let, častěji využíváme jednotky Ma - milion let [2], [12].



Obrázek 5: Základní dendrogram s časovou osou (tedy fylogram).
(a) Se stejnou délkou větví, (b) S různou délkou větví.

V zásadě existují dva přístupy ke konstrukci fylogenetických stromů. Distanční metody a znakové metody. Každá z těchto metod má své výhody i nevýhody a jsou popsány v následujících kapitolách podkapitolách [2], [8].

2.4.1. Znakové metody konstrukce fylogenetických stromů

Znakové metody používají jako vstup přímo sekvenci znaků. Následně pracují s pravděpodobností změny tohoto znaku. Algoritmus vytvoří všechny možné stromy a poté vybere strom, který splňuje předem dané kritérium optimality. Pokud uvažujeme bifurkovaný (z jednoho společného předka divergují vždy dva noví jedinci), zakořeněný strom, je počet možných topologií dán vztahem (2):

$$\frac{(2m-3)!}{2^{m-2}(m-2)!}, \tag{2}$$

kde m je počet vstupních sekvencí. V případě nezakořeněného bifurkovaného stromu je počet možných topologií dán vztahem (3):

$$\frac{(2m-5)!}{2^{m-3}(m-3)!} \tag{3}$$

Množství všech možných stromů tedy roste s počtem vstupních sekvencí. Pokud bychom vytvářeli všechny možné topologie stromů a ty mezi sebou teprve porovnávali, bylo by řešení velice výpočetně náročné. Naproti tomu tyto metody konvergují z hlediska kritéria vždy k optimálnímu řešení. Následuje stručné představení dvou nejběžnějších znakových metod, ze kterých další znakové metody většinou vycházejí [2], [8].

Maximum parsimony (MP)

Jedná se o metodu, jejímž kritériem je minimální počet substitucí, které jsou potřeba pro vysvětlení dané topologie. Zjednodušeně řečeno tedy konstruujeme všechny možné topologie a na základě kritéria vybereme substitučně nejúspornější řešení. Existuje mnoho modifikací této metody, které nevyžadují prohledávání celého stromového prostoru [2], [8], [9], [30].

Maximum likelihood (ML)

Metoda maximum likelihood, nebo-li metoda maximální věrohodnosti, předpokládá, že se do náhodného výběru dostávají znaky s různou pravděpodobností. Pracuje tedy s pravděpodobnostními substitučními modely. Vybrán je strom, který má největší pravděpodobnost (je nejvěrohodnější) pro danou topologii. Opět existují modifikace, při kterých neprohledáváme celý stromový prostor [2], [8], [9].

2.4.2. Distanční metody konstrukce fylogenetických stromů

Distanční metody vycházejí z matice distancí, viz tabulka 7, která udává vzájemné evoluční vzdálenosti mezi všemi dvojicemi taxonomických jednotek, které uvažujeme. Je tedy odvozená na základě distančních modelů, viz kapitola 2.3. V jejich průběhu je vytvářen zpravidla jen

jeden strom, zkonstruovaný na základě daného algoritmu. Jsou tedy zpravidla méně výpočetně náročné. Naproti tomu vedou k suboptimálnímu řešení, u kterého nemůžeme s určitostí tvrdit, že je zároveň i řešením optimálním. Následuje stručné představení tří nejběžnějších distančních metod, ze kterých další distanční metody většinou vycházejí [8], [9], [12].

Tabulka 7: Distanční matice tří taxonů.

	Pan paniscus	Hylobates muelleri	Hylobates klossii
Pan paniscus	0	0,1132	0,1174
Hylobates muelleri	0,1132	0	0,0334
Hylobates klossii	0,1174	0,0334	0

Poznámka: Distanční matice byla vytvořena na základě Jukes-Cantorova modelu aplikovaného na zarovnané sekvence (pomocí skórovací matice NUC44).

Minimum evolution (ME)

Kritériem optimality pro tuto distanční metodu je suma délek všech větví dané topologie. Na základě distanční matice jsou tedy zkonstruovány všechny možné topologie. Nakonec je vybrána ta, která má nejmenší sumu délek všech větví. Metoda má opět mnoho modifikací, které prohledávání celého stromového prostoru urychlují [2], [8], [9], [31].

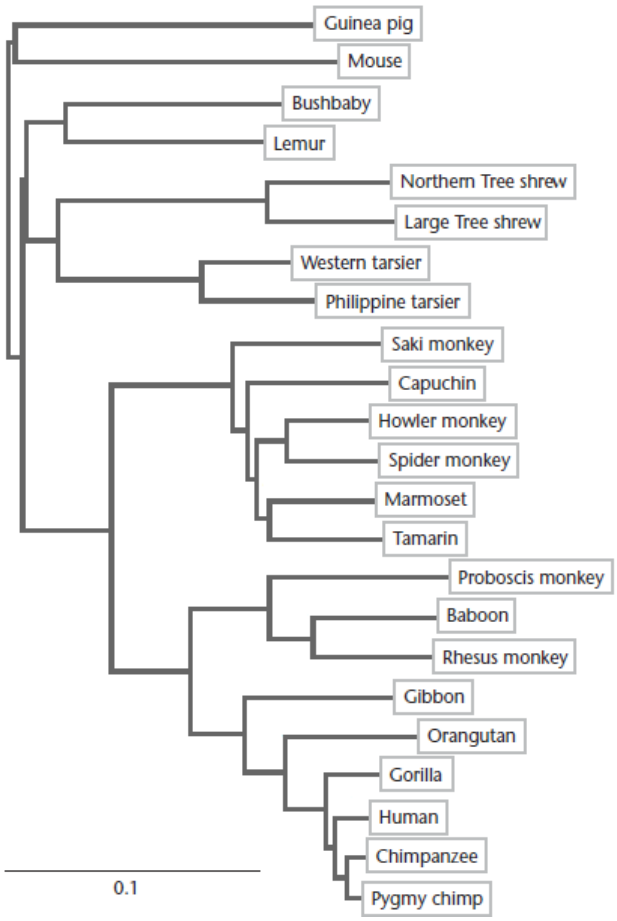
Unweighted Pair Group Method with Aritmetic mean (UPGMA)

Nejjednodušší modifikace metody minimální evoluce. Metoda vybere dvojici taxonů, které mezi sebou mají nejmenší evoluční vzdálenost a spojí je do uzlu. Následně se spočítají vzdálenosti vytvořeného uzlu od ostatních taxonů a kroky se opakují, dokud nevznikne kompletní strom. Uzel vždy umístíme do středu obou taxonů. To znamená, že vzdálenosti v podstatě průměrujeme. Délky větví pro oba taxony budou stejné. Metoda neuvažuje rozdílnou evoluční rychlost jednotlivých OTU, což samozřejmě nemusí odpovídat skutečnosti [2], [8], [9].

Neighbor-joining (NJ)

Metoda spojování sousedů vychází taktéž z metody minimální evoluce. Vycházíme z multifurkovaného stromu, kdy jsou všechny taxony spojeny do jednoho uzlu. Poté spojíme do nového uzlu dva taxony a spočteme sumu délek větví od jejich společného uzlu ke zbylému hvězdicovému uspořádání. Aplikací pro každý pár OTU vytvoříme sumační matici. Minimální hodnota v sumační matici poukazuje na pár, který má nejmenší sumu délek větví stromu. Tento pár tvoří nejbližší sousedy (je spojen uzlem). Zbývá vypočítat konečnou délku větví zvolených sousedních OTU. Jedná se o odhad evoluční vzdálenosti na základě metody nejmenších čtverců, kterou dosahujeme minimální evoluční cesty ve stromu. Nakonec přepočítáme distanční matici s nově vytvořeným uzlem. Dále postupujeme stejným způsobem, dokud nepospojujeme všechny OTU do výsledného fylogenetického stromu.

Jedná se o metodu, která uvažuje rozdílnou evoluční rychlost jednotlivých listů stromu. Navíc není výpočetně náročná. Metoda spojování sousedů, či některé její modifikace, je tedy jednou z nejpoužívanějších rekonstrukčních metod [2], [8], [9], [32].



Obrázek 6: Fylogenetický strom sestrojený metodou Neighbor-Joining.

Zdroje nepřesností pro odhad fylogeneze

Při rekonstrukci fylogeneze může vznikat hned několik artefaktů, které se odrazí v chybné topologii stromu. Citlivost na vznik těchto artefaktů je závislá na vlastní konstrukční metodě.

Problém představují hlavně velmi vzdálené sekvence. Pokud předpokládáme, že jsou vstupní data homologní, je jejich příčinou fakt, že u některých sekvencích probíhala substituce výrazně vyšší rychlostí než u ostatních, případně jedna sekvence vznikla fúzí dvou různých genů s různou evoluční historií. Následkem toho se vytvářejí **artefakty dlouhých větví** [2], [3], [12]:

a) Přitahování dlouhých větví (LBA – long branch attraction): Dvě velmi rozdílné sekvence (mezi sebou i vůči ostatním) jsou přitahovány k sobě a směrem ke kořeni stromu. Zvláště metoda maximální parsimonie je na tento problém citlivá.

b) Odpuzování dlouhých větví (LBR – long branch repulsion): Dvě velmi rozdílné sekvence vůči ostatním, avšak mezi sebou podobné, jsou od sebe odpuzovány (nejsou spojeny v uzel). Zejména metoda maximum likelihood často obě dlouhé větve od sebe oddělí.

c) Vyrušování dlouhých větví (LBD – long branch distract): Jedna dlouhá větev ovlivňuje topologii jiné části stromu, nikoliv však to, kam se sama zařadí [3], [12].

Pro odstranění těchto artefaktů můžeme použít jinou, méně citlivou metodu konstrukce stromů. Další možností je použít jiný algoritmus pro výpočet evoluční vzdálenosti, který uvažuje rozdílnou evoluční rychlost. Problém však zůstává v tom, že stejný model musíme aplikovat vždy na všechny sekvence. Další možností redukce dlouhých větví je vypuštění třetí pozice kodonu. Toto řešení je založeno na představě, že se třetí pozice příliš rychle vyvíjejí, protože jejich změna nemusí znamenat zároveň i změnu aminokyseliny a tedy i znaku (tzv. synonymní substituce). Přestože se díky tomu sníží vliv dlouhých větví, ztratíme tím také značnou část informací a snížíme tak výsledné rozlišení stromu. Také úplné vyloučení problematické větve není vždy řešením, poněvadž někdy může být našim cílem právě správné zařazení dlouhé větve. V poslední době velmi doporučovaný přístup je naopak přidání nových větví do analýzy, které redukují dlouhé větve (dlouhých větví bude větší množství). Omezením samozřejmě může být, když relevantní data již nemáme k dispozici (studovaný vzorek taxonů již například vymřel, či jsme již odebrali DNA všech exemplářů, či jiné důvody). Přidání nových dat zároveň nemusí mít na artefakty vliv, případně může produkovat artefakty nové. Další možností je kombinovat přístupy morfologické a molekulární [29], [33].

Dalším problémem je samotné odhalení dlouhých větví. Ty totiž nemusejí být, zvláště u rozsáhlejší analýzy, vždy hned zřejmé. Nejčastější možností detekce je předpoklad, že problematické větve budou mít nízkou statistickou podporu uzlů či větví (kapitola 3).

Pro detekci artefaktu přitahování dlouhých větví je vhodná metoda parametrického bootstrapingu (parametric simulation), která zkoumá, zda jsou dvě větve, které spolu nesousedí, dostatečně dlouhé, aby se přitahovaly. Metoda však nedokáže jednoznačně určit, zda se artefakt projeví, pouze určí pravděpodobnost, s jakou se může projevit.

Existují i další modifikace této metody, ať už více, či méně uznávané. Další metodou pro odhalení dlouhých větví je metoda RASA (Relative Apparent Synapomorphy Analysis). Metoda původně vyvinuta k měření fylogenetického signálu může dlouhé větve odhalit jako odlehlé hodnoty se zvýšeným fylogenetickým signálem. Metoda je však náchylná k falešně pozitivním výsledkům a dalším problémům a na základě posledních studií není doporučována.

Další možností jsou metody, založené na spektrální analýze dat, které dokážou problematické úseky v datech identifikovat. Přestože nebyla tato metoda původně navržena za tímto účelem, zdá se, že by její výsledky mohly být slibné a aplikace spektrální analýzy pro identifikaci dlouhých větví je předmětem intenzivního zkoumání [29], [34], [35].

Při konstrukci fylogenetických stromů je nutné si uvědomit, že stejná data produkují při různých metodách rozdílnou topologii. Nelze sestavit obecné pravidlo pro výběr optimální konstrukční metody stromu. Ta se bude vždy lišit v závislosti na charakteru vstupních dat. V dnešní době je pro identifikaci dlouhých větví, ale i dalších nepřesností v odhadu fylogeneze nejčastěji využíváno kombinovaných přístupů, založených na konstrukci stromů různými metodami při současném testování spolehlivosti topologií, pro které nám slouží statistické testy [2], [3], [12], [29].

3. Resamplingové testy fylogenetické analýzy

Většina chyb fylogenetické analýzy nejde předem odhalit. V dnešní době již však existují metody, které testují statistickou podporu větvení, jako je bootstrapping či jackknifing. Chyby analýzy se pak mohou projevit nízkou podporou uzlů. Díky tomu můžeme získat o vytvořeném fylogenetickém stromu nové informace, přestože je jejich interpretace mnohdy složitá a ne zcela explicitní. S jiným přístupem počítá metoda PTP (permutation tail probability), která testuje vytvořený strom z hlediska délky evoluční cesty. Poslední rozebíranou statistickou metodou bude OTU jackknifing, která zkoumá vliv jednotlivých větví na topologii celého stromu.

Všechny jmenované testy patří do rodiny tzv. resamplingových testů. Chápeme je jako metody založené na opakovaném výběru. Vycházejí z původních zarovnaných sekvencí, které postupně převzorkovávají, nebo-li je záměrně pozměňují. Na základě převzorkovaných dat poté konstruují stromy, které mezi sebou srovnávají. Metodika převzorkování a srovnání je charakteristická a individuální dle zvolené statistické metody [6], [8], [9], [12].

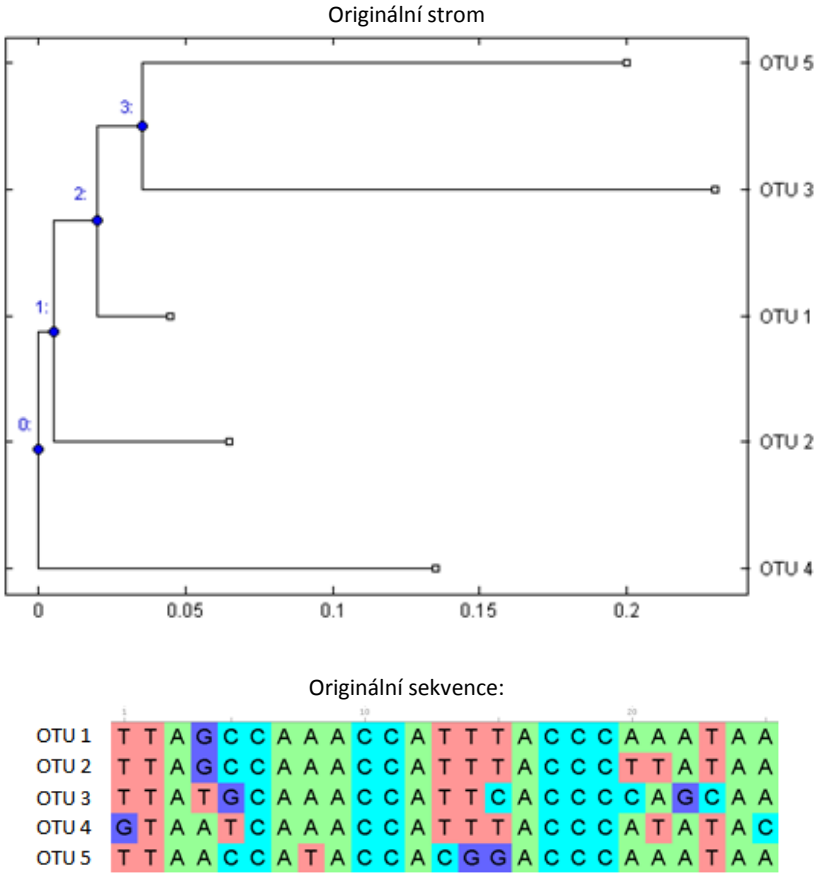
3.1. Bootstrapping

Bootstrapping je v dnešní době asi nejpoužívanější fylogenetickou statistickou metodou. Jedná se o obecnou neparametrickou metodu z roku 1979, která našla své využití v mnoha odvětvích a oborech. Využívá se pro odvození robustnosti populace plynoucích ze standardních chyb. Využívá se zejména tam, kde je využití parametrických testů nemožné nebo příliš složité. Velikou výhodou je fakt, že nemusíme znát rozložení vstupních dat.

Výsledkem je bootstrapová hodnota (bootstrap percentage, bootstrap p-value), která vyjadřuje stupeň podpory větvení (tedy uzlů), viz obrázek 10. V dalším pokračování této kapitoly si v jednotlivých krocích na ilustrovaném příkladu ukážeme, jak k procentuálním hodnotám dojdeme [6], [8], [38].

1) Sestrojení originálního stromu

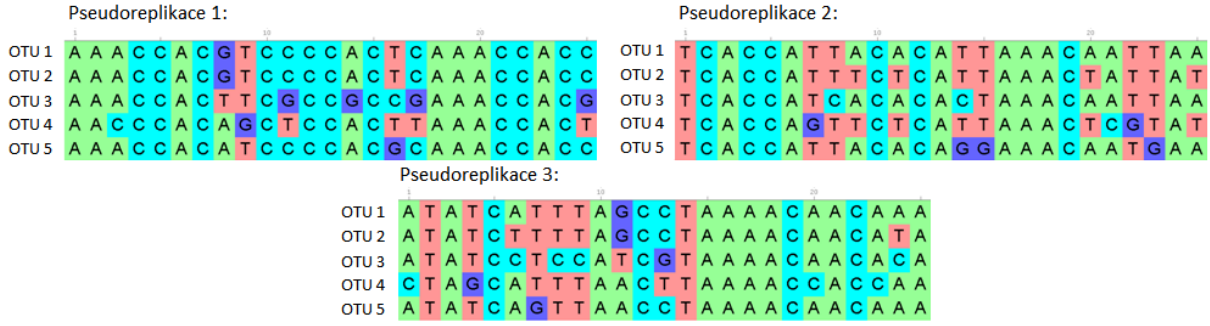
U bootstrapingu vycházíme z originálních zarovnaných sekvencí, které můžeme vidět na obrázku 7. Na základě evolučního modelu (zde byla využita proporcionální vzdálenost) mezi sekvencemi vytvoříme distanční matici a poté pomocí vybrané konstrukční metody (zde byla využita metoda neighbor-joining) sestrojíme originální strom [6], [8], [9].



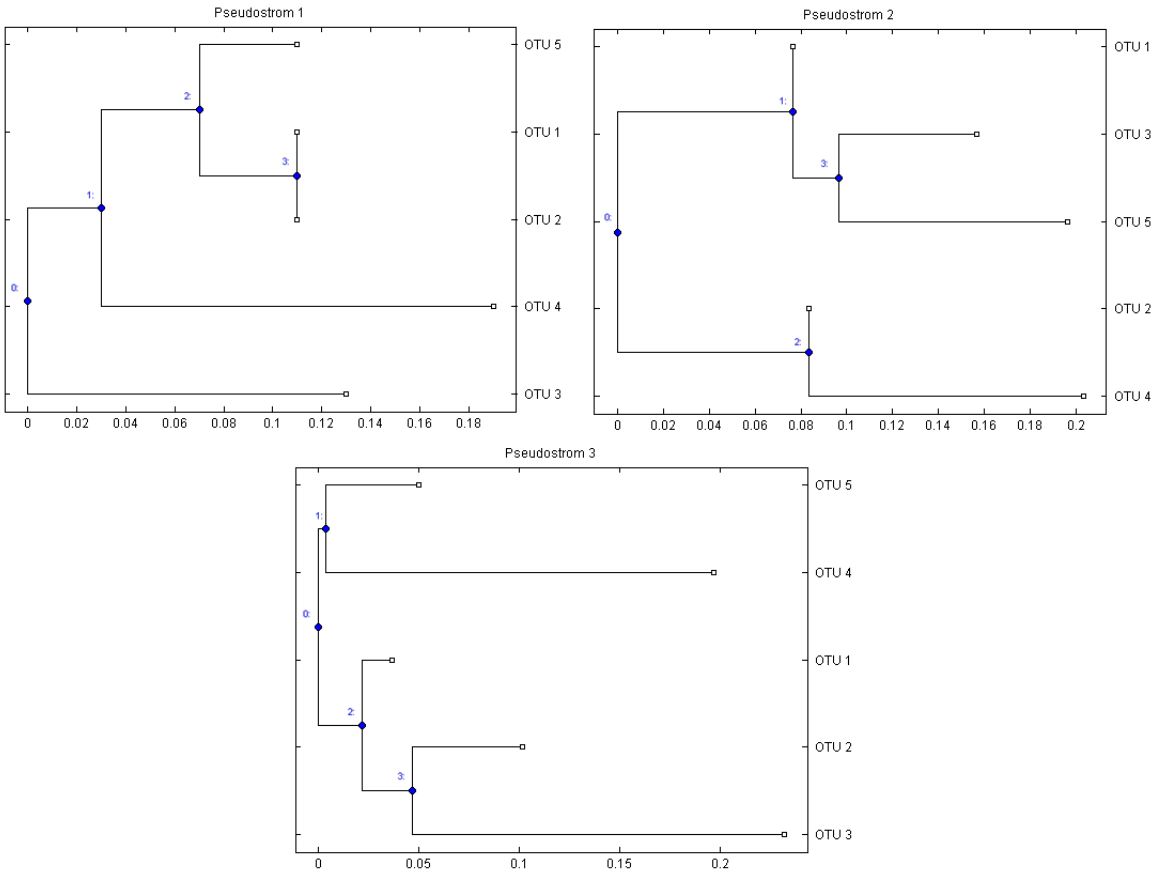
Obrázek 7: Originální strom konstruován na základě originálních sekvencí.

2) Vytvoření pseudoreplikací

V dalším kroku vytvoříme pseudoreplikaci původních sekvencí tak, že provedeme náhodný výběr sloupců originálních sekvencí s opakováním. Na obrázku 8 vidíme, že došlo ke změně pořadí některých sloupců, některé se opakují a některé byly zcela vypuštěny. Z takto získaných sekvencí konstruujeme stejným způsobem jako u originálních dat strom. Těchto pseudoreplikací a příslušných pseudostromů vytvoříme větší množství, v našem případě byly pro názornost vytvořeny 3, viz obrázek 8 a 9 [6], [8], [9].



Obrázek 8: Tři pseudoreplikace originálních sekvencí.



Obrázek 9: Tři pseudostromy konstruované na základě příslušných pseudoreplikací.

3) Porovnání uzlů

Na základě tří nově vzniklých pseudostromů porovnáme uzly originálního stromu s uzly všech pseudostromů, při čemž nám nebude záležet na poloze uzlu, ale jen na taxonech, které uzly spojují, a to bez ohledu na jejich pořadí. Tabulka 8 znázorňuje všechny porovnávané uzly s vzestupně seřazenými taxony. Stejnou barvou jsou rozlišeny ty, které se shodují. Uzel 0 (kořen stromu) bude logicky zastoupen vždy. V pseudostromu 2 byli vytvořeny shodné uzly číslo 2 a 3. Uzel 1 se neshoduje ani s jedním uzlem pseudostromů. Relativní a procentuální zastoupení uzlů je shrnuto v tabulce 9 [6], [8], [9].

Tabulka 8: Přehled uzlů originálního stromu a tří pseudostromů.

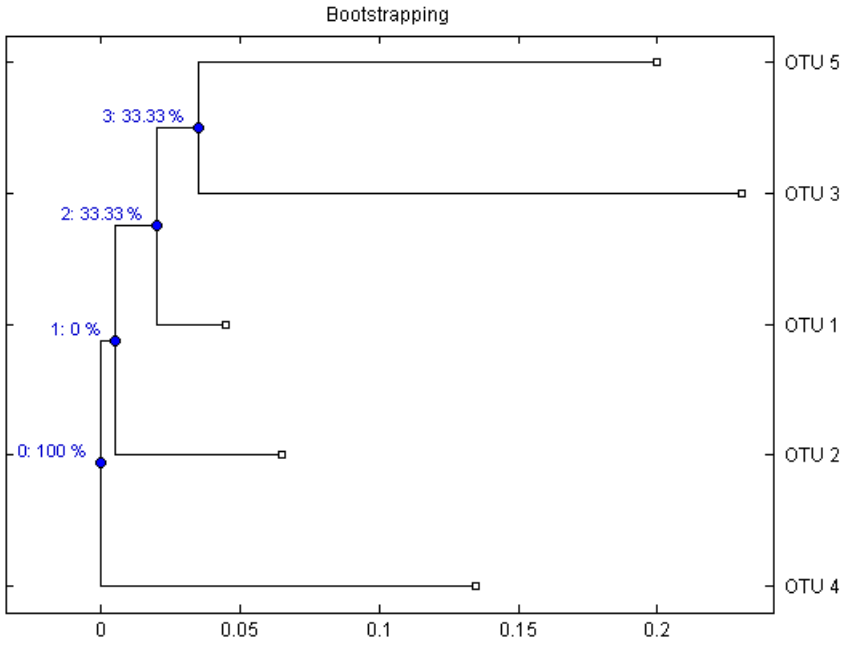
	Originální strom	Pseudostrom 1	Pseudostrom 2	Pseudostrom 3
Uzel 0 (kořen)	OTU 1, 2, 3, 4, 5	OTU 1, 2, 3, 4, 6	OTU 1, 2, 3, 4, 7	OTU 1, 2, 3, 4, 8
Uzel 1	OTU 1, 2, 3, 5	OTU 1, 2, 4, 5	OTU 1, 3, 5	OTU 1, 2, 3
Uzel 2	OTU 1, 3, 5	OTU 1, 2, 5	OTU 2,4	OTU 2, 3
Uzel 3	OTU 3, 5	OTU 1, 2	OTU 3, 5	OTU 4, 5

Tabulka 9: Relativní a procentuelní zastoupení uzlů u tří pseudoreplikací.

	Relativní zastoupení uzlu	Procentuelní zastoupení uzlu
Uzel 0 (kořen)	3/3 = 1	100 %
Uzel 1	0/3 = 0	0 %
Uzel 2	1/3 = 0,3333	33,33 %
Uzel 3	1/3 = 0,3333	33,33 %

4) Vytvoření výsledného stromu s vyznačenou bootstrappingovou podporou

Výsledný strom tedy bude mít shodnou topologii s originálním stromem, ale navíc bude mít u příslušného uzlu vyznačenou procentuální hodnotu bootstrappingové podpory z tabulky 9, viz obrázek 10 [6], [8], [9].



Obrázek 10: Fylogenetický strom s vyznačenou Bootstrappingovou podporou.

Ukázku výsledku bootstrappingové analýzy lze vidět na obrázku 12. Pro srovnání jsou na témže obrázku uvedeny i hodnoty jackknifingové podpory, probírané v následující kapitole 3.2.

Přesnost výsledků a jejich interpretace

V našem příkladě jsme pro názornost uvažovali jen tři pseudoreplikace. V praxi jich ovšem musíme vytvořit více, abychom mohli výsledky považovat za věrohodné. Konkrétně je doporučováno použití alespoň 500 pseudovzorků. Tato hodnota vychází z normálního rozložení bootstrappingových vzorků, které je logické, jelikož jsme při převzorkování využívali náhodného výběru. Pro požadovanou přesnost a vypočteme potřebný počet vzorků n ze vztahu (4):

$$n = \mathrm{BP} \cdot (1 - \mathrm{BP}) \cdot \frac{\sigma}{a^2}. \qquad (4)$$

Směrodatná odchylka normálního rozložení σ má pro 95 % interval spolehlivosti hodnotu 1,96. BP je vypočítaná hodnota bootstrappingové podpory. Pro velký interval BP je při použití 500 pseudovzorků výsledná přesnost +- 4 %. Se snižujícím se počtem vzorků přesnost samozřejmě klesá, naopak se vzrůstajícím počtem je přesnost vyšší, avšak za cenu zvýšené výpočetní náročnosti. Stanovit přesný počet pseudoreplikací, který by měl být použit u jakékoli studie, samozřejmě nelze. Přesto je doporučováno využívat alespoň 500 pseudoreplikací, jako hodnoty zlatého středu mezi náročností a přesností [39], [36].

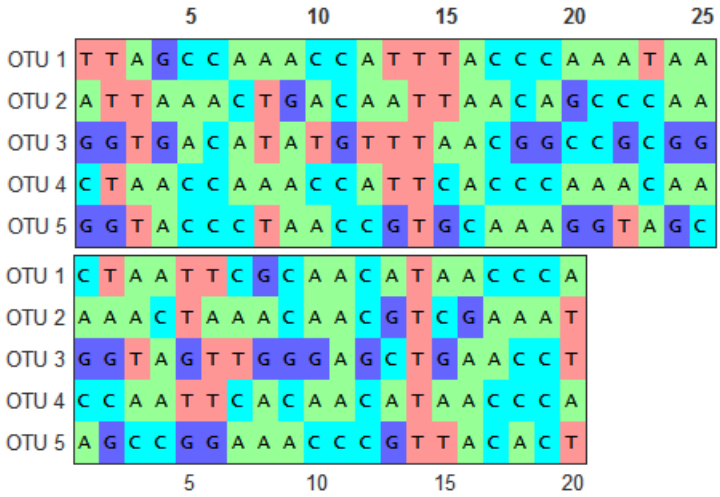
Interpretace bootstrappingové podpory z objektivního, statistického hlediska je velmi složitá a doposud se jí nepodařilo přesně definovat. Odhalit skutečné rozložení všech fylogenetických dat totiž není možné. Existují mnohé studie, které testování stromů touto metodou kritizují, stejně jako existují ty, které tuto statistickou metodu protěžují. Všeobecně je však tato metoda uznávaná a hojně využívaná. Její interpretace však může být pouze obecná.

Obecnou interpretací bootstrappingové podpory chápeme jako statistickou podporu námi vytvořeného větvení pro námi zvolená data (se všemi kroky, které vedly k vytvoření fylogenetického stromu, viz kapitola 2). Procentuální hodnotu podpory v žádném případě nesmíme chápat jako pravděpodobnost toho, že v minulosti k tomuto větvení došlo. Fylogenetickým stromem se snažíme popsat historii, kterou neznáme a vždy ji pouze odhadujeme. Žádná metoda nám nikdy neprozradí, jak dobře jsme onen odhad provedli. Statistická podpora nám tedy v podstatě udává to, zda jsme námi vybranými metodami vytvořili strom, který naše vstupní data vhodně vysvětluje. Fylogramy, které mají podporu 70 % a vyšší jsou považovány za dobře obhájitelné. Hodnota pod 50 % je brána jako nedostatečná. Naše data, či způsob jejich interpretace, by neobsahovala dostatečnou fylogenetickou informaci pro vytvoření daného větvení [1], [6], [8], [36], [37], [40].

3.2. Jackknifing

Jackknifing je historicky starší metodou, ze které bootstrapping v podstatě vychází. Byla publikována roku 1949 a rozšířena roku 1958. Jedná se rovněž o neparametrickou statistickou metodu, která je využívána v mnoha odvětvích pro odhad standardních chyb výběru, jehož rozložení neznáme.

Samotný algoritmus výpočtu je ve své podstatě téměř totožný s bootstrappingem. Jediný rozdíl je ve způsobu vytváření pseudovzorků. Zatímco u bootstrappingu vybíráme sloupce s opakováním, jackknifing vybere každý sloupec maximálně jednou. Navíc dochází ke zkrácení původní délky sekvencí o uživatelem zvolený počet sloupců. Ukázka vytvořené jackknifingové pseudoreplikace je na obrázku 11. Nahoře je možné vidět původní data set, pod ním jsou vytvořené nové pseudovzorky, které jsou o pět nukleotidů zkráceny. Každý sloupec z původního data setu se v novém objeví maximálně jednou. Další kroky jsou shodné s bootstrappingem popsaným v kapitole 3.1. Shodná je i interpretace výsledků jackknifingové podpory [6], [9], [41], [42].

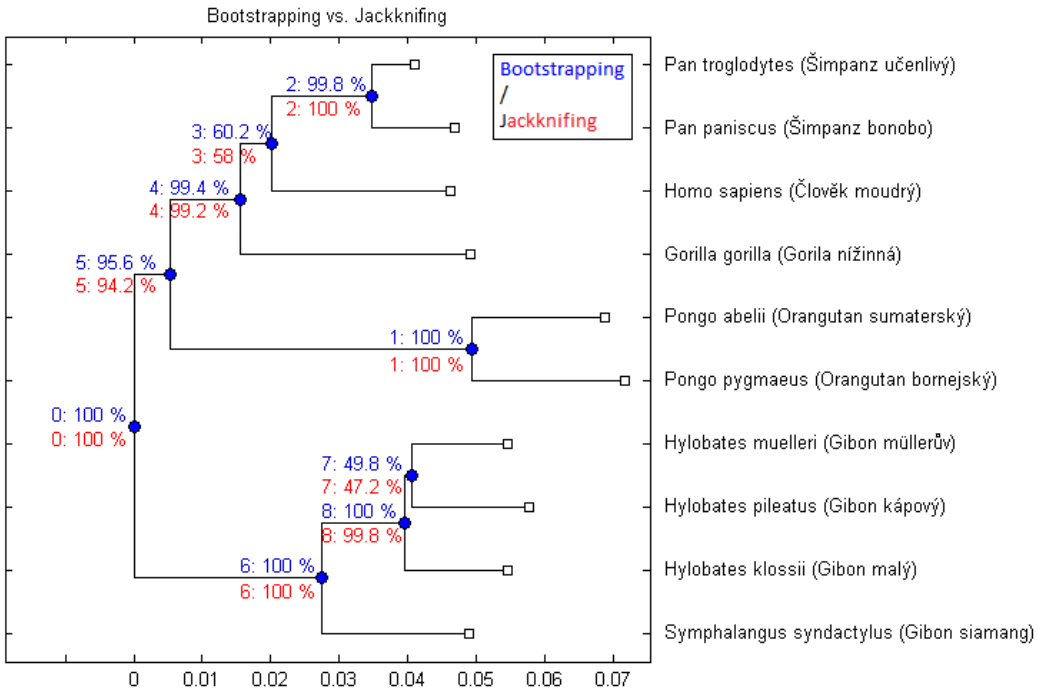


Obrázek 11: Originální sekvence a jejich jackknifingové pseudoreplikace.
Nahoře: originálních sekvence. Dole: jackknifingové pseudoreplikace.

Rozdíly mezi jackknifingem a bootstrappingem a interpretace výsledků

Obě metody vycházejí z jiných statistických předpokladů. Jackknifing přináší informaci o rozptylu vstupní populace, zatímco bootstrapping nejprve také odhaduje rozptyl vstupní populace, od něhož však poté počítá odchylku. Obě metody však přinášejí podobné numerické hodnoty statistické podpory. Ukázku výsledku jackknifingové analýzy lze vidět na obrázku 12 (dole). Pro srovnání je na témže obrázku (nahoře) uveden i strom s bootstrappingovou podporou.

Jedním z problémů jackknifingu je odhalit ideální počet nukleotidů, o které budeme sekvence zkracovat. Podle Felsteina (1985) se doporučuje zkrátit sekvence přibližně o 50 % původní délky. Další rozdíl je v reprodukovatelnosti výsledků. Zatímco jackknifing dává obdobné výsledky téměř pokaždé a nepotřebuje k tomu takové množství pseudoreplikací, bootstrapping je při použití menšího počtu pseudoreplikací méně stabilní. Největším problémem jackknifingu je však v tom, že jackknifing odhaduje správně pouze populace s hladkým, konzistentním rozptylem. Tato podmínka se také stala nejčastějším argumentem odborníků, kteří upřednostňuje použití bootstrappingu před jackknifingem. Zjednodušeně lze tedy říct, že je bootstrapping výpočetně náročnější, ale není u něj potřeba řešit problém ideální délky krácení pseudoreplikací a navíc je univerzálnější, vzhledem k požadavkům na vstupní populaci dat. Z těchto, ale i z dalších důvodů, dnes většina studií dává přednost bootstrappingu [37], [38], [43], [44].

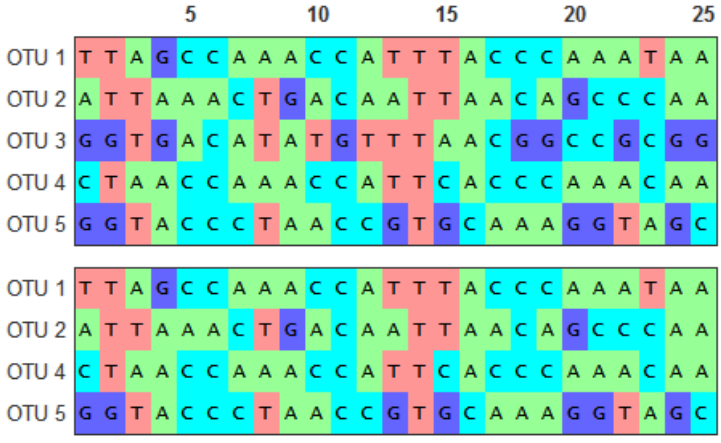


Obrázek 12: Ukázka dvou rozdílných resamplingových testů fylogenetické analýzy aplikovaných na stejné vstupní nukleotidové sekvence.

Poznámka: Uvedené stromy jsou konstruovány metodou neighbor-joining, distanční maticí je matice proporcionálních vzdáleností a vstupními sekvencemi jsou úseky 18s rRNA příslušných živočichů. Pro oba testy bylo použito 500 pseudoreplikací, u Jackknifingu byly sekvence zkráceny na 50 % své původní délky.

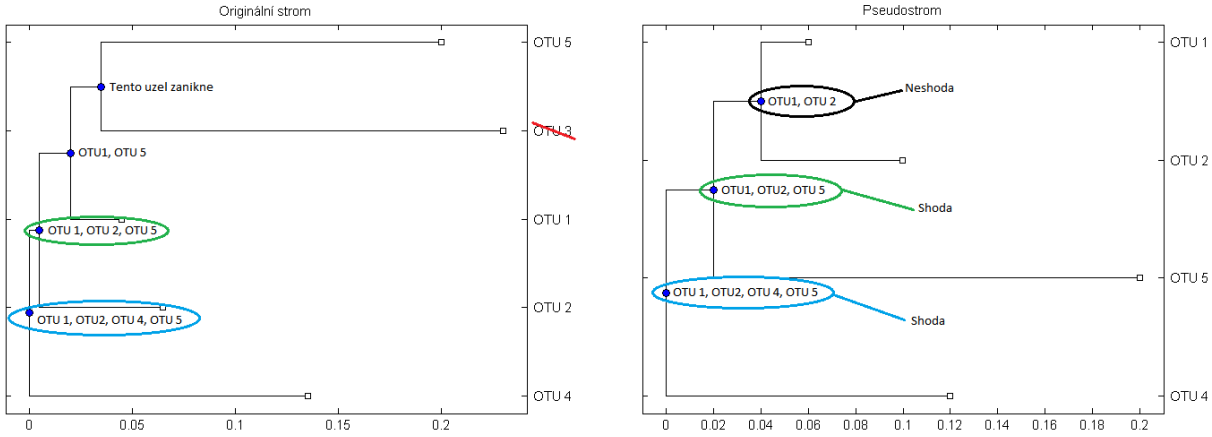
3.3. OTU jackknifing

Jak název napovídá, metoda vychází z principu jackknifingu. Nyní ovšem nedochází ke krácení původní délky sekvencí, ale v jednotlivých krocích odebíráme postupně vždy jeden řádek z originálního setu vstupních sekvenci. To znamená, že odebíráme v podstatě celou jednu větev, jednu taxonomickou jednotku, viz obrázek 13. Pro příklad jsme odebrali třetí OTU. Pořadí sloupců zůstává zachováno. Sekvence poté znovu zarovnáme.



Obrázek 13: Originální sekvence a jejich OTU jackknifingové pseudoreplikace.
Nahoře: originální sekvence. Dole: OTU jackknifingová pseudoreplikace při odebrání OTU 3.

Na základě takovéto pseudoreplikace sestrojíme pseudostrom (zde například metodou neighbor-joining, s použitím proporcionálních vzdáleností), který porovnáme s původním originálním stromem, obsahujícím všechny OTU, viz obrázek 14.

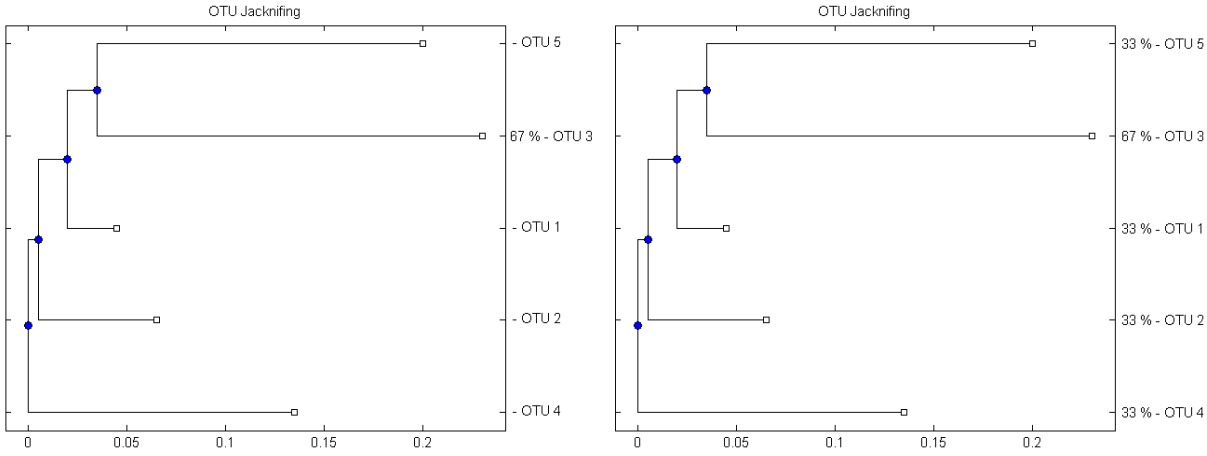


Obrázek 14: Srovnání originálního stromu s pseudostromem bez OTU 3.

Metodika srovnání je shodná s bootstrappingem i jackknifingem. Spočteme procentuální zastoupení shodných uzlů, tedy takových, které obsahují shodné OTU, při čemž nezaleží na pořadí větvení, viz rovnice (5). Příslušnou, v daném kroku odebranou, větev samozřejmě při srovnání s originálním stromem nebereme v úvahu:

$$\frac{\textbf{Počet shodných uzlů Pseudostromu}}{\textbf{Počet všech uzlů pseudostromu}} \cdot \mathbf{100}\ \% = \frac{\mathbf{2}}{\mathbf{3}} \cdot \mathbf{100}\ \% = \mathbf{67}\ \%. \qquad (5)$$

Hodnotu OTU jackknifingové podpory poté připíšeme k příslušné, v daném kroku odebrané, větvi v originálním stromě, viz obrázek 15 vlevo [12].



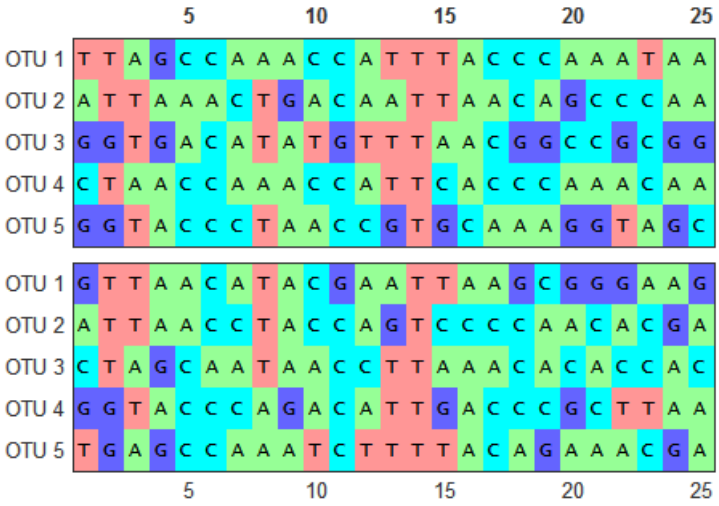
Obrázek 15: Vznik výsledného stromu u OTU jackknifingu.
Vlevo: Strom s OTU jackknifingovou podporou OTU 3. Vpravo: výsledný strom s OTU jackknifingovou podporou všech větví.

Interpretace výsledků

Interpretace této statistické metody je od předchozích odlišná. Hypotéza tohoto testu je založena na předpokladu, že odebrání jedné větve neovlivní topologii celého stromu, pokud je tento strom dostatečně robustní. Je zřejmé, že čím je procentuální hodnota OTU jackknifingové podpory vyšší, tím méně je strom danou větví ovlivněn, tedy tím více je strom robustnější. Například 100 % podpora větve by znamenala, že je originální strom a strom bez příslušné větve naprosto shodný (vyjma oné odebrané větve), topologie by se vůbec nezměnila. Metoda může být velmi užitečná pro odhalení artefaktů dlouhých větví probíraných v kapitole 2.4 [9], [12], [45].

3.4. PTP test

Poslední resamplingovou metodou je test permutation tail probability, zkráceně PTP test. Znovu vycházíme z originálních, pod sebou zarovnaných, sekvencí. Nyní však provedeme permutaci bez opakování všech znaků postupně v jednotlivých sloupcích data setu, viz obrázek 16.

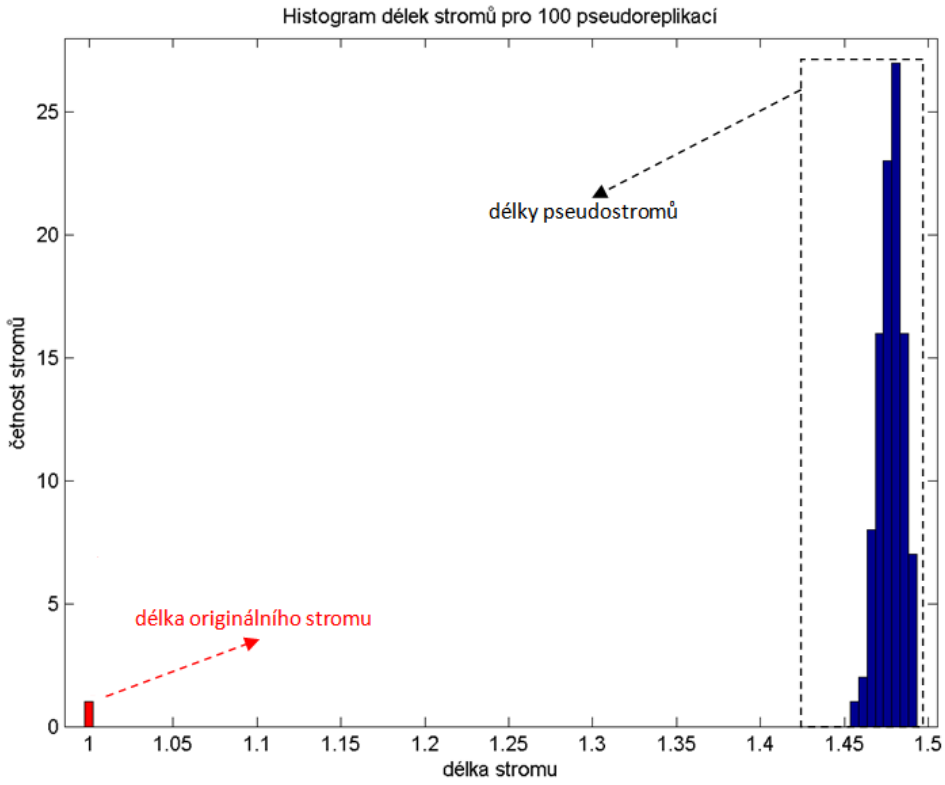


Obrázek 16: Originální sekvence a jejich PTP pseudoreplikace.
Nahoře: originální sekvence. Dole: PTP pseudoreplikace.

Tímto postupem v podstatě znehodnotíme informaci obsaženou v původním data setu. Na základě takovéhoto psedovzorku sestrojíme pseudostrom a spočteme jeho délku (sečteme délky všech jeho větví). Tímto způsobem sestrojíme a určíme délku alespoň 100 pseudostromů. Tato číslovka vychází ze stejné teorie, rozebírané v kapitole 3.1 Bootstrapping - přesnost výsledků a jejich interpretace. Zde se však spokojíme i s nižším počtem pseudoreplikací (v případě bootstrappingu to bylo 500), protože u PTP testu nepotřebujeme dosáhnout tak vysoké přesnosti. Jednotlivé délky pseudostromů jsou totiž od původní délky originálního stromu zpravidla velmi vzdáleny. Rovněž interpretace výsledků je odlišná. Navíc využíváme

permutaci u každého sloupce zvlášť, což činí metodu výpočetně náročnější než například u bootstrappingu, kde provedeme permutaci sice všech sloupců, ale pouze v jednom kroku.

V ideálním případě by délka originálního stromu měla být vždy kratší než délky všech pseudostromů. Pro názornost bývá výsledkem PTP testu histogram. Pro větší přehlednost je možné histogram normalizovat. Pokud vydělíme délku všech pseudostromů délkou originálního stromu, bude histogram ukazovat, kolikrát jsou pseudostromy delší v porovnání s originálním stromem. Ten bude mít délku vždy rovnu jedné, viz obrázek 17 [9], [12], [46].



Obrázek 17: Normalizovaný PTP histogram pro 100 pseudoreplikací.

Interpretace výsledků

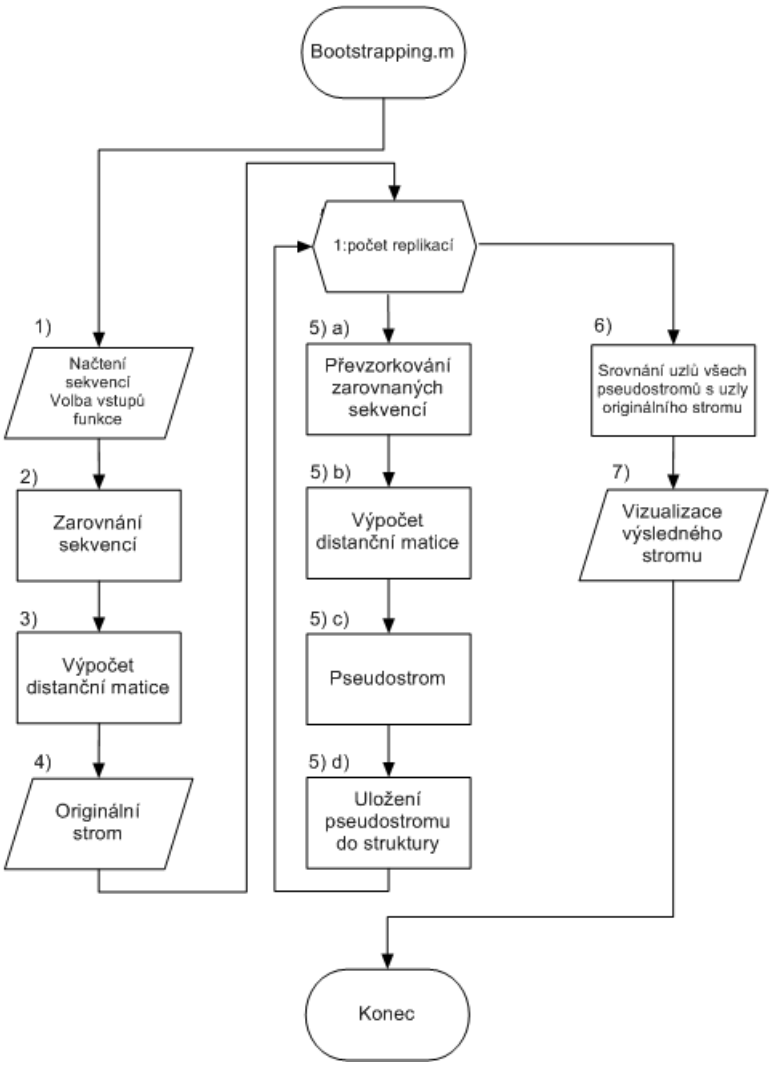
Předpokládáme, že originální fylogenetický strom sestrojený na základě našich dat není náhodný, ale je evolučně podložený za předpokladu, že evoluce probíhá vždy nejkratší cestou (obdobný předpoklad využívá například i konstrukční metoda minimum evolution, viz kapitola 2.4.2). Stejná množina dat znehodnocená permutacemi tedy musí dávat vždy horší výsledek. Výsledek interpretujeme tak, že data obsahují dostatečnou fylogenetickou informaci pro odhad evoluce, pokud alespoň 95 % pseudostromů má větší délku než originální strom [12], [46].

4. Realizace algoritmů resamplingových testů fylogenetické analýzy

Všechny resamplingové testy popsané v kapitole 3 byly realizovány v programovém prostředí Matlab s Bioinformatickým toolboxem jako samostatné funkce. Jako referenční data, která byla používána pro testování navržených algoritmů, jsme zvolili geny 18s rRNA deseti primátů. Ke konstrukci stromů byla vždy využita konstrukční metoda neighbor-joining. Vstupem všech funkcí mohou být aminokyselinové nebo nukleotidové sekvence (DNA nebo RNA) ve formátu FASTA (standardní formát dat pro práci v bioinformatice).

4.1. Bootstrapping

Funkce `Bootstrapping.m` pracuje dle následujícího schématu (obrázek 18):



Obrázek 18: Vývojový diagram funkce Bootstrapping.m

Vstupy funkce lze najít v tabulce 10. Výstupem je výsledný strom s uloženými hodnotami bootstrappingové podpory uzlů (obrázek 20) a dále pak vykreslený původní, originální strom (obrázek 19).

1) Načtení sekvencí, volba vstupů funkce

```
Bootstrapping(Seq,Dist_model,Subst_matice,PocetReplikaci)
```

Tabulka 10: Možnosti zadání vstupních argumentů funkce `Bootstrapping.m` včetně jejich vysvětlení.

Název vstupní proměnné	Možnosti zadání	Vysvětlení
Seq	Příklad použití: `'Sequence.fasta'`	Návez souboru ve formátu fasta, včetně přípony
Dist_model	`'p'`	Volba distančního modelu: Proporcionální
	`'jc'`	Volba distančního modelu: Jukes-Cantorův
	`'k'`	Volba distančního modelu: Kimurův
	`'t'`	Volba distančního modelu: Tamurův
	`'tn'`	Volba distančního modelu: Tamura-Neiův
	`'poisson'`	Poissonův model pro aminokyseliny
	`'gamma'`	Gamma model pro aminokyseliny
	`a`	Parametr řídící evoluční rychlost u gamma modelu
	`'jc_AK'`	Jukes-Cantorův model pro aminokyseliny
Subst_matice	`'BLOSUM62'`	Volba substituční matice pro aminokyseliny
	`'BLOSUM30' až 'BLOSUM90' s krokem 5`	
	`'BLOSUM100'`	
	`'PAM10' až 'PAM500' s krokem 10`	
	`'NUC44'`	Volba substituční matice pro nukleotidy
PocetReplikaci	Příklad použití: 500	Počet požadovaných pseudoreplikací

2) Zarovnání sekvencí

Sekvence jsou zarovnány pomocí funkce `multialign`, s nastavením požadované substituční matice.

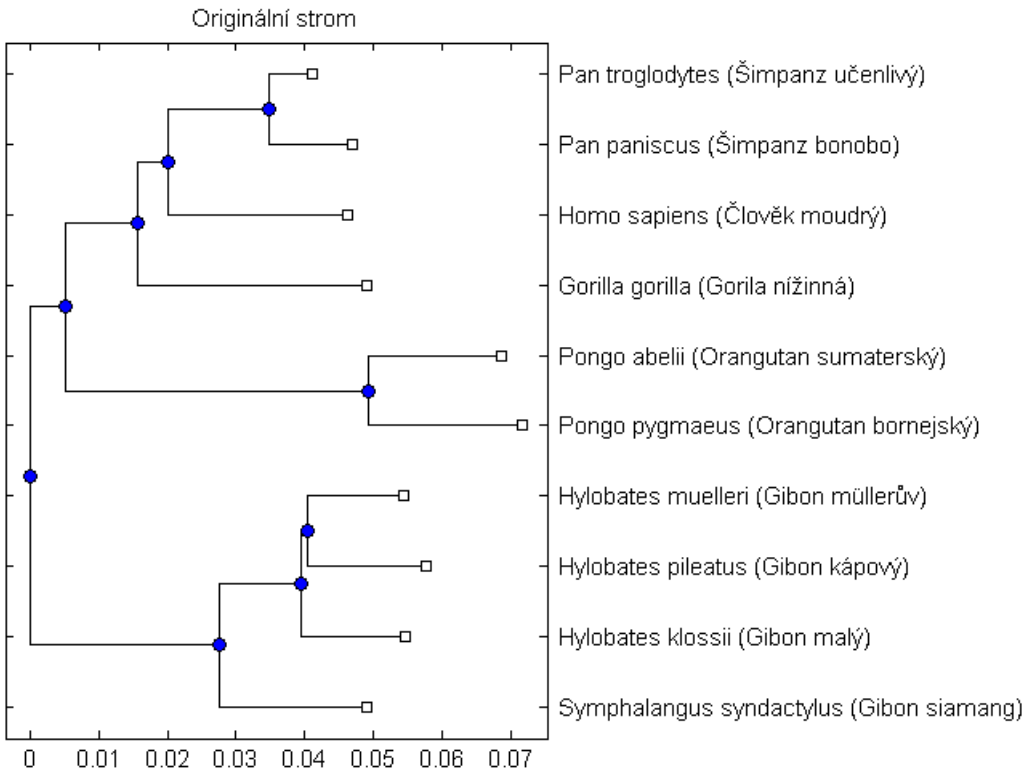
3) Výpočet distanční matice

Za tímto účelem byla vytvořena samostatná funkce `dist_model.m`, která dokáže spočítat evoluční vzdálenost dvojice sekvencí na základě požadovaného distančního modelu dle vzorců uvedených v tabulkách 5 a 6. Jednotlivé hodnoty vzdáleností jsou pomocí for cyklu řazeny do odpovídající distanční matice.

4) Originální strom

Hodnoty distanční matice je třeba nejprve přepsat po řádcích na vektor. Tento vektor je pak použit jako vstup funkce `seqneighjoin`, spolu s argumentem

`'equivar'`, což odpovídá metodě Neighbor-Joining s modifikací dle Studiera-Kepplera. Vytvoří se speciální typ objektu phytree. Pomocí funkce plot je poté originální strom vykreslen, viz obrázek 19 [32].



Obrázek 19: Originální fylogenetický strom.

5) For cyklus funkce Bootstrapping.m

a) Převzorkování zarovnaných sekvencí

Bootstrappingové pseudoreplikace sekvencí jsou vytvořeny na základě původního setu zarovnaných sekvencí `DataSeq` z kroku 2). Příkaz `pom = DataSeq (:,randsample (1:delkas,delkas,'true'))` vytvoří požadované zpřeházení sloupců s možným opakováním a uloží jej pod pomocnou proměnnou `pom`.

b) Výpočet distanční matice

Výpočet je shodný s krokem 3) až na to, že tentokrát počítáme distance v novém pseudoreplikovaném data setu `pom`.

c) Pseudostrom

Vytvoření matlabovského objektu typu phytree je shodné s krokem 4). Vychází z distanční matice obdržené v kroku 5) b). Nyní ovšem obdržený objekt phytree nevykreslujeme.

d) Uložení pseudostromu do struktury

Pseudostrom z kroku 5) c) si uložíme do struktury `Pseudo_tree`. Veškeré úkony 5) a) - 5) d) jsou opakovány ve for cyklu, při čemž se veškeré proměnné v nich vytvořené neustále přepisují, vyjma struktury `Pseudo_tree`. Ta se neustále rozšiřuje o nové pseudostromy v závislosti na uživatelem zvoleném počtu replikací.

6) Srovnání uzlů všech pseudostromů s uzly originálního stromu

Originální strom je nejprve rozložen na všechny své podmnožiny (podstromy, jejichž kořeny jsou postupně všechny uzly původního stromu) pomocí funkce `subtree`. Z těchto podstromů získáme názvy listů příslušných uzlů a ty pak seřazené podle abecedy uložíme do struktury `Leaf_Origin` pod příslušným indexem `i` (index koresponduje s pořadím uzlu originálního stromu). Stejný postup aplikujeme také na všechny pseudostromy a jejich uzly uložíme do struktury `Leaf_pseudo`, tentokrát pod indexy `j` a `i`, kde `j` určuje pořadí pseudostromu a `i` určuje pořadí uzlu daného speudostromu.

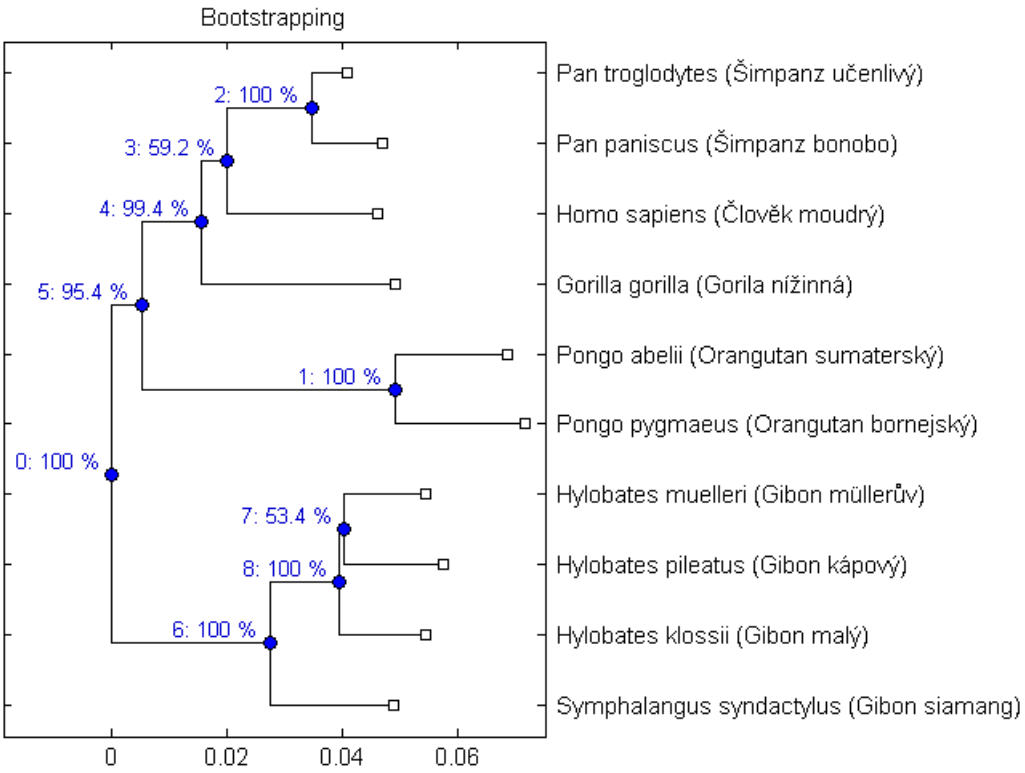
Samotné srovnání spočívá v porovnání struktur `Leaf_Origin` a `Leaf_pseudo` pomocí dvojitého for cyklu:

```
for i = 1 : PocetVstupu-1
    for j = 1 : PocetReplikaci*(PocetVstupu-1)
            if isequal(Leaf_Origin{i},Leaf_Pseudo{j})
                srovnani(i) = srovnani(i) + 1;
            end
    end
end
```

Pokud nastane shoda listů uzlů, k proměnné `srovnani` se přičte jednička na pozici pod indexem i, která koresponduje s pořadovým číslem uzlu originálního stromu. Nakonec tedy dostaneme vektor `srovnani`, kde na pozici každého uzlu originálního stromu bude spočten výskyt shodného uzlu u všech vytvořených pseudoreplikací.

7) Vizualizace výsledného stromu

Nejprve si spočteme procentuální zastoupení jednotlivých uzlů v pseudostromech podělením hodnot vektoru `srovnani` počtem replikací a vynásobením 100. Poté změníme originálnímu stromu názvy uzlů (`'NODENAMES'`), kde připíšeme procentuální zastoupení příslušných uzlů a nakonec vykreslíme originální strom s povoleným zobrazením názvu uzlů, viz obrázek 20.



Obrázek 20: Originální fylogenetický strom s hodnotami bootstrappingové podpory uzlů.

4.2. Jackknifing

Funkce `Jackknifing.m` pracuje dle shodného schématu jako bootstrapping, viz kapitola 4.1, obrázek 18. Liší se v podstatě jen v krocích 1) a 5) a), další kroky jsou naprosto shodné s bootstrappingem. Výsledný strom s vyznačenou jackknifingovou podporou uzlů je na obrázku 21.

1) Načtení sekvencí, volba vstupů funkce

```
Jackknifing(Seq,Dist_model,Subst_matice,PocetReplikaci,zkraceni)
```

Oproti bootstrappingu přibyl vstup funkce `zkraceni`. Jedná se o číselnou hodnotu udávající požadované zkrácení původní délky zarovnaného data setu sekvencí nukleotidů (aminokyselin). Nabývá hodnot od 0 do 1. Například zkrácení 0,45 tedy zkrátí původní data set na 45 % své původní délky. Ostatní vstupy jsou shodné s bootstrappingem, viz tabulka 10.

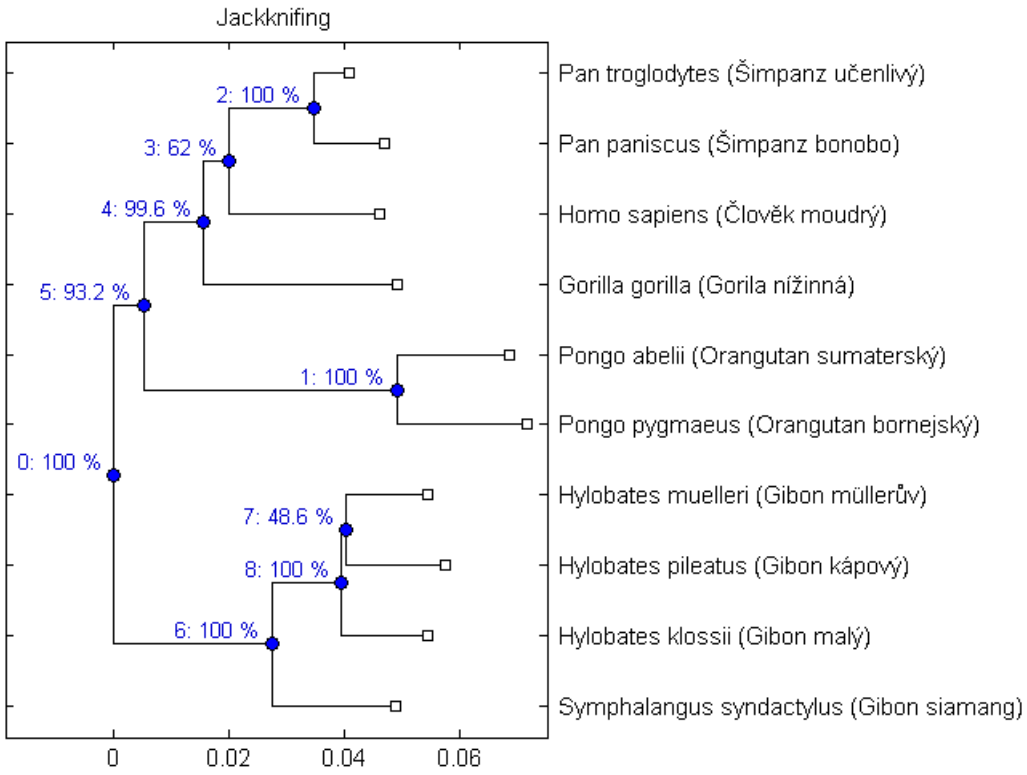
5) For cyklus funkce Bootstrapping.m

a) Převzorkování zarovnaných sekvencí

Před samotným převzorkováním data setu zarovnaných sekvencí dochází k jejich zkrácení. Původní sekvence se nejprve uloží pod pomocnou proměnnou `pom`. Poté se

pomocí funkce `randi` ve for cyklu vymaže zvolený počet sloupců v data setu. Takto zkrácený data set převzorkujeme příkazem `pom = pom(:,randsample(1:delkas,delkas))`. Vytvoří se tak požadované zpřeházení sloupců ve zkráceném data setu bez opakování.

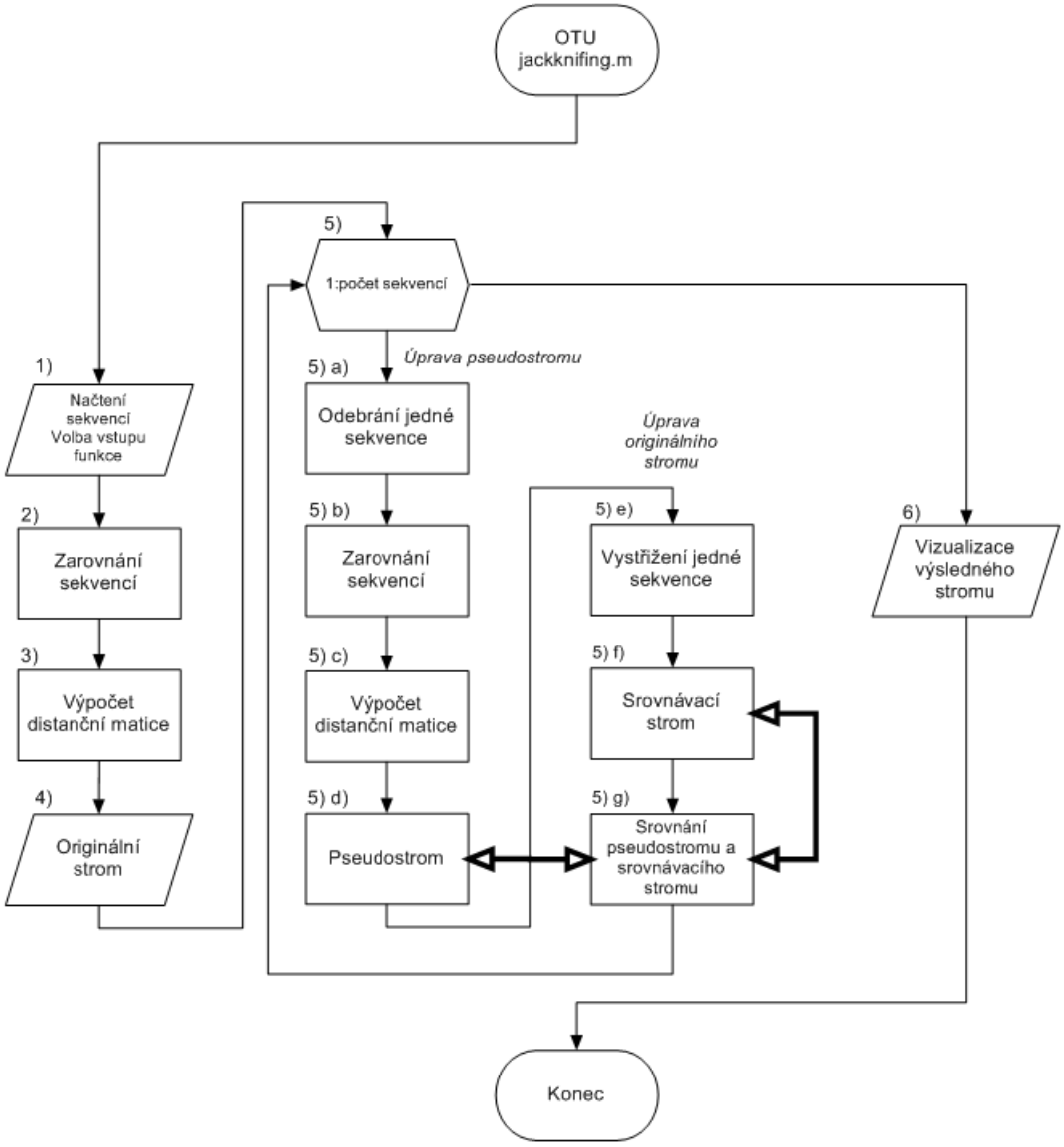
7) Vizualizace výsledného stromu



Obrázek 21: Originální fylogenetický strom s hodnotami jackknifingové podpory uzlů.

4.3. OTU jackknifing

Funkce `OTU_Jackknifing.m` pracuje dle následujícího schématu (obrázek 22). Funkční bloky zarovnání sekvencí, výpočet distanční matice a vizualizace originálního stromu jsou shodné s funkcí `bootstrapping.m`, viz kapitola 4.1, kroky 2) až 4). Pokračovat tedy budeme popsáním vstupů funkce a dále pak vysvětlením for cyklu funkce `OTU_Jackknifing.m`.



Obrázek 22: Vývojový diagram funkce OTU_Jackknifing.m

1) Načtení sekvencí, volba vstupů funkce

```
OTU_Jackknifing(Seq,Dist_model,Subst_matice,PocetReplikaci)
```

Vstupy jsou shodné s funkcí bootstrapping.m, viz tabulka 10.

5) For cyklus funkce OTU_Jackknifing

a) Odebrání jedné sekvence

V jednotlivých krocích for cyklu dojde k vymazání postupně vždy jednoho řádku vstupního data setu zarovnaných sekvencí. Takto upravený data set se uloží pod proměnnou `pom`.

b) Zarovnání sekvencí

Pomocná proměnná `pom` s upraveným data setem sekvencí se následně zarovná pomocí funkce `multialign`, s nastavením požadované substituční matice.

c) Výpočet distanční matice

Samostatná funkce `dist_model.m`, spočítat evoluční vzdálenost dvojic sekvencí pomocné `pom` na základě požadovaného distančního modelu dle vzorců uvedených v tabulkách 5 a 6. Jednotlivé hodnoty vzdáleností jsou pomocí for cyklu řazeny do odpovídající distanční matice.

d) Pseudostrom

Hodnoty distanční matice z kroku 5) c) je třeba nejprve potřeba přepsat po řádcích na vektor. Tento vektor je pak použit jako vstup funkce `seqneighjoin`, spolu s argumentem `'equivar'`. Ten udává konstrukční metodu stromu neighbor-joining. Vytvoří se speciální typ objektu phytree.

e) Vystřižení jedné sekvence

Dojde k vystřižení té větve originálního stromu, která byla v daném kroku for cyklu vymazána v bodě 5) a). Zbytek originálního stromu zůstává beze změny, čehož docílíme použitím funkce `prune`.

f) Srovnávací strom

Vzniklý srovnávací strom je nejprve rozložen na své podstromy a jednotlivé větve seřazené podle abecedy každého uzlu jsou uloženy do struktury `Leaf_Origin`, podobně jako je popsáno u kapitoly 4.1 Bootstrapping v kroku 6.

g) Srovnání pseudostromu a srovnávacího stromu

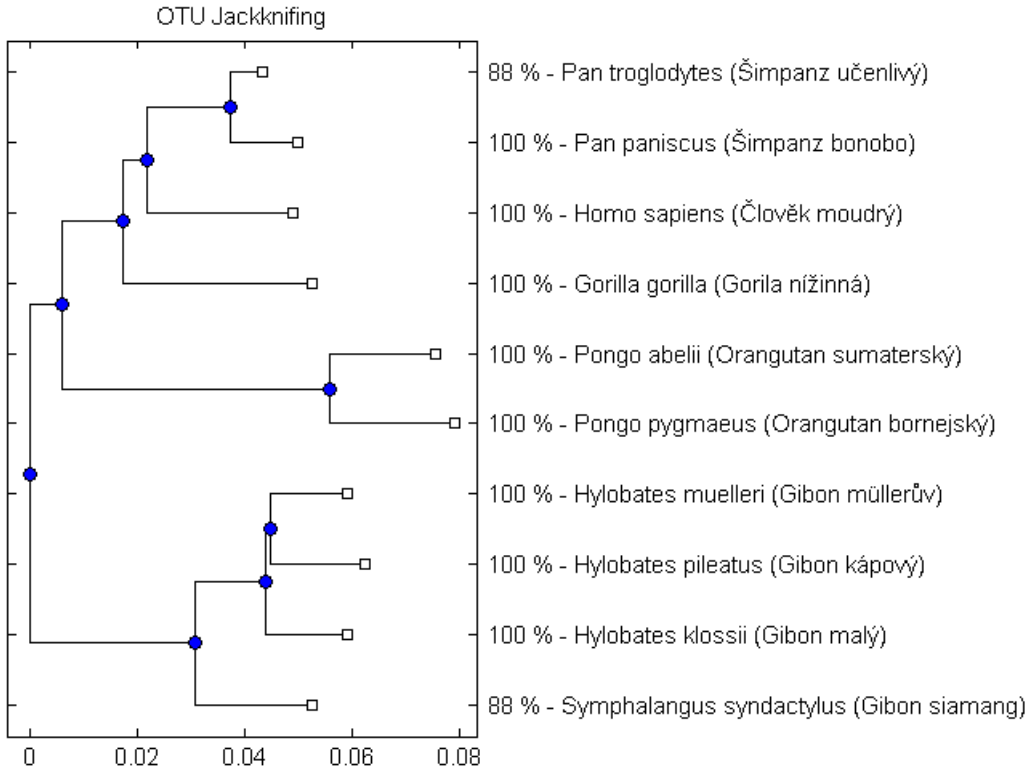
Stejným způsobem vytvoříme strukturu `Leaf_Pseudo`, která obsahuje uzly pseudostromu vytvořeného v kroku 5) d). Následně dojde ke srovnání vytvořených struktur `Leaf_Origin` a `Leaf_Pseudo` pomocí dvojitého for cyklu:

```
for j=1:PocetVstupu-2
        for k=1:PocetVstupu-2
            if isequal(Leaf_Origin(1,j),Leaf_Pseudo(1,k))
               srovnani(x)=srovnani(x)+1;
            end
        end
end
```

Pokud nastane shoda listů uzlů, k proměnné `srovnani` se přičte jednička na pozici pod indexem x, která koresponduje s pořadovým číslem odebrané větve v daném kroku for cyklu. Na konci celého for cyklu tedy dostaneme vektor `srovnani`, kde na pozici každé větve originálního stromu bude spočten výskyt shodných uzlů příslušných pseudostromů a srovnávacích stromů.

6) Vizualizace výsledného stromu

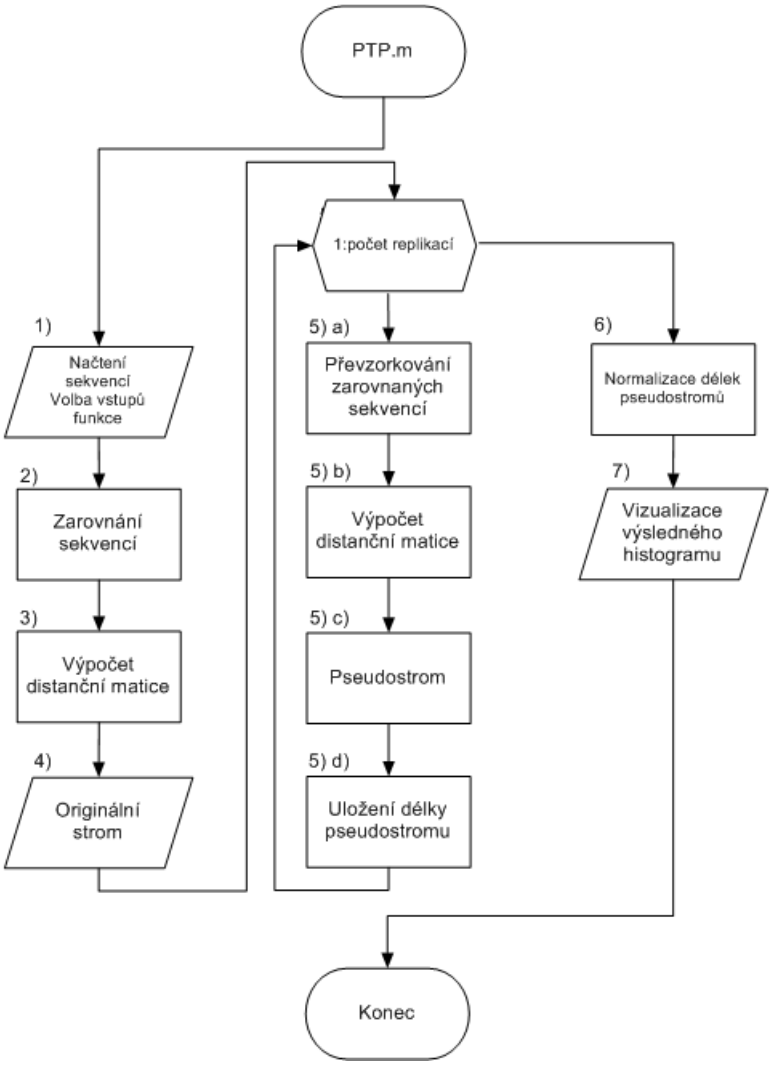
Nejprve si přepočítáme zastoupení shodných uzlů ve vektoru srovnani na procentuální hodnoty a ty uložíme jako novou proměnnou `zastoupeni`. Poté změníme originálnímu stromu názvy větví, kde připíšeme procentuální zastoupení příslušných větví a nakonec vykreslíme originální strom s novými názvy větví, viz obrázek 23.



Obrázek 23: Originální fylogenetický strom s hodnotami OTU jackknifingové podpory větví.

4.4. PTP

Funkce `PTP.m` pracuje dle podobného schématu jako bootstrapping, viz kapitola 4.1. Vstupy obou těchto funkcí jsou shodné a lze je nalézt v tabulce 10. Také převzorkování zarovnaného data setu sekvencí probíhá naprosto shodně, tedy zpřeházením jednotlivých sloupců s opakováním. Rozdíl je však v bodě 5) d), 6) a 7), viz obrázek 24.



Obrázek 24: Vývojový diagram funkce PTP.m

5) For cyklus funkce OTU_Jackknifing

d) Uložení délky pseudostromu

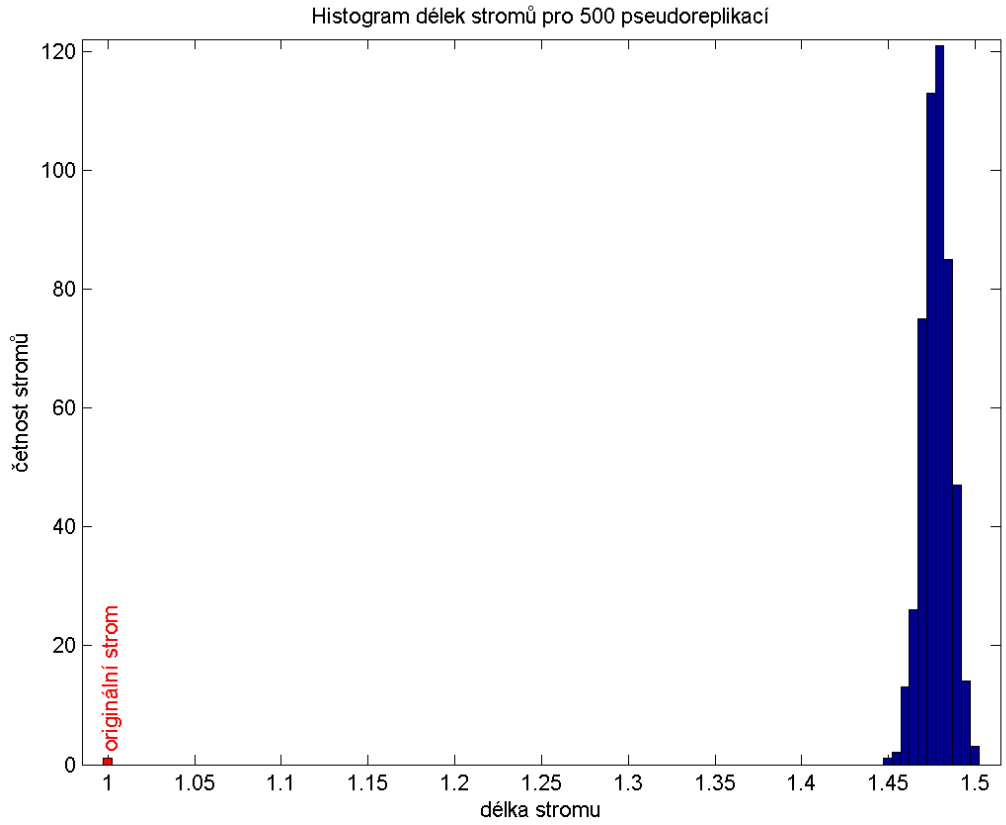
Zde ukládáme pouze informaci o celkové délce pseudostromu (součet délek všech jeho větví).

6) Normalizace délek pseudostromů

V bodě 6 vydělíme délky všech pseudostromů délkou originálního stromu. Dostaneme tak normalizované délky stromů, které udávají, kolikrát jsou delší než originální strom. Ten bude mít logicky normalizovanou délku vždy 1.

7) Vizualizace výsledného histogramu

Nakonec necháme vykreslit histogram normalizovaných délek stromů společně s označeným originálním stromem, viz obrázek 25.

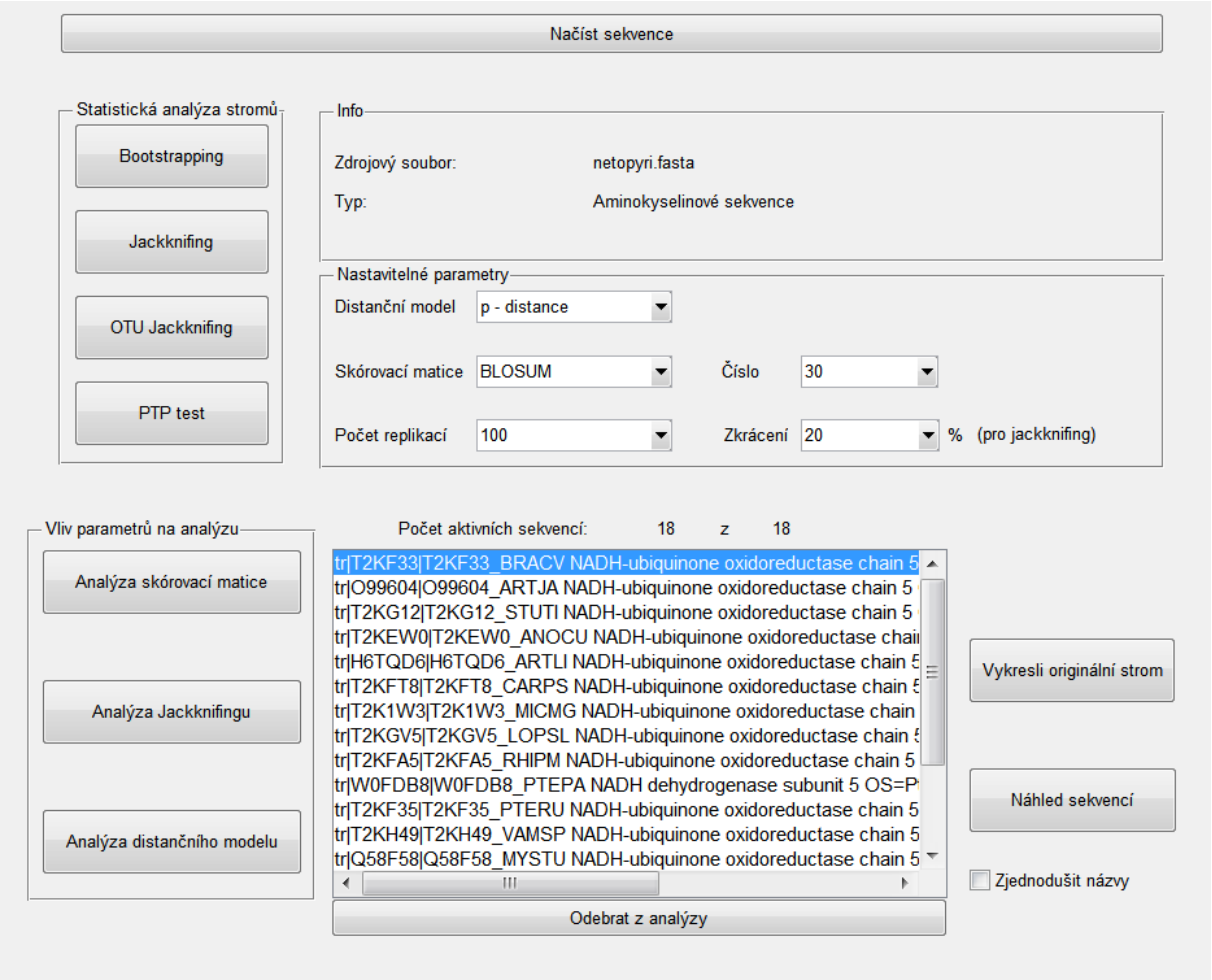


Obrázek 25: Výsledek PTP testu - normalizovaný histogram délek stromů.

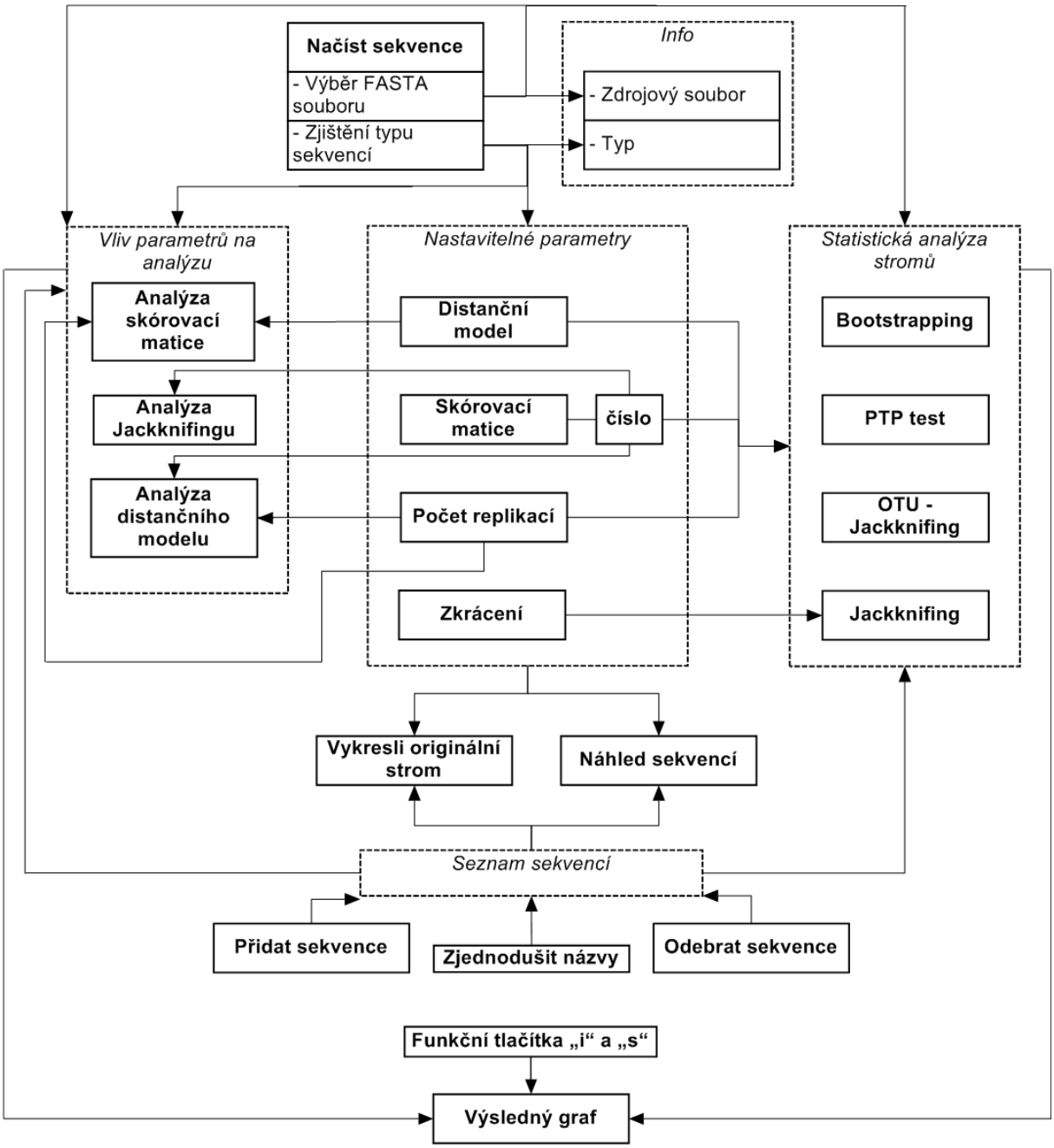
4.5. Programové rozhraní pro analýzu resamplingových testů

Za účelem rychlého, snadného a přehledného hodnocení a zpracování výsledků resamplingových testů bylo vytvořeno uživatelské programové rozhraní, umožňující ze souboru biologických sekvencí ve FASTA (standardní formát dat pro práci v bioinformatice) kódu vykreslit a analyzovat fylogenetický strom s nastavitelnými parametry v programovém prostředí Matlab s bioinformatickým toolboxem. Veškeré fylogenetické stromy, či analýzy založené na nich, jsou konstruovány pomocí metody Neighbor-Joining.

Program v podstatě sdružuje funkce předchozích podkapitol 4.1 - 4.4. Interface programu můžeme vidět na obrázku 26. Schéma propojení jednotlivých ovládacích prvků programu je naznačeno blokovým schématem na obrázku 27.



Obrázek 26: Interface programu pro analýzu resamplingových testů.



Obrázek 27: Schéma propojení jednotlivých ovládacích prvků programu.

Načtení sekvencí

Po spuštění programu musíme nejprve načíst sekvence ve FASTA kódu. Po stisknutí tlačítka „načíst sekvence“ se otevře nové okno, které umožňuje vybrat soubor uložený kdekoli na pevném disku počítače. Teprve po otevření tohoto souboru se zobrazí celý interface programu jako na obrázku 26.

Panel info

V panelu info vidíme název souboru, který je načten a také informaci o tom, zda se jedná o aminokyselinové nebo nukleotidové sekvence.

Panel nastavitelné parametry

Panel nastavitelné parametry umožňuje vybrat distanční model, skórovací matici a počet replikací, které mají být realizovány. Nabídka distančních modelů a skórovacích matic se mění v závislosti na tom, zda jsme načetli aminokyselinové nebo nukleotidové sekvence a je shodná s údaji uvedenými v kapitolách 2.2 a 2.3. Pokud zvolíme model gamma pro aminokyseliny, program umožní výběr gamma parametru řídícího evoluční rychlost. Posledním nastavitelným údajem je zkrácení, které udává o jakou procentuální část si přejeme zkrátit zarovnané sekvence u jackknifingového testu.

Seznam načtených sekvencí

Ve spodní části programu se nachází posuvný seznam načtených sekvencí. Pro analýzy založené na odebírání a zkoumání vlivu jednotlivých větví na resamplingové testy program umožňuje jejich odebrání z analýzy (případně zpětné přidávání do analýzy). Program také umožňuje zjednodušit názvy sekvencí pomocí zaškrtávacího pole „zjednodušit názvy“. Zkrácení je založeno na principu vyhledání pozice znaků „OS=“ nebo posledního znaku „|“. Od následujícího znaku se začnou zapisovat jednotlivá písmena až do druhé nalezené mezery. Takovýto zápis totiž odpovídá standardizovanému zápisu názvů sekvencí stažených z databází: http://www.ncbi.nlm.nih.gov/genbank/ nebo http://www.uniprot.org/. Před zkrácený zápis se ještě doplní pořadová čísla sekvencí, aby bylo zabráněno případnému vzniku duplicitních názvů větví. Pokud má uživatel sekvence stažené z jiných databází nebo nebude se zjednodušeným zápisem spokojen, má možnost ohraničit úsek názvu sekvencí, který má být zobrazen, do složených závorek {úsek}. Je však nutno takto ohraničit úsek u všech sekvencí daného FASTA souboru.

Napravo od seznamu sekvencí se nacházejí dva funkční tlačítka. Tlačítko „vykresli originální strom“ vykreslí fylogenetický strom na základě nastavených parametrů metodou Neighbor-Joining. Tlačítko náhled otevře matlabovskou aplikaci pro náhled zarovnaných sekvencí, ze kterých budou všechny analýzy vycházet. Tento náhled může být užitečný pro jednoduché hodnocení podobností sekvencí a také pro hodnocení množství substituovaných pozic.

Panel statistická analýza stromů

Panel statistická analýza stromů nabízí realizaci resamplingových testů na základě funkcí uvedených v kapitole 4. Výsledný strom (či histogram u PTP testu) se otevře v novém okně a zůstává otevřen dokud jej sám uživatel nezavře. To umožňuje mít najednou spuštěné výsledky analýz s různými parametry či vstupními sekvencemi.

Funkční klávesy „i“ a „s“

Pokud má uživatel aktivní okno s výsledkem analýzy a stiskne tlačítko „i“, zobrazí se mu podrobné informace o dané analýze. Při opětovném stisknutí tlačítka „i“ podrobnosti zmizí. Klávesa slouží k lepší orientaci v mnoha otevřených výsledcích analýz.

Klávesa „s“ slouží k rychlému uložení výsledku jako obrázku ve formátu png. Soubor se uloží do adresáře spuštěného programu. Pro snadnou orientaci jsou v názvu souboru obsaženy informace o dané analýze ve formátu:

název souboru _ název resamplingového testu a počet replikací _ (jackknifingové zkrácení) _ konstrukční metoda _ zkratka použitého distančního modelu _ skórovací matice a její číslo _ pořadové číslo obrázku se shodným názvem.png

Příklad názvu obrázku:

netopyri.fasta_bootstrapping100_N-J_jc_AK_BLOSUM50(0).png

Panel vliv parametrů na analýzu

Tento panel nabízí tři analýzy pro aminokyselinové sekvence a dvě pro nukleotidové (zde je analýza skórovací matice bezpředmětná, jelikož existuje pouze jedna, NUC44). Výsledkem všech analýz je vždy graf srovnávající vliv daných parametrů na výsledek fylogenetického testu. Před spuštěním analýzy je dobré mít na paměti, že se jedná o časově náročné výpočty, závisle na počtu požadovaných replikací a množství a délce vstupních sekvencí.

Analýza skórovací matice zkoumá závislost průměrné hodnoty bootstrappingu na skórovací matici BLOSUM a PAM. Uživatel volí pouze počet replikací a distanční model.

Analýza jacknifingu zkoumá vliv počtu replikací a zkrácení na hodnotu jackknifingu. Volíme tedy pouze distanční model a skórovací matici.

Analýza distančního modelu ukazuje závislost bootstrappingové podpory na volbě distančního modelu. Uživatel volí pouze skórovací matici a počet replikací.

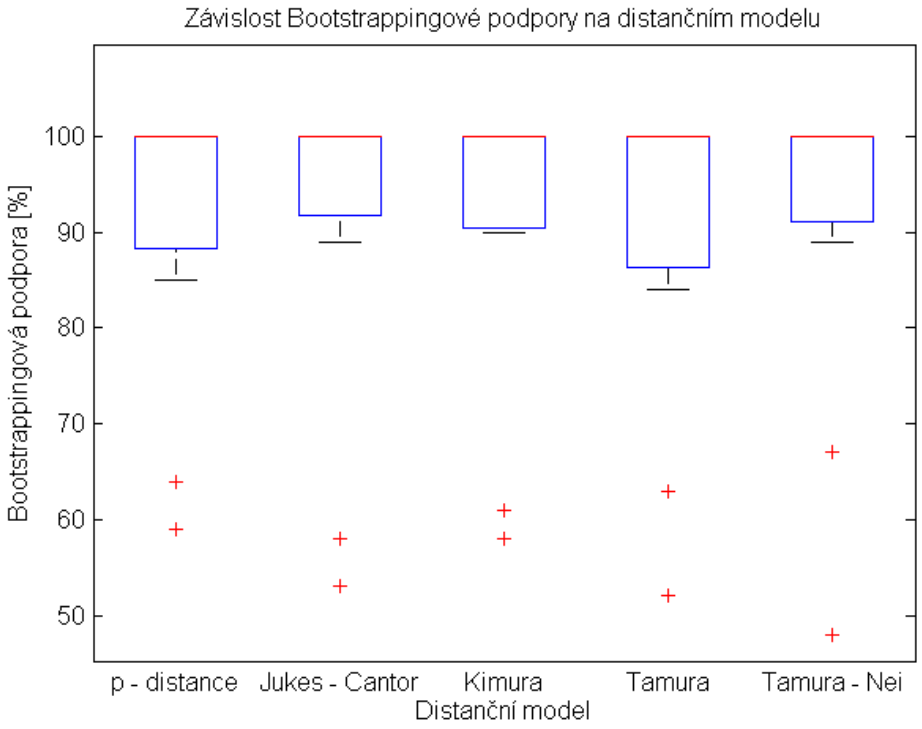
5. Analýza resamplingových testů

Resampligové testy představené a realizované v této práci jsou známou problematikou a jejich aplikace na verifikaci správnosti fylogenetického stromu by měla být vždy nezbytnou součástí samotné konstrukce. V teoretických popisech metod však často chybí vysvětlení, co které metody testují a jak je možné zkvalitnit výsledek volbou parametrů fylogenetické analýzy.

V této poslední kapitole jsme tedy využili realizované programové rozhraní pro konstrukci a vyhodnocení správnosti fylogenetického stromu k analýze vlivu změny parametrů fylogenetické analýzy na výsledek statistických testů. Testovali jsme vliv distančního modelu a skórovací matice na výsledek bootstrappového testu, vliv výběru a počtu sekvencí na hodnotu PTP testu a nakonec jsme se zaměřili na určení správného počtu replikací a zkrácení u jackknifingového testu.

5.1. Vliv distančního modelu na hodnotu bootstrapové podpory

Abychom mohli hodnoty bootstrapové podpory pro různé modely snadno a přehledně porovnat, zvolili jsme vykreslení procentuelních hodnot podpory pro dané modely do formy krabicových grafů (box plotů). Jako vstupní data jsme zvolili různé typy nukleotidových nebo aminokyselinových sekvencí. Bootstrappingový test byl prováděn s počtem replikací 500 a skórovací maticí byla NUC44 nebo BLOSUM 50. Nejprve jsme zkoumali vliv na nukleotidové sekvence, které jsou homologní, bez výrazných substitucí, delecí a inzercí a jsou pro fylogenetickou analýzu velmi vhodné. Jedná se fylogenetický marker, gen 18s rRNA pro skupinu dvanácti primátů. Výsledný graf vidíme na obrázku 28 [20], [21], [22].

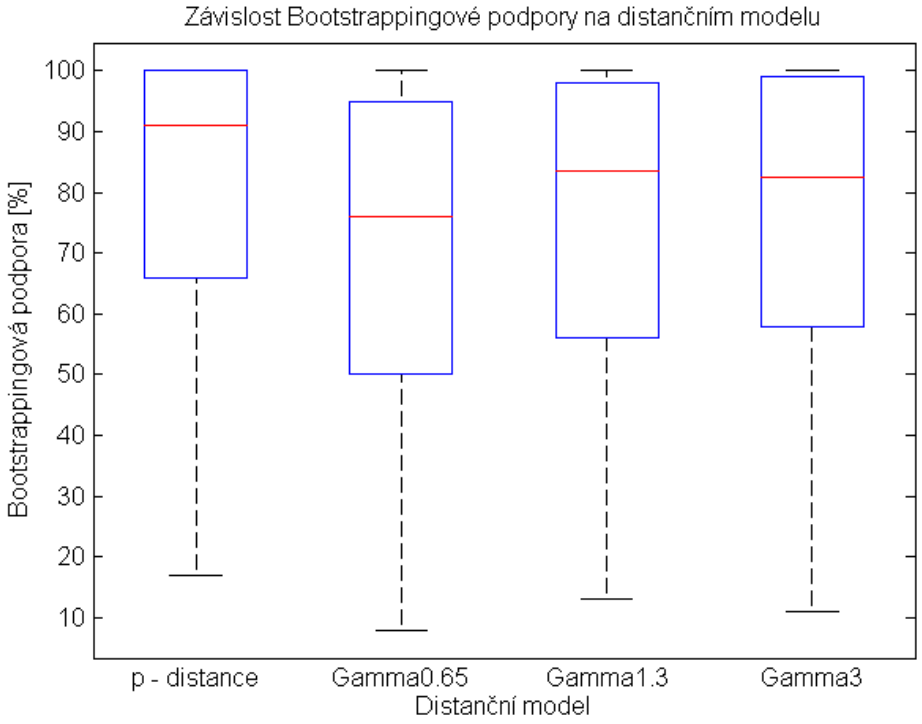


Obrázek 28: Závislost bootstrappingové podpory na distančním modelu pro nukleotidové sekvence genu 18s rRNA pro skupinu dvanácti primátů.

Hodnota mediánu zůstává u všech modelů na 100 %. Spodní hranice kvartilu je vždy okolo 90 %. Hodnoty minima a maxima jsou rovněž srovnatelné. Výraznější rozdíly lze spatřit u hodnot odlehlých bodů. Pravděpodobně se u každé srovnávané skupiny jedná o shodné dva uzly, které jsou volbou distančního modelu výrazněji ovlivněny. Jejich podpora je však vždy poměrně nízká. V celkovém kontextu se tedy zdá, že volba distančního modelu nemá v tomto případě na bootstrappingovou analýzu výrazný vliv.

Pro podporu této hypotézy jsme udělali obdobnou analýzu pro aminokyselinové sekvence. Tentokrát jsme ale záměrně vybrali sekvence, které nejsou pro fylogenetickou analýzu vůbec vhodné. Jsou sice homologní, použili jsme však dvě rozdílné isoformy téhož proteinu - superoxidu dismutázy. Bylo použito 33 sekvencí různých živočichů isoformy SOD 1 a 14 isoformy SOD 3. Díky tomu se u zarovnaných sekvencí objevuje mnoho delecí a substitucí. Zajímalo nás, jaký bude mít tento fakt vliv na naši analýzu. Obdrželi jsme dva výsledné grafy 29 a 30.

Obrázek 29: Závislost bootstrappingové podpory na distančním modelu pro aminokyselinové sekvence superoxid dismutázy isoforem SOD1 a SOD 3.



Obrázek 30: Závislost bootstrappingové podpory na hodnotě gamma parametru pro aminokyselinové sekvence superoxid dismutázy isoforem SOD1 a SOD 3.

Nejhorších výsledků, co se týče mediánu, jsme dosáhli při použití gamma modelu s parametrem 0,65. Tento parametr je záměrně nastaven na nejmenší možnou hodnotu a odpovídá velmi rozdílným sekvencím. Změna však ani zde a ani u nejvyšší možné hodnoty s pa-

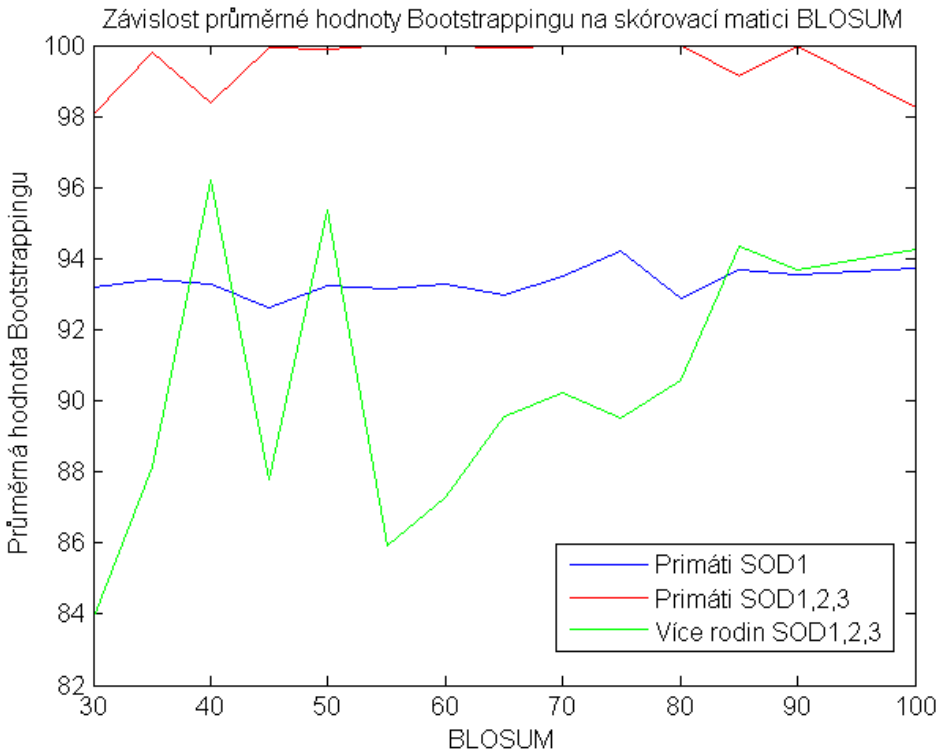
rametrem 3 není velká. Výsledky ukazují, že ani u těchto aminokyselinových sekvencí neměly změny modelu výrazný vliv na celkovou bootstrappingovou podporu.

Shrnutí

Volba správného distančního modelu je bezesporu velmi důležitým předpokladem pro konstrukci správného fylogenetického stromu. Ovlivňuje však hlavně samotnou délku větví, respektive celého stromu. Má tedy vliv na odhad výsledné evoluční evzdálenosti organismů ve stromu, samotnou klasifikaci do shluků ale neovlivní. Celková hodnota bootstrapové podpory se tak podle uvedených zjištění při použití aminokyselinových i nukleotidových sekvencí příliš nemění.

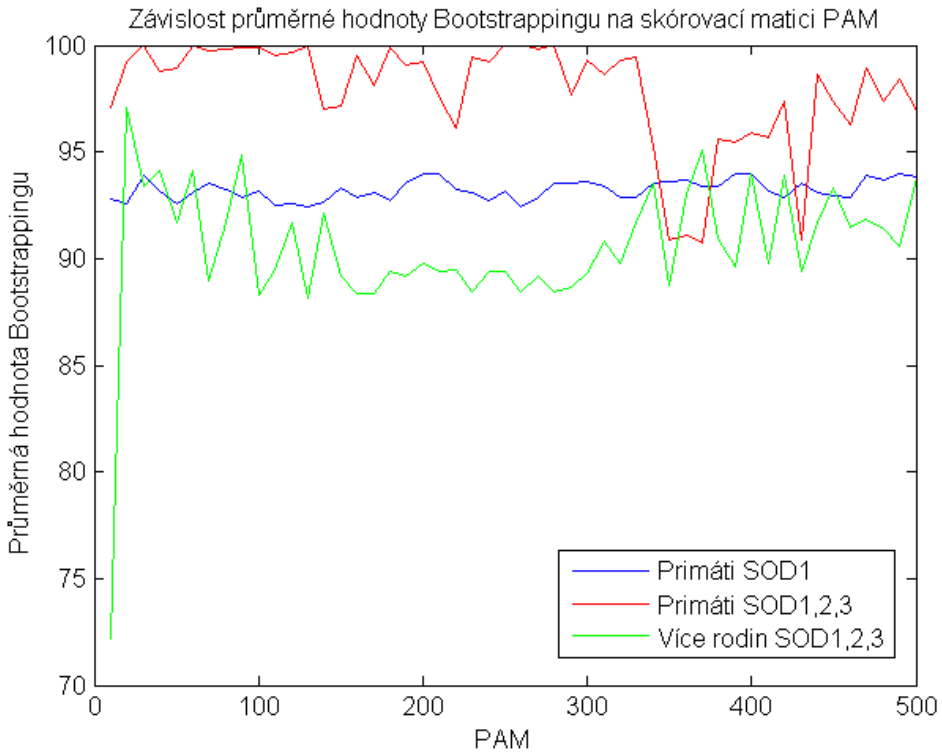
5.2. Vliv skórovací matice na hodnotu bootstrapové podpory

Vliv jsme v tomto případě zkoumali jen u aminokyselinových sekvencí, které umožňují volbu skórovací matice. Pro srovnání vlivu na různé typy sekvencí byly použity sekvence proteinu superoxid dismutázy tří isoforem SOD 1, 2 a 3 vždy pro skupinu sedmi živočichů stejného nebo různého živočišného řádu. Pro výpočet distanční vzdálenosti byl zvolen model p - distance, počet replikací byl 500. Výsledné grafy 31 a 32 ukazují závislost průměrné hodnoty bootstrappingu na skórovacích maticích BLOSUM a PAM.



Obrázek 31: Závislost průměrné hodnoty bootstrappingu na skórovací matici BLOSUM pro různé typy sekvencí.

Nejvyšších hodnot bootstrappingu jsme dosáhli při použití sekvencí jednoho řádu a všech tří isoforem (primáti SOD 1, 2 , 3). Průměrná podpora se pohybuje nad 98 % a je velice stálá. Podobně vyrovnaný charakter průběhu mají i sekvence stejného řádu i isoformy (primáti SOD 1). V tomto případě však byla průměrná hodnota nižší a pohybovala se v rozmezí 92 % až 93 %. Velmi nevyrovnaný průběh jsme zaznamenali u poslední zkoumané skupiny sekvencí, obsahující aminokyseliny šesti živočišných řádů a všech tří isoforem (více rodin SOD 1, 2, 3). Hodnota kolísala v rozmezí 84 % až 96 %. Zajímavé je, že se leckdy i sousední matice BLOSUM liší o téměř 10 %, viz obrázek 31 pro hodnoty BLOSUM 40 a 45. Je to způsobeno jednoznačně nesprávným zarovnáním vstupních sekvencí a ukazuje na významný vliv volby skórovací matice na hodnotu bootstrappingové podpory.



Obrázek 32: Závislost průměrné hodnoty bootstrappingu na skórovací matici PAM pro různé typy sekvencí.

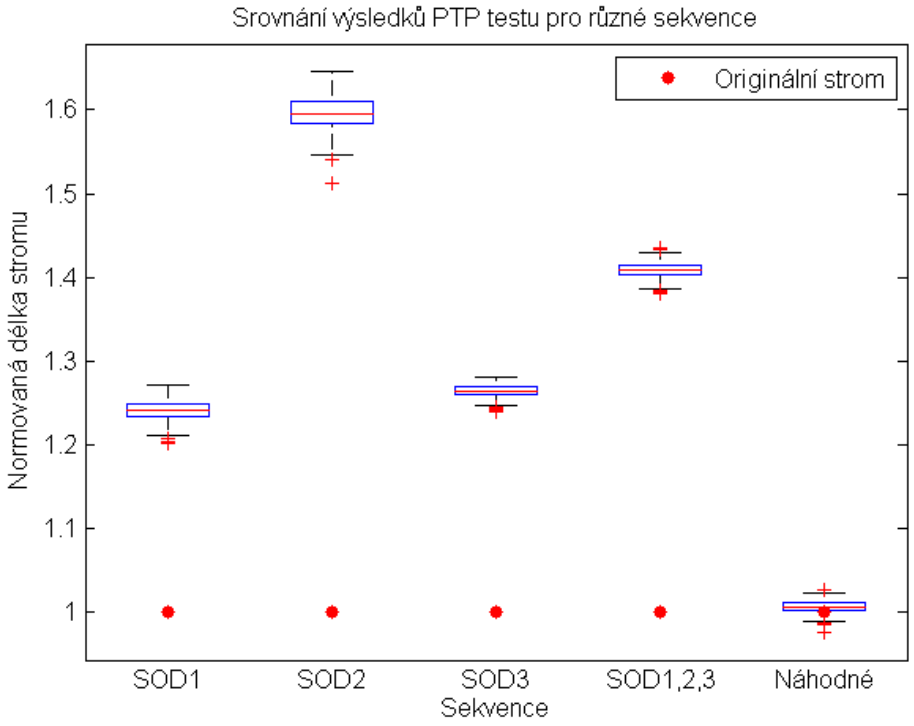
Při zkoumání vlivu matice PAM (obrázek 32) jsme obdrželi nejvyrovnanější charakter průběhu u křivky pro jednu isoformu i řád (primáti SOD 1). Poměrně vyrovnaný charakter má i křivka pro jednu živočišnou rodinu a tři isoformy (primáti SOD 1, 2 3). Ta začíná výrazněji kolísat až u poměrně vysokých hodnot matice PAM350 až PAM500. Poslední zkoumaná skupina sekvencí pro více isoforem i řádů (více rodin SOD 1, 2, 3) kolísá nejvýrazněji u velmi nízkých a vysokých hodnot matice PAM. Také z průběhu tohoto grafu je tedy patrné, že chybné zarovnání výrazně ovlivňuje hodnotu bootstrappingové podpory.

Shrnutí

Nejvyrovnanějšího charakteru dosahovaly výsledky u sekvencí jedné skupiny živočichů a jedné isoformy kódujícího proteinu. Tento data set obsahoval nejméně substitucí a delecí a tím pádem byl na změnu substituční matice nejvíce rezistentní. Podobnou rezistenci vykazovala i druhá skupina vstupních sekvencí pro jeden živočišný řád a tři isoformy u matice BLOSUM. U matice PAM se však ukázalo, že i tento data set je do značné míry ovlivňován výběrem skórovací matice a to zejména pro vyšší čísla matice PAM. Nejspíše je to způsobeno tím, že zarovnaný data set již obsahoval větší množství substitucí a delecí. Nejchoulostivější na změnu skórovací matice byly sekvence různých živočišných rodin i isoforem. Tento výběr obsahoval největší množství substitucí a delecí. Všechny výsledky této analýzy potvrzují předpoklad, že volba skórovací matice významným způsobem ovlivňuje výsledky bootstrappingového testu a to zejména pro sekvence s velkým množstvím delecí a substitucí.

5.3. Vliv výběru sekvencí na hodnotu PTP testu

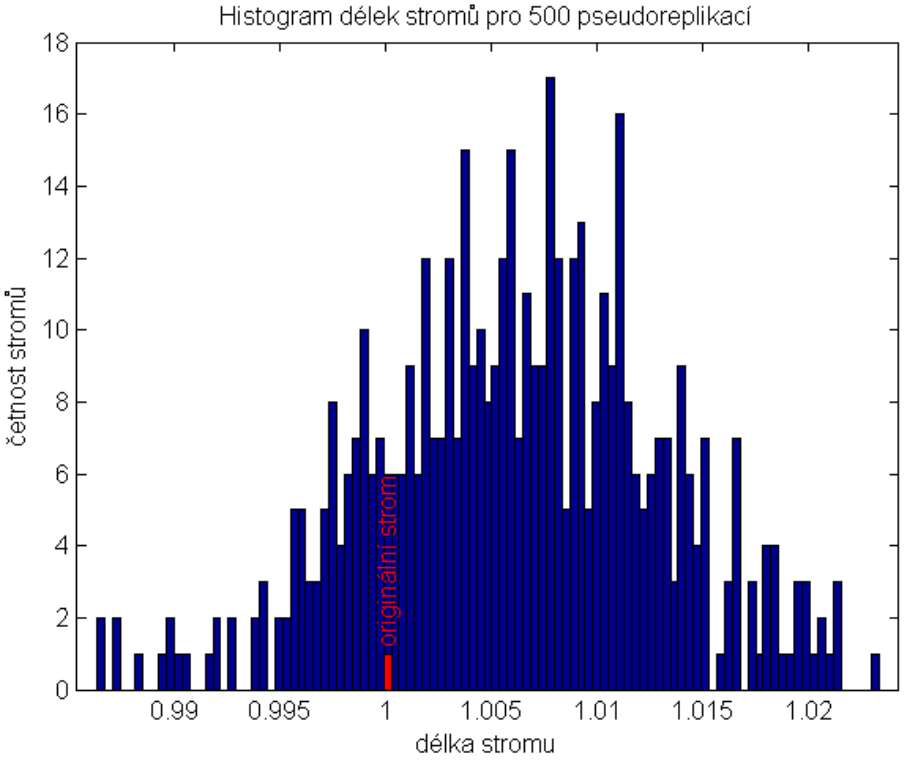
Pro tuto analýzu byly použity aminokyselinové sekvence vždy dvanácti živočichů různých taxonomických kategorií kódujících protein superoxid dismutázy v isoformách 1, 2 a 3. Pro srovnání jsou výsledky doplněny skupinou dvanácti náhodných aminokyselinových sekvencí. Jako distanční model byla použita p - distance. Skórovací matice byla BLOSUM 50 a počet replikací činil 500.



Obrázek 33: Srovnání výsledků PTP testu pro aminokyselinové sekvence superoxid dismutázy v různých isoformách.

Výsledný graf 33 ukazuje, že se výsledky PTP testu liší při použití sekvencí různých isoforem téhož proteinu. Účinnost odhalení fylogenetické informace v setu sekvencí jsme demonstrovali ve srovnání s náhodně generovanými sekvencemi o stejné délce s rovnoměrným rozložením nukleotidů. U náhodného setu sekvencí vyšly délky pseudostromů vždy blízké hodnotě originálního stromu, proto je jejich normalizovaná délka vůči originálu blízká jedné. Normalizované délky pseudostromů ze setů obsahujících fylogenetickou informaci vycházeli vždy výrazně vyšší než jedna. Zdá se, že nejvíce fylogenetické informace obsahují sekvence s isoformou SOD 2, což se dá vysvětlit mnohem větší konzervovaností proteinů této isoformy. Isoformy SOD 1 a SOD 3 pak obsahují podobné množství fylogenetické informace. Shoda není náhodná, tyto dvě isoformy jsou si totiž na molekulární úrovni mnohem více vzájemně podobné.V organismu se shodují i funkcí kterou plní tj. vazba iontů kovů Cu a Zn, zatímco druhá isoforma váže ionty Mn. Skupina obsahující sekvence všech tří isoforem se nachází mezi SOD 2 a SOD 1, 3. SOD 2 tedy prodloužilo délku pseudostromů v PTP testu. Ná-

hodné sekvence obsahovaly stejný počet znaků jako ostatní srovnávané skupiny. Přesto je z výsledku patrné, že neobsahují dostatečnou fylogenetickou informaci pro věrohodnou fylogenetickou analýzu. Náhodné sekvence nejsou totiž s jistotou homologní a PTP test tuto skutečnost názorně demonstruje. Pro detailnější představu o výsledku PTP testu pro náhodné sekvence je výsledný histogram vykreslen na obrázku 34 [48].



Obrázek 34: Výsledný histogram PTP testu pro náhodné sekvence.

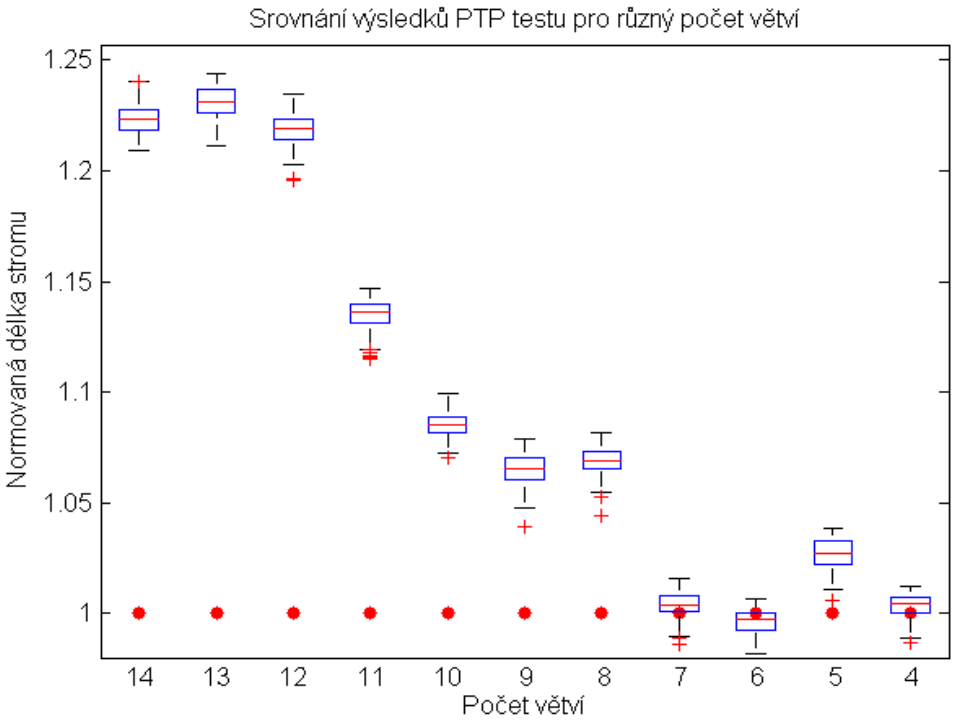
Shrnutí

Naše výsledky dokazují, že použití různých isoforem téhož proteinu mění výsledek PTP testu. Navíc se ukázalo, že obdržené výsledky odpovídají představám o množství fylogenetické informace daných isoforem. U velmi podobných isoforem jsme totiž obdrželi srovnatelný výsledek. Vzhledem k umístění srovnávaných skupin SOD 1, 3 a SOD 2 jsem mohli rovněž předpokládat délku pseudostromů pro všechny tři SOD skupiny, které jsou delší než pro samostatné skupiny SOD 1 i SOD 3. Dokázali jsme tedy, že PTP test jakožto fylogenetická analýza funguje podle našich předpokladů a že je velmi citlivý na výběr vhodných sekvencí pro fylogenetickou analýzu. To činí PTP test velmi důležitým a užitečným fylogenetickým nástrojem.

5.4. Vliv počtu sekvencí na hodnotu PTP testu

Pro tuto analýzu byly použity aminokyselinové sekvence čtrnácti živočichů různých řádů kódujících protein superoxid dismutázy v isoformě 3. Postupně jsme vždy jeden taxon z analýzy vyjmuli a výsledky PTP testů jsme vykreslili do box plotů, viz obrázek 35. Jako distanční

model byla použita p - distance. Skórovací matice byla BLOSUM 50 a počet replikací činil 100.



Obrázek 35: Srovnání výsledků PTP testu pro různý počet větví.

Předpokládali jsme, že výsledný graf bude mít sestupnou tendenci, což se nám spíše potvrdilo. Vždy záleží na charakteru a typu sekvence, což jsme si ukázali i v předchozí kapitole Sestupný charakter však rozhodně není pravidelný. Navíc mělo například odstranění šesté větve dokonce mírně pozitivní vliv na výsledek PTP testu.

Shrnutí

Množství použitých sekvencí téhož data setu má na PTP test výrazný vliv. Předpoklad o sestupném trendu výsledného grafu se nám sice potvrdil, ale v některých případech může odebrání nevhodné větvě naopak zlepšit celkový výsledek PTP testu. To může být způsobeno například použitím jedné sekvence u které probíhala evoluce jinými mechanismy než u zbylých. Je vždy problém vytvořit datový set tak, aby rychlost evoluce i mechanismy jež ji způsobují byly ve všech sekvencích shodné.

5.5. Volba správného množství pseudoreplikací a zkrácení u jackknifingového testu

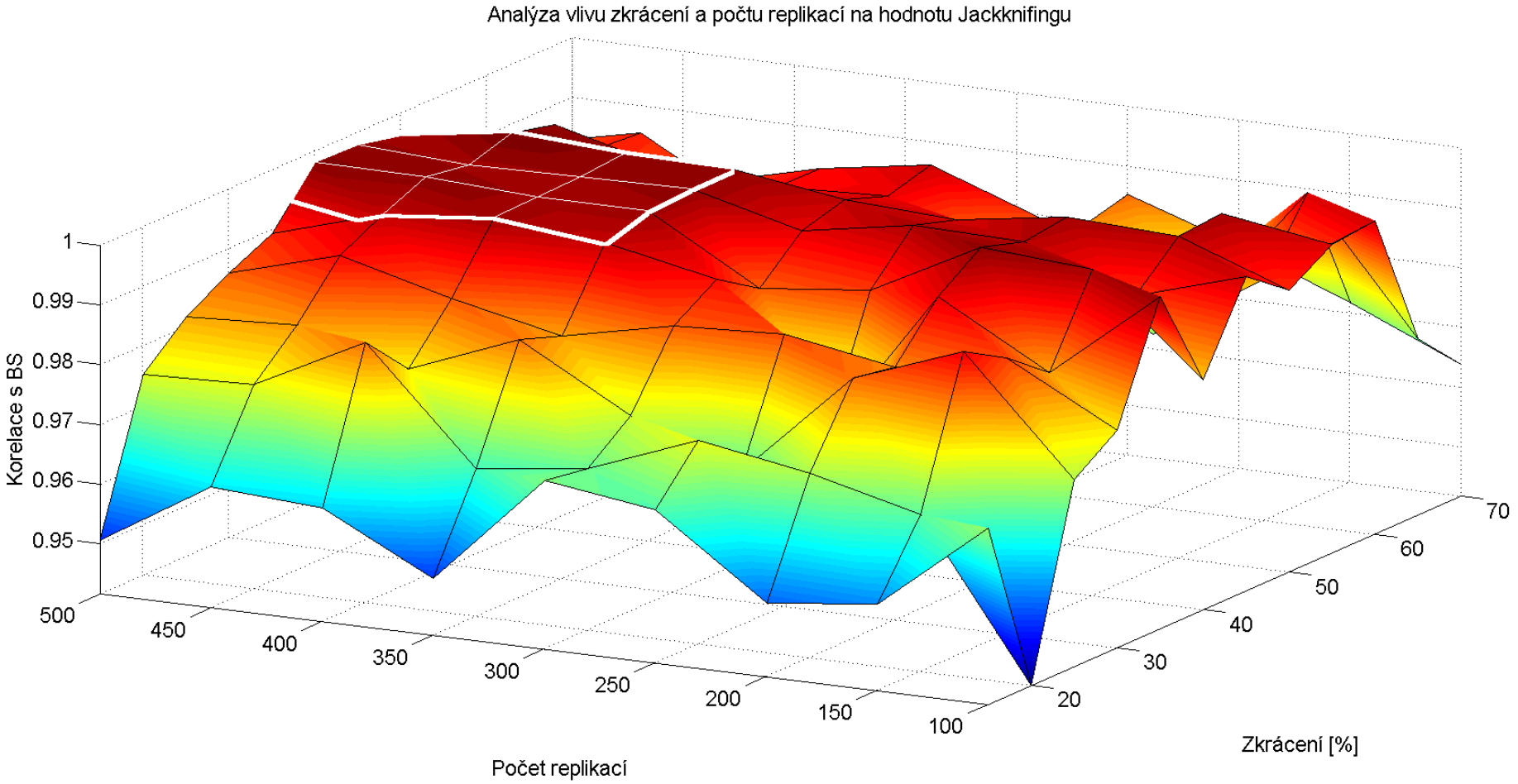
Jackknifingový a bootstrappingový test vychází z podobných předpokladů a hodnota podpory by v ideálním případě měla být pro oba shodná, viz kapitola 3.2. Přesto je v praxi používanější bootstrapping, protože vede k přesnějšímu výsledku s menším počtem pseudoreplikací, tzn.

s nižšími výpočetními nároky. U této analýzy jsme se snažili tento předpoklad ověřit a vyhodnotit, jaký počet pseudoreplikací a jaké procentuální zkrácení zarovnaného data setu sekvencí je pro jackknifingový test ideální, pokud chceme dosáhnou co nejshodnějšího výsledku s bootstrappingovým testem o sta pseudoreplikacích. Abychom mohli míru podobnosti obou testů snadno vyhodnotit, srovnávali jsme dosažené výsledky jackknifingové podpory s bootstrapovou pomocí pearsonova korelačního koeficientu. Čím více se bude korelační koeficient blížit jedné, tím větší shoda nastala. Výsledkem je třídimenzionální graf na obrázku 36. Pro analýzu byly použity aminokyselinové sekvence kódující čtvrtou podjednotku NADH dehydrogenázy patnácti živočichů. Sekvence jsou homologní a pro fylogenetickou analýzu živočichů velmi vhodné [47].

Z grafu (obrázek 36) je patrné, že hodnoty od sta do tří set pseudoreplikací jsou velmi nevyrovnané a tudíž nevhodné pro použití. Tři sta pseudoreplikací by již bylo možné použít, ovšem jen pro zkrácení od 35 % do 50 %. Nižší i vyšší procentuální hodnoty zkrácení jsou již příliš nepřesné a vzdalují se od bootstrappingové hodnoty.

Shrnutí

Abychom dosáhli u jackknifingového testu obdobných výsledků jako u bootstrappingového o stu pseudoreplikacích, musíme zkrátit zarovnaný data set sekvencí o 35 % až 50 %. Minimální počet replikací by měl činit alespoň 300. Jedině s takovými parametry dostáváme storvnatelné a stabilní výsledky. Dokázali jsme tedy, že pro dosažení srovnatelných výsledků je jackknifing výpočetně náročnější než bootstrapping. Oblast ideálních parametrů jackknifingového testu je pro přehlednost bíle zvýrazněna na výsledném grafu (obrázek 36).



Obrázek 36: Analýza vlivu zkrácení a počtu replikací na hodnotu jackknifingu.
Bíle je zvýrazněna je oblast ideálního zkrácení a počtu replikací pro dosažení dostatečně přesného výsledku vzhledem k bootstrappingu o 100 pseudorepli kacích.

ZÁVĚR

Výsledkem fylogenetické analýzy je zpravidla dendrogram, znázorňující schéma kladogeneze. Pro správné pochopení statistických testů, které popisují robustnost fylogenetických stromů, je nejprve potřeba správně pochopit metodiku konstrukce fylogenetických stromů od výběru molekulárních dat, přes zarovnání sekvencí a evoluční modely až po jejich samotné konstrukční metody. V každém z těchto konstrukčních kroků totiž mohou nastat problémy, které znemožňují správný popis fylogeneze. Tyto problémy jsme však do jisté míry schopni detekovat a případně odstranit. Za tímto účelem je v dnešní době nejčastěji využívána konstrukce stromů různými metodami při současném testování spolehlivosti topologií, pro které nám slouží statistické testy.

Nejčastěji využívanými statistickými testy fylogenetické analýzy jsou resamplingové testy. Ty mají hlavní výhodu v tom, že jsou neparametrické a tedy nemusíme znát rozložení vstupních dat, které do nich vstupují. Mezi tyto testy patří bootstrapping, jackknifing, OTU jackknifing a PTP test. V textu práce je uveden jejich princip a zejména interpretace jejich výsledků. Tu chápeme jako statistickou podporu námi vytvořeného větvení pro námi zvolená data. Procentuální hodnotu podpory v žádném případě nesmíme chápat jako pravděpodobnost toho, že v minulosti k tomuto větvení došlo. Fylogenetickým stromem se snažíme popsat historii, kterou neznáme a vždy ji pouze odhadujeme. Statistická podpora nám tedy v podstatě udává to, zda jsme námi vybranými metodami vytvořili strom, který naše vstupní data vhodně vysvětluje.

Praktická část práce se věnuje realizaci a hodnocení uvedených resamplingových testů v programovém prostředí matlab. Za účelem rychlého, snadného a přehledného načtení, hodnocení a zpracování výsledků resamplingových testů jsme vytvořili uživatelské programové rozhraní. Vytvořený program umožňuje načíst sekvence ve FASTA kódu a na základě těchto sekvencí zkonstruovat fylogenetický strom metodou neighbor-joining, s možností změny distančního modelu a substituční matice. Na tento strom pak můžeme aplikovat resamplingové testy, jejichž parametry můžeme také měnit a díky tomu sledovat jejich vliv na výsledek analýzy. Díky programu jsme poté provedli analýzu vlivu některých parametrů na výsledky resamplingových testů na různých typech sekvencí.

Zjistili jsme, že vliv změny distančního modelu na celkovou hodnotu bootstrapové podpory, při použití různých aminokyselinových i nukleotidových sekvencí, je minimální. Výsledek se dal předpokládat. Volba distančního modelu ovlivňuje spíše samotnou délku větví originálního stromu, samotná topologie není většinou ovlivněna. Zarovnané sekvence se totiž nemění vůbec. Při vytváření pseudoreplikací se pak replikují vždy ty stejné sloupce zarovnaných sekvencí a přestože poté dochází k výpočtu distanční matice na základě jiného modelu, topologie pseudostromů zůstávají téměř nezměněny.

Rozdíl nastal při hodnocení vlivu skórovací matice na výsledek bootstrappingu. Všechny výsledky této analýzy potvrzují, že volba skórovací matice významným způsobem ovlivňuje výsledky bootstrappingového testu a to zejména pro sekvence s velkým množstvím delecí a substitucí. Při změně skórovací matice se totiž mění způsob zarovnání sekvencí. V zarovnaném data setu tak pro různé skórovací matice vznikají odlišné sloupce znaků, které se pak v testu replikují. Díky tomu se pak budou lišit i pseudostromy a celkové výsledky analýzy pro různé skórovací matice. Největší vliv pak lze očekávat u sekvencí s nízkou vzájemnou identitou. U těch se totiž při změně skórovací matice nejvýrazněji změní data set zarovnaných sekvencí.

Naše výsledky také ukazují, že použití různých isoforem téhož proteinu mění výsledek PTP testu. Při použití různých isoforem však jsme schopni odhalit určitou analogii vzhledem ke znalostem o podobnosti jednotlivých isoforem. PTP test také jasně odhalil nevhodnost náhodných sekvencí, které rozhodně nejsou homologní, pro fylogenetickou analýzu. Dokázali jsme si, že je PTP test velmi citlivý na volbu analyzovaných sekvencí a že je tedy velmi užitečnou pomůckou fylogenetické analýzy.

Rovněž množství použitých sekvencí téhož data setu má na PTP test výrazný vliv. Výsledky však ukazují, že nelze předpokládat, že s nižším množstvím sekvencí obdržíme nutně horší výsledek PTP testu. V některých případech může odebrání nevhodné větvě naopak zlepšit celkový výsledek testu. Vždy bude záležet na vztahu celého data setu vůči konkrétní sekvenci, kterou odebíráme či přidáváme.

V poslední analýze jsme se zabývali stanovením vhodného počtu replikací a vhodné procentuální hodnoty zkrácení zarovnaných sekvencí u jackknifingového testu. Výsledky ukazují, že ideálním zkrácením u jackknifingu je 35 % až 50 %. Minimální počet replikací by měl činit alespoň 300. Pro tyto hodnoty dostáváme dostatečně přesné, stabilní a srovnatelné výsledky, vzhledem ke korelaci tohoto testu s bootstrappingovým o stu pseudoreplikacích. Dokázali jsme tedy, že pro dosažení shodných výsledků obou testů musíme u jackknifingu provést více replikací a že je tedy jackknifing výpočetně náročnější než bootstrapping.

SEZNAM LITERATURY:

[1] HOLMES, Susan. *Bootstrapping Phylogenetic Trees: Theory and Methods*. Statistical Science. 2003, roč. 18, č. 2, s. 241-255.

[2] NEI, Masatoshi a Sudhir KUMAR. *Molecular Evolution and Phylogenetics*. New York: Oxford University Press, 2000. ISBN 0-19513585-7.

[3] FLEGR, Jaroslav. *Evoluční biologie*. 2., opr. a rozš. vyd. Praha: Academia, 2009. ISBN 978-80-200-1767-3.

[4] CHANDRA, Girish. Anagenesis & Cladogenesis. *An online guidance in Zoology* [online]. 2011 [cit. 2013-11-09]. Dostupné z: http://www.iaszoology.com/anagenesis-cladogenesis/

[5] My Name is LUCA—The Last Universal Common Ancestor. *Actionbioscience* [online]. 2002 [cit. 2013-11-09]. Dostupné z: http://www.actionbioscience.org/newfrontiers /poolepaper.html

[6] HAMPL, Vladimír. MOLEKULÁRNÍ TAXONOMIE - PŘEDNÁŠKA. *Protistologie* [online]. 2.3. 2012, s. 5 [cit. 2013-11-09]. Dostupné z: http://web.natur.cuni.cz/~vlada/ moltax/

[7] ŘEHULKA, Pavel. Základy bioinformatického zpracování dat v proteomice. *Bioinformatika* [online]. 2009, s. 118 [cit. 2013-11-09]. Dostupné z: http://www.pmfhk.cz /WWW/UMP/aplikovana_proteomika/bioinformatika_pr_cv.pdf

[8] HIGGS, Paul G a Teresa K ATTWOOD. *Bioinformatics and molecular evolution*. Malden, MA: Blackwell Pub., c2005, xiii, 365 p. ISBN 14-051-0683-2.

[9] CVRČKOVÁ, Fatima. *Úvod do praktické bioinformatiky*. Vyd. 1. Praha: Academia, 2006, 148 s. ISBN 80-200-1360-1.

[10] International Union of Pure and Applied Chemistry. THE INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY. *Iupac* [online]. 2013 [cit. 2013-11-09]. Dostupné z: http://www.iupac.org/

[11] SVRŠEK, Jiří. Molekulární biologie: Základní rysy genetického kódu. *Natura plus* [online]. 1997 [cit. 2013-11-09]. Dostupné z: http://natura.baf.cz/natura/1996/5/9605-6.html

[12] ŠKUTKOVÁ, Helena. *Analýza biologických signálů*. Brno, 2013. Přednáška. Vysoké učení technické Brno, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství.

[13] NEČÁSEK, Jan. *Genetika*. 2. vyd. Praha: Scientia, 1997, 112 s. ISBN 80-718-3085-2.

[14] ŠÍPEK, Antonín. Genetika-biologie: Váš zdroj informací o genetice a biologii. [online]. [cit. 2013-11-09]. Dostupné z: http://www.genetika-biologie.cz/

[15] PHILLIPS, Aloysius, Daniel JANIES, Ward WHEELER. *Multiple sequence alignment in phylogenetic analysis.* Molecular Phylogenetics and Evolution [online]. 2000, č. 16, 17–330 [cit. 2013-11-17]. Dostupné z: http://www.ncbi.nlm.nih.gov/pubmed/ 10991785

[16] OGDEN, T. HEATH a MICHAEL S. ROSENBERG. *Multiple Sequence Alignment Accuracy and Phylogenetic Inference. Syst. Biol.* [online]. 2006, č. 55, 314–328 [cit. 2012-11-17]. Dostupné z: http://www.ncbi.nlm.nih.gov/pubmed/16611602

[17] ROST, B. *Twilight zone of protein sequence alignments*. Protein Eng [online]. 1999, roč. 12, č. 2, s. 85-94. Dostupné z: http://www.ncbi.nlm.nih.gov/pubmed/10195279

[18] KLEYWEGT, G. J. A world about homology. [online]. 2005 [cit. 2013-11-20]. Dostupné z: http://xray.bmc.uu.se/~kurs/BSBX2/documents/gk_1_homology.pdf

[19] NOVOTNÝ, Marián. MOLEKULÁRNÍ TAXONOMIE - PŘEDNÁŠKA. *Alignment* [online]. 9.3. 2012, s. 109 [cit. 2013-11-20]. Dostupné z: http://www.protistologie.cz/ files/MolTax/MolTax2012-3.pdf

[20] FIELD, K., G. OLSEN, D. LANE, S. GIOVANNONI, M. GHISELIN, E. RAFF, N. PACE a R. RAFF. Molecular phylogeny of the animal kingdom. *Science* [online]. 1988-02-12, vol. 239, issue 4841, s. 748-753 [cit. 2013-12-06]. DOI: 10.1126/science. 3277277. Dostupné z: http://www.sciencemag.org/cgi/doi/10.1126/ science.3277277

[21] XIA, Xuhua, Zheng XIE a Karl M. KJER. 18S Ribosomal RNA and Tetrapod Phylogeny. *Systematic Biology* [online]. 2003-6-1, vol. 52, issue 3, s. 283-295 [cit. 2013-12-06]. DOI: 10.1080/10635150390196948. Dostupné z: http://sysbio.oxfordjournals.org/cgi /doi/10.1080/10635150390196948

[22] NODA, R., C. G. KIM, O. TAKENAKA, R. E. FERRELL, T. TANOUE, I. HAYASAKA, S. UEDA, T. ISHIDA a N. SAITOU. Mitochondrial 16S rRNA Sequence Diversity of Hominoids. *Journal of Heredity* [online]. 2001-11-01, vol. 92, issue 6, s. 490-496 [cit. 2013-12-06]. DOI: 10.1093/jhered/92.6.490. Dostupné z: http:// jhered.oxfordjournals.org/cgi/doi/10.1093/jhered/92.6.490

[23] Jukes TH a Cantor CR Evolution of protein molecules. In Munro HN, editor, Mammalian Protein Metabolism, *Academic Press*, New York. 1969, pp. 21-132 [cit. 2013-12-06].

[24] KIMURA, Motoo. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* [online]. 1980, vol. 16, issue 2, s. 111-120 [cit. 2013-12-06]. DOI: 10.1007/BF01731581. Dostupné z: http://link.springer.com/10.1007/BF01731581

[25] TAMURA, Koichiro. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+ C-content biases. Molecular *Biology and Evolution* [online]. 1992, 9.4: 678-687 [cit. 2013-12-06].

[26] TAMURA, Koichiro; NEI, Masatoshi. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular biology and evolution*. 1993, 10.3: 512-526 [cit. 2013-12-06].

[27] NEI, Masatoshi; GOJOBORI, Takashi. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution*, 1986, 3.5: 418-426 [cit. 2013-12-06].

[28] Li, Wen-Hsiung, Chung-I. Wu, a Chi-Cheng Luo. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution 2*, no. 2 1985, 150-174 [cit. 2013-12-06].

[29] BERGSTEN, Johannes. A review of long-branch attraction. *Cladistics* [online]. 2005, vol. 21, issue 2, s. 163-193 [cit. 2013-12-13]. DOI: 10.1111/j.1096-0031.2005.00059.x. Dostupné z: http://doi.wiley.com/10.1111/j.1096-0031.2005.00059.x

[30] FITCH, Walter M. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 1971, 20.4: 406-416 [cit. 2013-12-06].

[31] RZHETSKY, Andrey; NEI, Masatoshi. A simple method for estimating and testing minimum-evolution trees. *Mol. Biol. Evol*, 1992, 9.5: 945-967 [cit. 2013-12-06].

[32] STUDIER, James A.; KEPPLER, Karl J. A note on the neighbor-joining algorithm of Saitou and Nei. Molecular biology and evolution, 1988, 5.6: 729-731.

[33] SULLIVAN, Jack a David L. SWOFFORD. Are Guinea Pigs Rodents? The Importance of Adequate Models in Molecular Phylogenetics. *Journal of Mammalian Evolution*. 1997, roč. 4, č. 2, s. 77-86. DOI: 10.1023/A:1027314112438. Dostupné z: http://link.springer.com/10.1023/A:1027314112438

[34] LYONS-WEILER, J. Escaping from the Felsenstein Zone by Detecting Long Branches in Phylogenetic Data. *Molecular Phylogenetics and Evolution* [online]. 1997, vol. 8, issue 3, s. 375-384 [cit. 2013-12-14]. DOI: 10.1006/mpev.1997.0450. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S1055790397904504

[35] HENDY, Michael D. a David PENNY. Spectral analysis of phylogenetic data. *Journal of Classification* [online]. 1993, vol. 10, issue 1, s. 5-24 [cit. 2013-12-14]. DOI: 10.1007/BF02638451. Dostupné z: http://link.springer.com/10.1007/BF02638451

[36] SOLTIS, Douglas E. a Pamela S. SOLTIS. Applying the Bootstrap in Phylogeny Reconstruction. *Statistical Science*. 2003, roč. 18, č. 2, s. 256-267. DOI: 10.1214/ss/1063994980. Dostupné z: http://projecteuclid.org/Dienst/ getRecord?id=euclid.ss/1063994980/

[37] EFRON, Bradley a Robert TIBSHIRANI. *A introduction to the bootstrap*. Boca Raton: Chapman, c1994. Monographs on statistics and applied probability. ISBN 04-120-4231-2.

[38] EFRON, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* [online]. 1979, vol. 7, issue 1, s. 1-26 [cit. 2013-12-14]. DOI: 10.1214/aos/1176344552. Dostupné z: http://projecteuclid.org/euclid.aos/1176344552

[39] HEDGES SB. The number of replications needed for accurate estimation of the bootstrap P value in phylogenetic studies. *Molecular Biology and Evolution*. 1992;9(2):366-369.

[40] HILLIS DM, BULL JJ. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*. 1993;42(2):182-192.

[41] QUENOUILLE MH. Approximate tests of correlation in time-series 3. *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 45, no. 03, pp. 483-484. Cambridge University Press, 1949.

[42] TUKEY JW. Bias and confidence in not quite large samples (abstract). *Ann Math Stats.* 29: 614, 1958.

[43] KUNSCH, Hans R. The Jackknife and the Bootstrap for General Stationary Observations. *The Annals of Statistics* [online]. 1989, vol. 17, issue 3, s. 1217-1241 [cit. 2013-12-14]. DOI: 10.1214/aos/1176347265. Dostupné z: http://projecteuclid.org/euclid.aos/1176347265

[44] FELSENSTEIN, Joseph. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 783-791. 1985.

[45] SAWYER, S. Resampling Data: Using a Statistical Jackknife. [online]. 2005, s. 5 [cit. 2014-01-02]. Dostupné z: http://www.math.wustl.edu/~sawyer/handouts/Jackknife.pdf.

[46] BRYANT, H. N. The Role of Permutation Tail Probability Tests in Phylogenetic Systematics. Systematic Biology [online]. 1992-06-01, vol. 41, issue 2, s. 258-263 [cit. 2014-01-02]. DOI: 10.1093/sysbio/41.2.258. Dostupné z: http://sysbio.oxfordjournals.org/cgi/doi/10.1093/sysbio/41.2.258

[47] RUSSO, C. A.; TAKEZAKI, Naoko; NEI, Masatoshi. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Molecular Biology and Evolution [online], 1996, 13.3, 525-536.

[48] ŠKUTKOVÁ, Helena, BABULA, Petr, PROVAZNÍK, Ivo. Bioinformatic study of genetic variability of superoxide dismutase isoforms. In: *2nd International Conference on Chemical Technology*. Mikulov: Česká společnost průmyslové chemie, 2014, s. 1-6. ISBN ISBN 978-80-86238-61-6.

SEZNAM PŘÍLOH

PŘÍLOHA 1: Základní fylogenetické pojmy

Pro snazší pochopení tohoto textu je potřeba seznámit se s některými základními pojmy fylogenetiky. Hesla jsou popsána zjednodušeně jen pro účely pochopení práce.

Taxon Též systematická jednotka, taxonomická jednotka či operační taxonomická jednotka - Skupina konkrétních organismů, které mají společné určité znaky. Např. čeleď, řád i jiné.

OTU Z anglického Operational Taxonomical Unit - zkratka operační taxonomické jednotky, viz Taxon.

RNA Z anglického ribonucleic acid - Makromolekula složená z řetězce nukleotidů, obsahujících cukr ribózu. Nejčastěji tvoří jednovlánovou strukturu. Rozlišujeme transverovou RNA (tRNA), mediátorovou RNA (mRNA) a ribozomální RNA (rRNA). Mezi nejdůležitější funkce RNA patří přenos genetické informace při transkripci a translaci.

DNA Z anglického deoxyribonucleic acid - Makromolekula složená z řetězce nukleotidů, obsahujících cukr deoxyribózu. Nejčastěji tvoří dvoušroubovici, v níž jsou jednotlivé řetězce uspořádány dle komplementarity bazí. Je nositelkou genetické informace, kdy díky procesu transkripce do mRNA a následné translace utváří primární strukturu proteinu dle genetického kódu.

Nukleotid Biologické molekuly, skládající se z cukru (ribóza nebo deoxyribóza), fosfátového zbytku a nukleové báze. Názvosloví nukleotidů odpovídá bázi, kterou daný nukleotid obsahuje. Báze adenin, cytosin, guanin a thymin (zkratky A, C, G, T) se vyskytují u DNA. U RNA je thymin nahrazen uracilem (zkratka U). Jednotlivé nukleotidy jsou k sobě vázány fosfodiesterovou vazbou a vytváří řetězec DNA či RNA. Nukleotidy se dále mohou spojovat s dalším řetězcem pomocí vodíkových můstků dle komplementarity bazí.

Komplementarita bazí Způsob, jimiž jsou nukleotidy navzájem pospojovány pomocí vodíkových můstků. Dle Watson-Crikovských pravidel se páruje adenin s thyminem (případně s uracilem u RNA) za použití dvou vodíkových můstků a guanin s cytosinem za použití tří vodíkových můstků. Existují však i další možnosti párování.

Gen Specifický úsek DNA, který je exprimován do struktury proteinu. Soubor všech genů tvoří genotyp, který společně s prostředím utváří fenotyp daného organismu.

Genová exprese	Způsob přepisu genetické informace do sekvence aminokyselin. Jednotlivé geny se v procesu transkripce přepíší do molekuly mRNA, následně dochází k sestřihu. Introny jsou části mRNA které jsou vystřiženy, exony jsou pak části které jsou dále exprimovány. mRNA slouží jako přenašeč informace od DNA k ribozómům, na kterých probíhá translace. Pořadí aminokyselin se zde stanovuje tak, že ke každému kodonu (tripletu, trojici nukleotidů) se připojí tRNA s odpovídajícím antikodonem nesoucí aminokyselinu.
Start kodon	Též iniciační kodon. Kodon, u něhož začíná proces translace. Většinou se jedná o kodon AUG.
Stop kodon	Kondon, u něhož dochází k zastavení translace a tedy i celé proteosyntézy. Jedná se o kodony UAA, UAG, UGA.
Protein	Protein, neboli bílkovina, je makromolekula, jejíž primární strukturu tvoří řetězec aminokyselin. Jednotlivé aminokyseliny jsou pospojovány peptidovou vazbou (-NH-CO-). Proteiny tvoří podstatu všech živých organismů.
Aminokyselina	V užším slova smyslu je chápeme jako 23 základních stavebních jednotek proteinů.
Transice	Jedná se o bodovou mutaci, při níž dochází k záměně purinové báze za jinou, taktéž purinovou.
Transverze	Jedná se o bodovou mutaci, při níž dochází k záměně purinové báze za pyrimidinovou.
Homoplazie	Jev, ke kterému dochází pokud dva druhy sdílí stejnou formu znaku ale ta se u každého z nich vyvinula nezávisle, například v důsledku působení stejných selekčních faktorů.
Homologní geny	Jsou to takové geny, které jsou odvozeny z jednoho společného genu (předka). Rozdělujeme je na ortologní a paralogní.
Paralogní geny	Paralogní geny, které vznikly po duplikaci druhu (původního genu). Mají pozměněnou funkci.
Ortologní geny	Jsou výsledkem speciace původního genu, divergují po vzniku druhu. Čili u všech druhů mají stejnou funkci.
Xenologní geny	Geny vzniklé horizontálním přenosem (např. předáním části bakteriální DNA na hostitele)
Pseudogeny	Zbytky genů, které během evoluce ztratily svůj význam.