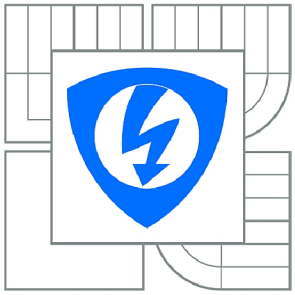
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGINEERING

METODY REKONSTRUKCE FYLOGENETICKÝCH SUPERSTROMŮ

METHODS FOR PHYLOGENETIC SUPERTREE RECONSTRUCTION

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

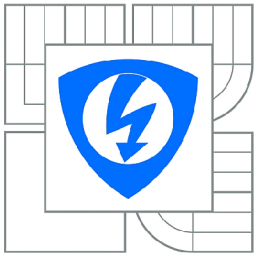
AUTOR PRÁCE — Bc. KAMIL KOSÍŘ

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. Kamil Kosíř	***ID:***	125041
Ročník:	2	***Akademický rok:***	2013/2014

NÁZEV TÉMATU:

Metody rekonstrukce fylogenetických superstromů

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši problematiky fylogenetických superstromů a jejich rekonstrukčních technik. 2) Proveďte srovnání algoritmů pro rekonstrukci fylogenetických superstromů ze zdrojových stromů na základě principu kompletace částečných vzdálenostních matic. 3) Zhodnoťte volbu zdrojových stromů a jejich datový popis. Vytvořte vhodný datový set biologických sekvencí z veřejných databází a vytvořte zdrojové stromy pro následnou srovnávací studii rekonstrukčních algoritmů superstromů. 4) Srovnání algoritmů a jejich výstupů demonstrujte na vlastních skriptech realizovaných v programovém prostředí Matlab s Bioinformatickým toolboxem. 5) Proveďte diskuzi získaných výsledků a zhodnoťte výhody a nevýhody jednotlivých metod.

DOPORUČENÁ LITERATURA:

[1] CREEVEY, C. J. , MCINERNEY, J. O. Trees from trees: construction of phylogenetic supertrees using clann. Methods In Molecular Biology Clifton Nj, vol. 537, pp. 139-161, 2009.
[2] BININDA-EMONDS, O. R. P. The evolution of supertrees. Trends in ecology & evolution (Personal edition), vol. 19, no. 6, pp. 315-22, Jun. 2004.

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:

Autor diplomové práce nesmí při vytváření diplomové práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

Abstrakt

V posledních 30 letech zaznamenala fylogenetika velký rozvoj. Počítače se staly silnější a lépe dostupné a spolu s více sofistikovanými algoritmy přichází snaha vědců zrekonstruovat z velkého množství fylogenetických dat kompletní strom života. Právě pro tyto účely vznikají fylogenetické superstromy, které umožňují kombinaci všech doposud získaných informací. Cílem této práce je nalezení a sestrojení metody konstrukce superstromu, která bude dávat přesné výsledky.

Klíčová slova

Fylogenetika, fylogenetický strom, fylogenetický superstrom, OTU

Abstract

The Phylogenetic reconstruction has seen great development in the last 30 years. Computers have become more powerful and more generally accessible, and computer algorithms more sophisticated. It comes the effort of scientists to reconstruct the entire tree of life from a large amount of phylogenetic data. Just for this purpose are formed phylogenetic supertrees that allow the combination of all information gathered so far. The aim of this work is to find a method to construct supertree that will give correct results.

Key words

Phylogenetics, phylogenetic tree, phylogenetic supertree, OTU

KOSÍŘ, K. *Metody rekonstrukce fylogenetických superstromů.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2014. 71 s. Vedoucí diplomové práce Ing. Helena Škutková.

Prohlášení

Prohlašuji, že svou diplomovou práci na téma Metody rekonstrukce fylogenetických superstromů jsem vypracoval samostatně pod vedením vedoucího semestrálního projektu 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 osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujícího autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

V Brně dne …… ……

Podpis autora

Poděkování

Mé poděkovaní patří v první řadě vedoucí diplomové práce Ing. Heleně Škutkové za trpělivost, věcné připomínky k práci a praktické rady. Dále bych rád poděkoval mým nejbližším, kteří mě ve studiu podporovali.

Obsah

Seznam obrázků

Úvod

Fylogenetika je věda studující evoluční příbuznost mezi různými druhy organismů. Podle moderní evoluční teorie pocházejí všechny organismy na Zemi ze společného předka. Vývoj druhů neboli evoluce organismů je zobrazena pomocí fylogenetického stromu. V dřívějších dobách se stromy konstruovaly na základě subjektivních dat. Používány byly především morfologické znaky. V dnešní době se díky objevení struktury DNA používají ve fylogenetice molekulární data, která jsou zcela objektivní. Obor zabývající se molekulárními daty se nazývá molekulární fylogenetika.

Pro konstrukci fylogenetických superstromů je zapotřebí sestrojit nejdříve stromy zdrojové. Pro tyto stromy existuje celá řada rekonstrukčních metod a je potřeba zvážit několik faktorů, před použitím té dané metody, jako například předpokládanou evoluční vzdálenost, typ dat určených pro rekonstrukci a výpočetní náročnost.

Fylogenetické superstromy řeší problém fylogenetické analýzy z nejistých nebo nekompletních vstupních dat. Jejich výsledkem je sestrojení celkově lepších superstromů, obsahujících kompletní data ze stromů vstupních, kombinací více fylogenetických stromů. Pro různé metody existují různé pravidla, jejichž výsledkem by neměla být pouze kombinace informací obsažených ve zdrojových stromech, ale taky odvození vztahů, které se v jednotlivých zdrojových stromech vyskytují.

Metody pro konstrukci fylogenetických superstromů se vzájemně liší především ve způsobu kombinace informací z různých zdrojů. V této práci jsou popsány a představeny tři metody pro rekonstrukci fylogenetických superstromů, založené na principu kompletace částečných vzdálenostních matic. První z nich je maticová reprezentace úspornosti. Tato metoda využívá pro konstrukci matice reprezentující vztahy mezi zdrojovými stromy kódovací schéma. Ke kódování jsou použity 0 a 1, které se přidělují taxonům na základě jejich polohy ve fylogenetickém stromu. Druhá popsaná metoda se nazývá MSSA. Jedná se o metodu, která hledá nejlepší superstrom bez průměrování informací ze zdrojových stromů. Využívá heuristické hledání superstromu s bodovací funkcí, které jsou minimalizovány a vrací strom, který se nejvíce podobá některému ze souboru zdrojových stromů. Třetí metodou a zároveň metodou, kterou se zabývá tato diplomová práce, je metoda průměrného konsensu, která počítá vzdálenost každého taxonu s každým skrz všechny zdrojové stromy. Průměrná vzdálenost každého taxonu je použita pro finální vzdálenostní matici, ze které je konstruován superstrom. Tato metoda je doplněna o dvě metody pro výpočet prázdných míst ve výsledné distanční matici, ultrametrickou a aditivní. Obě tyto metody využívají matematických metod k dopočítání chybějících částí ještě před fylogenetickou rekonstrukcí.

1 MOLEKULÁRNÍ FYLOGENETIKA

Hlavním cílem molekulární fylogenetické studie je vytvořit evoluční strom, který bude graficky znázorňovat vztahy mezi druhy a geny během jejich evoluce. Jedná se o velmi složitý proces, který dále komplikuje i skutečnost, že neexistuje jediný správný způsob řešení fylogenetických problémů. Soubory dat pro fylogenezi se mohou skládat ze stovek různých druhů, kde každý může mít rozdílnou míru mutace, která má vliv na evoluční změny. Právě z tohoto důvodu je k dispozici mnoho odlišných evolučních modelů a stochastických metod, zvolení optimální metody záleží na povaze studie a použitých datech [6],[20].

V roce 1859 vydal Charles Darwin publikaci s názvem On the Origin of Species by Means of Natural Selection, ort he Preservation of flavoured Races in Stuggle for Life (O vzniku druhů přírodním výběrem, neboli uchováním prospěšných plemen v boji o život), ve které byly přesvědčivé důkazy, které poukazovaly na to, že všechny živé organismy pocházejí ze společného předka. V této publikaci uvedl i schéma neboli strom života, který vysvětluje základní princip biologické evoluce. Od této chvíle byla evoluce živočišných druhů chápána jako postupný vývoj života od prvního výskytu na Zemi k mnoha různým formám života, které se dále vyvíjejí [6],[16].

1.1 Fylogenetika

Fylogenetika je věda studující evoluční příbuznost mezi různými druhy organismů, populacemi apod. Podrobněji se fylogenetika zabývá studiem fylogeneze, tedy analýzou vývojových linií (taxonů).

Podle moderní evoluční teorie pocházejí všechny organismy na Zemi ze společného předka, z čehož vyplývá, že jsou všechny druhy příbuzné. Tento vztah je nazýván jako fylogeneze a je graficky reprezentována fylogenetickým stromem, který znázorňuje historii evolučních vztahů.

V dřívějších dobách se stromy konstruovaly na základě subjektivních dat. Používány byly především morfologické znaky (počet končetin, barva srsti, tvar uší,…). V dnešní době se pracuje s objektivními daty, čímž jsou například molekulární znaky, kam patří i nukleotidové a proteinové sekvence. Této problematice se věnuje obor molekulární fylogenetiky [6],[11].

1.2 Molekulární fylogenetika

To že druhy pocházejí ze společného předka, je naznačeno podobností biologických funkcí a molekulárních organismů v živých organismech. K odvození těchto vztahů používá molekulární fylogenetika strukturu a funkci molekul. Tento obor studie vznikal na začátku 20. století v letech 1960 s příchodem proteinového sekvenování, elektroforézy, PCR a dalších technik molekulární biologie. Za posledních 30 let, kdy se počítače staly silnější a lépe dostupné spolu s více sofistikovanými algoritmy, byli vědci schopni řešit složité a pravděpodobnostní problémy, které definují evoluci na molekulární úrovni efektivněji. V posledním desetiletí bylo toto odvětví znovu reorganizováno a sekvenovaní celého genomu pro komplexní organismy se stalo rychlejší a méně nákladné. S velkým množstvím volně dostupných genomických dat molekulární fylogenetika neustále roste a nachází nové aplikace.

Základním cílem molekulární fylogenetické studie je obnovit pořadí evolučních událostí a reprezentovat je pomocí evolučních stromů, které graficky znázorňují vztahy mezi druhy nebo geny za určitý čas. Jedná se o velmi složitý proces, jelikož fylogenetické datové sady se můžou skládat ze stovek různých druhů, z nichž každý může mít různé mutace, které mají vliv na evoluční změny. V důsledku toho existuje mnoho různých evolučních modelů a stochastických metod. Výběr optimální metody pro fylogenetickou analýzu závisí na povaze dat, které jsou k dispozici [6].

1.2.1 Molekulární znaky

Informaci uchovanou v sekvenčních datech je možné získat dvěma způsoby, sekvenováním nebo použitím nepřímých metod.

Nejvýhodnějším zdrojem molekulárně biologických znaků z hlediska možnosti dalšího zpracování je sekvenování DNA. Sekvenování odhalí největší procento genetických znaků, ve kterých se studované druhy liší a současně s velkou pravděpodobností identifikuje podmnožinu selekčně neutrální znaků.

Méně kvalitní informaci poskytují nepřímé metody, které ale mohou být technicky výrazně jednodušší, rychlejší a levnější [12].

1.2.2 Výhody molekulárních znaků

Molekulární znaky mají oproti klasickým znakům mnoho důležitých výhod. Jednou z nich je jejich libovolné množství, bereme-li v potaz velikost genomu většiny organismů. Jsou to znaky zpravidla kvalitativní a v porovnání s klasickými (morfologickými) znaky, které jsou charakteru kvantitativního, jsou lépe popsatelné. V praxi to znamená, že lze přesně

popsat jednotlivé molekulární znaky dvou vzájemně lišících se druhů tak, že budou lehce pochopitelné.

Pomocí molekulárních znaků je možné porovnávat a třídit organismy s minimálním množstvím společných znaků a sdružovat tyto odlišné organismy do vyšších taxonů. Lze získat jednoznačnou odpověď na otázku, které z porovnávaných dvojic druhů jsou si vzájemně podobnější (muchomůrka-fialka, muchomůrka-vrabec, nebo fialka-vrabec).

„Většina molekulárně biologických znaků je selekčně neutrálních, proto počet společně sdílených znaků mezi dvěma druhy odráží míru příbuznosti těchto druhů, nikoli míru podobnosti selekčních tlaků, které na ně v minulosti působily“[12].

Molekulární znaky jsou vhodné především pro studium kladogeneze (kladogeneze = odštěpování vývojových linií), méně vhodné jsou pro poskytnutí informaci týkajících se anageneze organismů (anageneze = změna vlastností taxonu nebo vývojové linie). Klasické znaky jsou v některých oblastech fylogenetiky již z principu jediným použitelným nástrojem. Nelze tedy zcela nahradit studium klasických znaků studiem znaků molekulárních [12].

1.2.3 Zpracování molekulárních dat

Relativně levnou a jednoduchou technikou pro získání molekulárně biologických znaků se v současnosti stalo sekvenování DNA. Tuto techniku je možné použít na prakticky libovolnou skupinu organismů. Sekvenování DNA spočívá v určení sekvence nukleotidů v jednom z řetězců DNA. K určení přesné sekvence nukleotidů v úseku DNA slouží dvě metody: Sangerova a Maxam & Gilbertova metoda [12].

Pro vytvoření fylogenetického stromu je výhodnější použití dat proteomických. Protože pro účely taxonomie a fylogenetiky je sekvenování proteomu technicky nesrovnatelně náročnější a v současnosti se nevyužívá, tato data lze získat jednoduchým převodem z dat genomických [12].

Pro počítačové zpracování je nutné získané molekulární znaky srozumitelně popsat. V praxi se používá jednoduchý zápis ve formě IUPAC (International Union of Pure and Apllied Chemistry) kódu, který je znázorněn v tabulce 1.

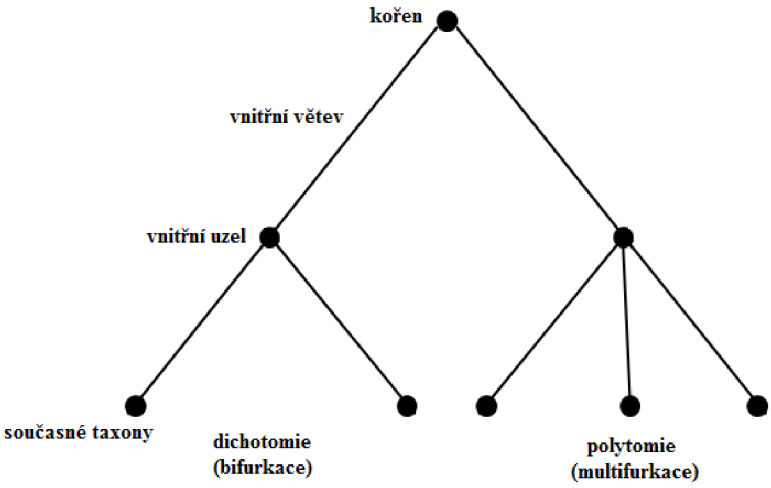
Tabulka 1: IUPAC kódy pro nukleové kyseliny [25]

IUPAC kód	význam
A	Adenin
C	Cytosin
G	Guanin
T	Thymin
U	Uracil
N	Cokoliv (any)
R	A, G (purine)
Y	C, T (pirimidine)

1.3 Fylogenetické stromy

Fylogenetický strom, nazýván také jako evoluční strom, ukazuje evoluční vztahy všech biologických druhů, které mají společného předka v minulosti. Délky větví takového stromu obvykle udávají dobu divergence dvou druhů od společného předka, který bývá znázorněn uzlem. Koncové uzly na špičkách stromů (listy) představují operační taxonomické jednotky (operational taxonomic unit; OTU). Tyto OTU odpovídají molekulární sekvenci nebo taxonu (druhu), ze kterého byl strom vytvořen. Označení taxon (systematická jednotka) označuje skupinu organismů, které mají společné znaky.

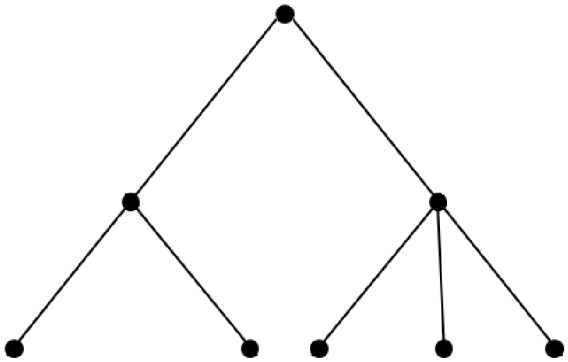
Každý strom je charakterizován prostorovým uspořádáním jednotlivých větví, takzvanou topologií. Stav, kdy se ve vnitřním uzlu nachází pouze tři větve, je označován jako bifurkace neboli dichotomie. Takový strom se označuje jako binární, plně vyřešený. V případě, že se ve vnitřním uzlu nachází více větví, jedná se o multifurkaci neboli polytomii [6], [31].



Obrázek 1: Anatomie fylogenetického stromu [10]

Zakořeněný strom

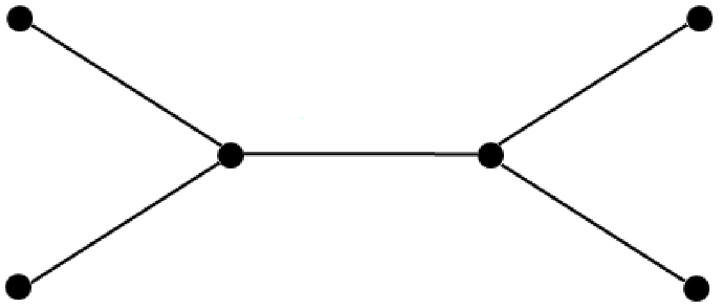
Zakořeněný strom je takový, který má uzel (kořen), ze kterého všechny OTU pocházejí. Každý další uzel, který se ve stromě nachází, představuje nejbližšího společného předka OTU z něj vycházejících. Tyto vnitřní uzly jsou většinou pouze hypotetické OTU, které není možné v současnosti pozorovat. Takový strom je potom zakončen takzvanými listy, což jsou reálné taxonomické jednotky[31].



Obrázek 2: Zakořeněný fylogenetický strom [31]

Nezakořeněný strom

Nezakořeněný strom ilustruje příbuznost uzlů bez znalosti jejich původu. Zatímco nezakořeněné stromy můžou vznikat ze zakořeněných stromů jednoduchým vynecháním kořene, tak pro vytvoření kořenového stromu z nezakořeněného je náročnější a je potřeba použít různé metody pro identifikaci původu daných OTU [31].



Obrázek 3: Nezakořeněný fylogenetický strom [31]

1.3.1 Typy stromů

Dendrogram

Stromový graf, který znázorňuje vzájemnou příbuznost či podobnost zkoumaných OTU. Sestavuje se na základě distančních (založených na matici genetických vzdáleností) nebo znakových metod.

Fenogram

Grafické znázornění stromu, vyjadřujícího vzájemnou podobnost druhů, znázorňuje tedy fenotypovou příbuznost jednotlivých OTU. Délka větví v tomto stromu reprezentuje dobu, za kterou se původní druh změní na nově vzniklý druh.

Kladogram

Strom, který znázorňuje pouze topologii. Zobrazuje pořadí a odlišnost vývojových linií. Kladogram má v systematické biologii jiný význam a není schématem kladogeneze, paradoxně ale je důležitým prvkem pro vytvoření tohoto schématu.

Fylogram

Délka větví fylogramu ukazuje evoluční vzdálenost (počet nukleotidových substitucí) mezi uzly. Je možné se setkat i s fylogramy, kde jsou důležité úhly, které svírají mezi sebou jednotlivé větve. S touto variantou se ale často nesetkáme a běžně jsou větve ve fylogramu v pravoúhlé formě [13].

1.3.2 Počet možných fylogenetických stromů

Početně náročný algoritmus, čímž bezesporu statistické vyhodnocování je, vybírá ze všech možných alternativních stromů to nejlepší možné řešení. S rostoucím množstvím taxonů roste i nárust matematických operací, které jsou potřebné pro dokončení vyhodnocení.

Pomocí čtyř taxonů je možné vytvořit pouze tři alternativní nekořenové stromy. V případě že bude určen kořen, je možné pro každý ze tří alternativních nekořenových stromů vytvořit pět různých stromů, čímž vznikne celkem 15 stromů. Počet taxonů *s*, odpovídající počtu možných nekořenových stromů je vyjádřen vztahem [31]:

$$\frac{(2s-5)!}{2^{s-3}(s-3)!} \tag{1.1}$$

Pro kořenové stromy bude platit tento vztah [28]:

$$\frac{(2s-3)!}{2^{s-2}(s-2)!} \tag{1.2}$$

Pro představu je vytvořena tabulka 3, která poskytuje výsledky pro různé počty taxonů.

Tabulka 2: Počet možných fylogenetických stromů pro daný počet sekvencí [30]

taxony	Stromy bez kořene	Stromy s kořenem
4	3	15
8	10395	135135
10	2027025	34459425
20	$2*10^{20}$	$8*10^{21}$
30	$9*10^{36}$	$5*10^{38}$
50	$3*10^{74}$	$3*10^{76}$

1.4 Metody konstrukce fylogenetických stromů

Ke konstrukci fylogenetických metod se používají především následující dvě metody. První z nich jsou distanční metody, která jsou založeny na výpočtu podobností mezi jednotlivými sekvencemi, a druhý způsob je pomocí znakových metod. V těchto metodách se používají přímo sekvence znaků s přesně vymezenými hodnotami. Obě dvě metody jsou blíže popsány v následujících podkapitolách.

Dříve, než budou popsány jednotlivé distanční a znakové metody, je třeba zmínit modely, které slouží k výpočtu evoluční vzdálenosti.

1.4.1 Modely pro výpočet evoluční vzdálenosti

Evoluční vzdálenosti jsou nezbytným parametrem pro konstrukci fylogenetických stromů. Jejich hodnoty mohou být vypočteny na základě modelů: Jukes-Cantor, Kimura, Tamura-Nei. Výběr modelu následně určuje charakteristiku výsledného stromu.

Jukes-Cantor

Jednoparametrický model označován jako J-C model. Tento model počítá s tím, že všechny čtyři nukleotidy (A, C, G a T) se vyskytují se stejnou frekvencí. Je stanovena α , jako frekvence změny kteréhokoliv nukleotidu na jiný. Z čehož vyplývá, že při přeměně báze na jinou dostaneme 3α.

Tabulka 3: Tabulka přechodů pro metodu Jukes-Cantor

	A	T	C	G
A	-3α	α	α	α
T	α	-3α	α	α
C	α	α	-3α	α
G	α	α	α	-3α

Při porovnání dvou sekvencí lze určit *D* (p-distance), místo na kterých se sekvence liší. Nelze ale určit skutečný počet změn mezi nukleotidy, protože tato metoda nepočítá s více substitucemi na jednom nukleotidu. Proto se počítá evoluční vzdálenost *d*, která je definována jako odhadovaný počet substitucí, ke kterým došlo na jednom místě. Pro výpočet průměrného počtu změn jednoduše vynásobíme frekvenci změn s časem *t*. Kde čas *t* je čas nutný k přechodu od předka k potomkovi. Celkový čas, při němž můžou nastat změny u potomků je 2*t*. Čímž dostaneme průměrný počet změn na jednom místě [17]:

$$d = 2t \times 3\alpha = 6\alpha t \tag{1.3}$$

Nicméně v praxi většinou nelze určit t nebo α dostatečně přesně. Z tohoto důvodu nemůžeme počítat *d* přímo z rovnice (1.3). Tento problém lze obejít následujícím způsobem: *d* i *D* jsou závislé na α*t* a sloučením rovnic je lze eliminovat, z čehož lze odvodit, že D závisí na čase následovně [17]:

$$D = \frac{3}{4} - \frac{3}{4}e^{-8\alpha t} \tag{1.4}$$

Další úpravou dostaneme [17]:

$$\ln\left(1 - \frac{4}{3}D\right) = -8\alpha t \tag{1.5}$$

Porovnáním s rovnicí (1.3) dostáváme [17]:

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}D\right) \tag{1.6}$$

Rovnice (1.6) nám umožňuje vypočítat evoluční vzdálenost d bez nutnosti znát frekvence výskytu změn a bez znalosti času. Parametr *D* popisuje množství pozorovaných rozdílů mezi dvěma danými sekvencemi [17],[18].

Kimura

Jedná se o dvouparametrický model, který počítá s pravděpodobností změny jednoho nukleotidu za jiný, tak jako předchozí model a navíc hodnotí pravděpodobnost změny nukleotidu v rámci své chemické skupiny (puriny nebo pirimidiny). Tento model tedy

rozlišuje frekvence transicí α a transverzí β, ale rozdílnou četnost purinů a pyrimidinů považuje za rovnocennou.

Zde je uvedena tabulka přechodů:

Tabulka 4: Tabulka přechodů pro metodu Kimura

	A	T	C	G
A	-α -2β	β	β	α
T	β	-α -2β	α	β
C	β	α	-α -2β	β
G	α	β	β	-α -2β

Vztah pro přepočet četnosti substitucí na evoluční vzdálenost je dán vztahem [19]:

$$d = -\frac{1}{2}\ln(1 - 2P - Q) - \frac{1}{4}\ln(1 - 2Q) \tag{1.7}$$

Kde P je četnost transicí a Q je četnost transverzí a zároveň platí: P + Q = D [19].

Tamura-Nei

Tento tříparametrický model rozšiřuje předchozí dvouparametrický model o třetí parametr θ, který bere v potaz, že Guanin a Cytosin mezi sebou tvoří silnou vazbu (tři vodíkové můstky), která se hůře naruší.

Tabulka 5: Tabulka přechodů pro metodu Tamura-Nei

	A	T	C	G
A	$-\alpha_1 g_G - 2\beta(g_T+g_C)$	βg_T	βg_C	$\alpha_1 g_G$
T	βg_A	$-\alpha_2 g_C - 2\beta(g_A+g_G)$	$\alpha_2 g_C$	βg_G
C	βg_A	$\alpha_2 g_T$	$-\alpha_2 g_T - 2\beta(g_A+g_G)$	Bg_G
G	$\alpha_1 g_A$	βg_T	βg_C	$-\alpha_1 g_A - 2\beta(g_T+g_C)$

Vztah pro přepočet četnosti substitucí na evoluční vzdálenost je dán vztahem [30]:

$$d = -2\theta(1-\theta)\ln\left(1 - \frac{P}{2\theta(1-\theta)} - Q\right) - \left(1 - 2\theta(1-\theta)\right)\frac{1}{2}\ln(1-2Q), \tag{1.8}$$

kde θ je proporcionální četnost všech C+G ku, všem znakům v sekvenci, P je frekvence transicí a Q frekvence translací [30].

1.4.2 Distanční metody

Základem těchto metod je matice vzdáleností, která obsahuje vzájemné vzdálenosti všech možných dvojic taxonů, pro které se fylogenetický strom konstruuje. Před sestavením matice vzdáleností je nutné sekvence zarovnat a vzájemně porovnat. Vzdálenost mezi dvěma OTU je popsána počtem bodových mutací. Tato distanční vzdálenost není skutečnou evoluční vzdáleností. Aby fylogram znázorňoval skutečnou evoluční vzdálenost, je potřeba udělat následnou korekci dat na zpětné a vícenásobné mutace se stejnou pozicí a na paralelní vzájemně nezávislé mutace.

Mezi nejčastěji používané distanční metody se řadí shluková analýza UPGMA (*Unweighted Pair-Group Method using Arithmetic averages/mean*), metoda nejbližšího souseda (*Neighbour-joining method*) a metoda minimální evoluce (*minimum evolution*) [30].

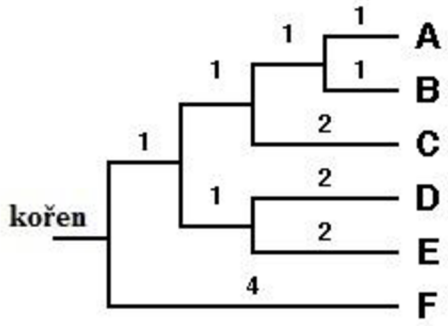
1.4.3 UPGMA

UPGMA (*Unweighted Pair-Group Method using Arithmetic averages/mean*) je nejjednodušší metodou pro konstrukci fylogenetického stromu. Fylogenetický strom konstruovaný touto metodou se nazývá fenogram. Fenogram proto, že byly původně používány jako reprezentanti rozsahu fenotypové podobnosti dané skupiny druhů organismů. Z překladu celého názvu lze odvodit základní charakteristiky této metody. *Unweighted*, neboli *neváženy* znamená, že všechny párové vzdálenosti mají stejný vliv na tvorbu stromu. *Pair-group*, v překladu *dvojice*, udává, že shluky jsou vytvářeny kombinací dvou hodnot a *Arithmetic average* přeloženo jako *aritmetický průměr*, který udává párové vzdálenosti ke každému shluku, které jsou střední hodnotou vzdálenosti ke všem členům shluku.

Konstrukce stromu metodou UPGMA začíná sestrojením trojúhelníkové matice vzdáleností každého druhu s každým. Po sestavení této matice se hledá pár taxonů, jehož vzájemná vzdálenost je ze všech nejmenší a spojení jejich větví uzlem. Poté je nalezená dvojice taxonů nahrazena shlukem (u=(x-y)). Dalším krokem je výpočet vzdáleností od ostatních taxonů do nové matice tak, že vzdálenost mezi novým shlukem *u* a dalším taxonem bude [24]:

$$d_{uk} = \frac{d_{xk} + d_{yk}}{2} \tag{1.9}$$

Z přepočítaných hodnot se vytvoří nová matice a postup se opakuje do doby, než se spojí všechny taxony. Výsledný fylogenetický strom (dendrogram) se sestrojí spojením jednotlivých shluků v pořadí, ve kterém byly spojovány [24].

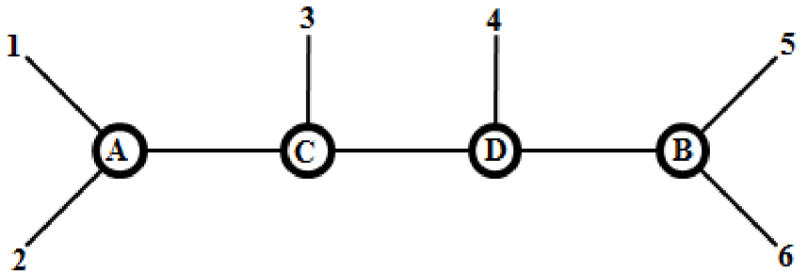


Obrázek 4: Výsledný fylogenetický strom metody UPGMA [24]

1.4.4 Neighbor-joining metoda

Spadá do distančních metod a je určena k rekonstrukci fylogenetických stromů z evoluční matice vzdáleností. Principem této metody je najít v každé fázi shlukování taxonů dvojici OTU.

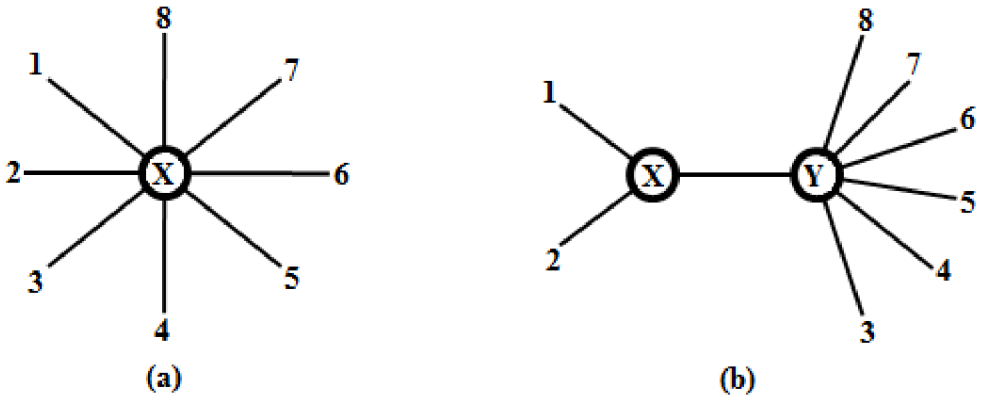
Základním konceptem NJ (Neighbor-joining method) metody jsou „sousedé“ (neighbors), kteří jsou definováni jako dva taxony, které jsou spojeny jediným uzlem v nekořenovém stromě. Například, taxon 1 a 2 ve stromě na obrázku 5 jsou sousedé, protože jsou spojeni jedním uzlem *A*. Stejně tak i taxon 5 a 6 jsou sousedé, zbylé taxony sousedé nejsou. Po spojení taxonu 1 a 2 získáme jeden taxon (1-2), který se stane sousedem taxonu 3. Je možné definovat topologii stromu spojováním sousedů a tím tvořit nové páry sousedů. Topologie stromu na obr 5 může být popsána následujícími dvojicemi sousedů: (1,2), (5,6), (1-2,3) a (1-2-3,4). Nalezením těchto párů sousedů je možné získat topologii stromu [24].



Obrázek 5: Fylogenetický strom s šesti OTU [24]

Konstrukce stromu metodou NJ začíná s hvězdicovým stromem, ve kterém nejsou žádné shluky taxonů (obrázek 6a). V praxi jsou některé taxony k sobě blíže než jiné. Tuto skutečnost popisuje obrázek 6b, ve kterém je pouze jeden vnitřní uzel *XY*, který spojuje taxony 1 a 2 s ostatními (3,4,…, N) spojenými do jednoho uzlu *Y*. Každý pár taxonů může být na pozici 1 a 2 daného stromu a je zde tedy N (N – 1)/2 možností vybrat je. Z těchto možností

je vybrána ta, jejímž výstupem je nejmenší součet délek větví. Tímto postupem jsou nalezeny všechny interní větve, kterých je N-3 [27].



Obrázek 6: a) hvězdicový strom b) strom, ve kterém byly OTU 1 a 2 sloučeny [27]

Pro výpočet sumy délek větví stromu je definováno D_{ij} a L_{ab} jako vzdálenost mezi taxony *i* a *j* délku větví mezi uzly *a* a *b* [27].

$$S_0 = \sum_{i=1}^{N} L_{ix} = \frac{1}{N-1} \sum_{i<j} D_{ij} \tag{2.0}$$

Každá větev je počítána N – 1 krát v případě že jsou přidány všechny vzdálenosti. Délka větve mezi uzly X a Y (L_{xy}) je dána [27]:

$$L_{xy} = \frac{1}{2(N-2)} \left[\sum_{k=3}^{N} (D_{1k} + D_{2k}) - (N-2)(L_{1x} + L_{2x}) - 2 \sum_{i=3}^{N} L_{iY} \right] \tag{2.1}$$

Po eliminaci vnitřní větve L_{xy} z obrázku 6b, dojde k vytvoření dvou stromů s hvězdicovou topologií. $L_{1x} + L_{2x}$ a $\sum_{i=3}^{N} L_{iY}$ můžou být získány použitím rovnice [27]:

$$L_{1x} + L_{2x} = D_{12} \tag{2.2}$$

$$\sum_{i=3}^{N} L_{iY} = \frac{1}{N-3} \sum_{3 \leq i < j} D_{ij} \tag{2.3}$$

Přidáním těchto délek větví se získá součet (S_{12}) všech délek větví stromu na obrázku 6b [27].

$$S_{12} = \frac{1}{2(N-2)} \sum_{k=3}^{N} (D_{1k} + D_{2k}) + \frac{1}{2} D_{12} + \frac{1}{N-2} \sum_{3 \leq i < j} D_{ij} \qquad (2.4)$$

Obecně není známo, které páry taxonů jsou skutečnými sousedy, a proto se počítá součet délek větví pro všechny dvojice taxonů. Dvojice sousedů je vybrána jako pár s nejmenší společnou hodnotou. V praxi může být i tato dvojice dvojicí nepravých sousedů. Pokud je zjištěno, že S_{12} je nejmenší ze všech hodnot *Sij*, potom taxony 1 a 2 jsou označeny jako pár sousedů a jsou spojeny do společného taxonu (1-2). Vzdálenost mezi tímto taxonem (1-2) a ostatními taxony je dána [27]:

$$D_{(1-2)j} = (D_{1j} + D_{2j})/2 \qquad (3 \leq j \leq N) \qquad (2.5)$$

Tímto krokem se sníží počet taxonů o jeden a vytvoří se nová distanční matice. Celý postup se pak opakuje, dokud se počet taxonů nerovná třem, nakonec zůstane pouze jeden nezakořeněný strom [27].

V dnešní době se mnohem více využívají modifikace NJ v podobě prezentované Studierem a Keplerem, známé jako metoda BioNJ (viz. [14]).

1.4.5 Minimální Evoluce

Metoda Minimální evoluce (ME) je založena na předpokladu, že strom s nejmenším odhadem součtu délek větví bude pravdivý. V minulosti byl tento předpoklad používán i přesto, že neměl žádný matematický důkaz. Teoretický základ této metody ukazuje, že odhad součtu délek větví pravého stromu je nejmenší ze všech možných stromů, za předpokladu že použité evoluční vzdálenosti jsou statisticky objektivní a délky větví jsou odhadnuty metodou nejmenších čtverců.

Metoda ME tedy zkoumá všechny možné topologie stromů, aby nalezla topologii s nejmenší hodnotou *S*. Jednou z možností je sestrojení stromu pomocí metody NJ a následně prozkoumat a najít ze souboru topologií blízkých tomuto stromu, topologii s nejmenší hodnotou *S* (dočasný ME strom). Nový soubor topologií blízký dočasnému ME stromu je znovu prozkoumán pro nalezení ještě menší hodnoty *S*. Tento proces se opakuje tak dlouho, dokud není nalezen další strom s menší hodnotou. Následně strom s nejmenší hodnotou *S*, je považován za ME strom [26], [3].

1.4.6 Znakové metody

Tyto metody používají sekvence znaků, kde na každé pozici je digitální znak s přesně vymezenou hodnotou. A pracují s pravděpodobností změny hodnot, které jsou rovny

substituci nukleotidu. Do znakových metod se řadí například metoda maximální parsimonie (Maximum Parsimony) nebo metoda maximální věrohodnosti (Maximum Likelihood), které jsou blíže popsány v následujících kapitolách.

1.4.7 Metoda maximální parsimonie

Existuje mnoho různých verzí metod maximální parsimonie (Maximum Parsimony). Tato kapitola je zaměřena pouze na metody, které jsou užitečné pro analýzu molekulárních dat. Eck a Dayhoff (viz.[7]) byli první, kteří použili MP metody pro konstrukci stromů z dat aminokyselinových sekvencí. Následně autoři Fitch (viz.[10]) a Hartigan (viz.[15]) upravili MP algoritmus pro nukleotidové sekvence. V těchto metodách jsou porovnávány čtyři nebo více zarovnaných nukleotidových sekvencí a nukleotidy předků jsou odvozeny odděleně pro každé místo dané topologie za předpokladu, že dojde k mutacím ve všech směrech mezi čtyřmi nukleotidy. Následně je vypočítán nejmenší počet nukleotidových změn, který popisuje evoluční proces topologie. Stejný výpočet je použit pro všechny potencionálně správné topologie a topologie, která vyžaduje nejmenší počet změn je vybrána jako nejlepší strom.

Tato metoda tedy porovnává vzniklé topologie a hledá fylogenetický strom vyžadující nejmenší počet nukleotidových záměn, které by musely nastat, aby tento strom odpovídal průběhu evoluce. Toto porovnání se provádí na základě skóre stromů, které je dáno součtem skóre jednotlivých pozic. V některých metodách se při hodnocení stromů přiřadí jednotlivým nukleotidovým mutacím různá váha na základě toho, že některé nukleotidy mutují snáze než ostatní.

Stejný výpočet je použit pro všechny potencionálně správné topologie a topologie, která vyžaduje nejmenší počet změn je vybrána jako nejlepší strom. Existuje více způsobů, jak je možné zefektivnit prohledávání stromů a to například metodou, která je schopna prohledávání omezit pouze na „nadějné“ stromy, tato metoda se nazývá Branch and bound [24].

1.4.8 Metoda maximální věrohodnosti

Metoda maximální věrohodnosti (Maximum Likelihood; ML) pro fylogenetické účely byla prvně prezentována autory Cavalli-Sforza a Edwards (viz.[3]) pro data tvořená genovými sekvencemi. Později s ohledem na nukleotidové sekvence, vynalezl Felstein (viz.[9]) algoritmus pro konstrukci fylogenetických stromů pomocí metody maximální věrohodnosti.

Tato metoda hledá nejpravděpodobnější topologii pro daná data. Strom, ve kterém jsou změny bazí nejvěrohodnější, je vybrán jako výsledný strom. Jinými slovy, stromu, který potřebuje pro popsání fylogenetických dat více změn, bude přiřazena menší věrohodnost než stromu, který pro popis stejných fylogenetických dat použije změn méně [24].

2 FYLOGENETICKÉ SUPERSTROMY

Metody rekonstrukce superstromů kombinují informace ze sad taxonomicky se překrývajících fylogenetických superstromů, které jsou někdy nazývány jako zdrojové nebo vstupní stromy a jejich výstupem je superstrom. Tato fylogenetická analýza vyžaduje, aby byl zdrojový strom propojen sadou sdílených taxonů. Zdrojový strom, ve kterém se nenachází společné taxony, nelze kombinovat. V případě dvou nepřekrývajících se zdrojových stromů je možné využít třetího stromu, který sdílí taxony s oběma a tím stromy zkombinovat.

První zmínka o metodách fylogenetických superstromů se objevila v 80. letech 20. století. Jednalo se o obecnou charakteristiku a soubor pravidel, jak mají být fylogenetická data ze zdrojových stromů kombinována. Pro různé metody existují různé pravidla, jejichž výsledkem by neměla být pouze kombinace informací obsažených ve zdrojových stromech, ale taky odvození vztahů, které se v jednotlivých zdrojových stromech vyskytují. Výsledný superstrom by neměl obsahovat protiřečící si vztahy jednotlivých zdrojových stromů.

Pro upřednostnění metody superstromů před alternativními metodami jsou, jak praktické, tak filozofické důvody. Z filozofického hlediska, čím více dat bude použito k vyřešení problému, tím větší bude pravděpodobnost, že bude lepší výsledek. Metody superstromů umožňují začlenění informací z různých zdrojů, jako například použití metody superstromů pro kombinaci genetických dat s morfologickými údaji. Z této kombinace vzniká datový soubor, který obsahuje jak makroevoluční tak i mikroevoluční informace umožňující prohlášení, zda vytvořit vztahy mezi evolučně vzdálenými taxony.

Z praktického hlediska, několik datových souborů obsahuje stejné druhy/kmeny i proteiny, a proto může být jejich kombinace obtížná. Distribuce skvrnitého genu znamená, že pouze několik (odhaduje se 1%) genů je univerzálně distribuováno v jedné kopii. Tradiční fylogenetické studie se spoléhají právě na toto 1% pro rekonstrukci fylogeneze organismů. V tomto případě by bylo ideální rekonstruovat fylogenetický strom z 100% dostupných dat. Superstromy jsou tedy jedinou reálnou možností využít 100% informací k rekonstrukci stromu života, jako kaskádově zarovnané geny ze všech organismů. Jediným omezením metody superstromů je, že data lze reprezentovat pouze jako stromy [2],[4].

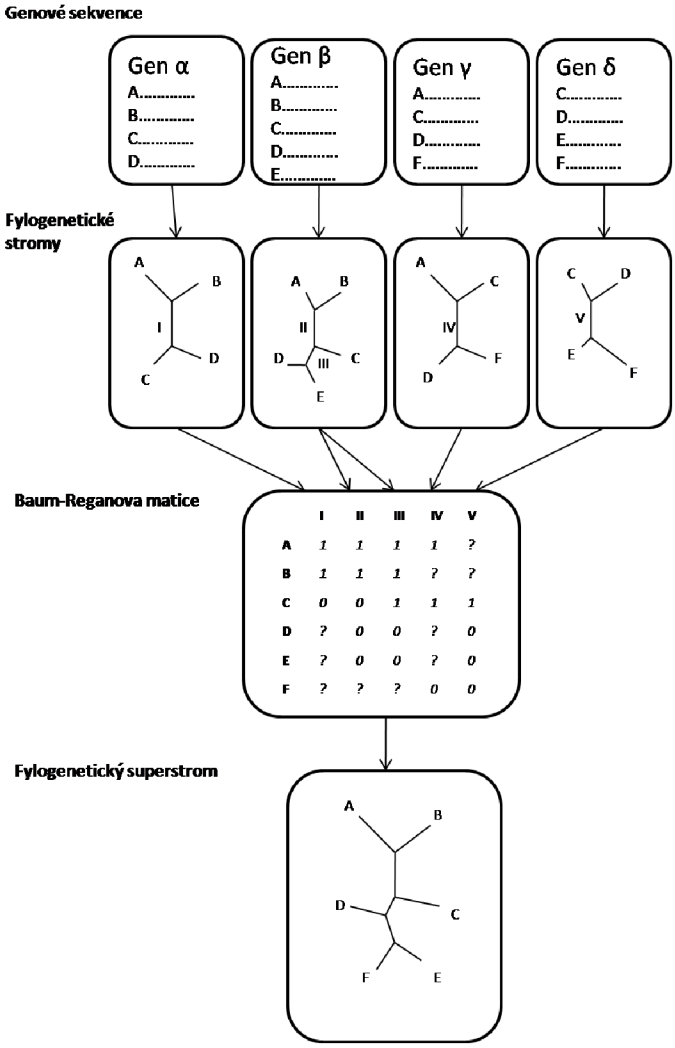
2.1 Metody rekonstrukce superstromů

Existuje mnoho metod pro rekonstrukci superstromů, které řeší problém s kombinací informací z mnoha stromů různými způsoby. V této kapitole jsou popsány tři nejpoužívanější metody založené na principu kompletace částečných vzdálenostních matic.

2.1.1 Maticová reprezentace úspornosti (MPR)

Jedná se o nejpoužívanější metodu pro rekonstrukci superstromů, kterou nezávisle na sobě popsali dva autoři, Baum a Ragan. Je nazývána jako maticová reprezentace úspornosti (MPR – Matrix Representation with Parsimony), využívá kódovací schéma pro konstrukci matice reprezentující vztahy mezi zdrojovými stromy. Využívá algoritmu maximální shody pro rekonstrukci superstromu z těchto matic.

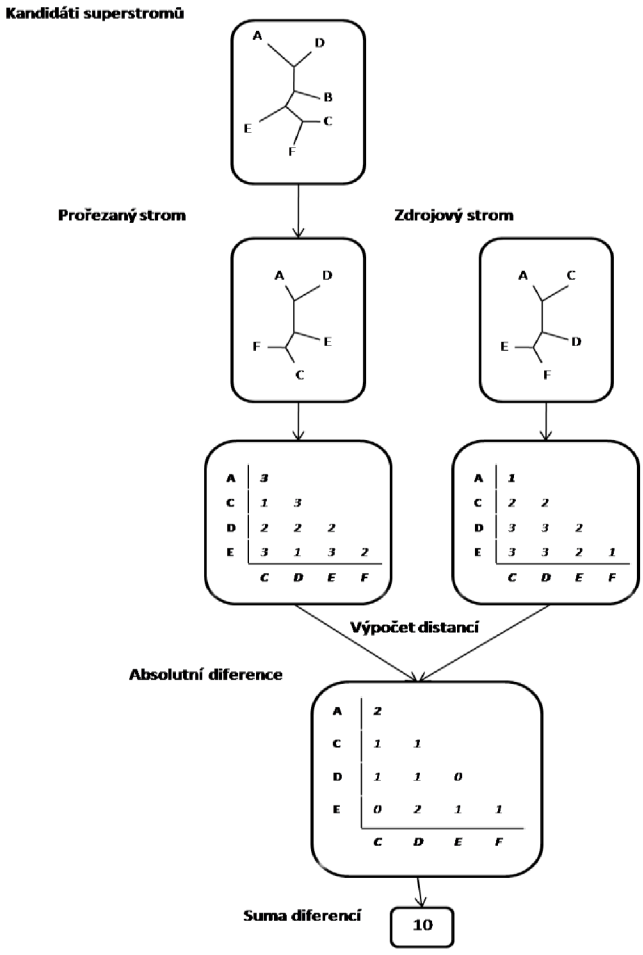
Principem této metody je určení vnitřních větví v jednotlivých zdrojových stromech a za pomoci jednoduchého kódovacího schématu 0 a 1 určí taxony na obou stranách větve. Všem taxonům nacházejících se na jedné straně je přidělena 1 a taxonům na straně druhé 0. Pokud se taxon nachází na neurčité pozici, je mu přidělen '?'. U nezakořeněných zdrojových stromů nezáleží, na které straně se použije označení 1 a 0. Kódování všech interních větví napříč všemi zdrojovými stromy je sloučeno do jedné matice a ta je použita pro rekonstrukci superstromu [2], [4].



Obrázek 7: Maticová reprezentace pomocí MPR [4]

2.1.2 MSSA

Metoda MSSA (The Most Similar Supertree Algorithm) hledá nejlepší superstrom bez průměrování informací ze zdrojových stromů. Využívá heuristické hledání superstromu s bodovací funkcí, které jsou minimalizovány a vrací strom, který se nejvíce podobá některému ze souboru zdrojových stromů. Bodovací funkce funguje tak, že kandidáta (superstrom) porovná s každým zdrojovým stromem samostatně. Vzhledem k tomu, že superstrom obsahuje všechny taxony a všechny zdrojové stromy je pravděpodobné, že obsahuje i potřebnou část pro porovnání kandidáta se zdrojovým stromem. Rozdíl mezi dvěma stromy je vypočítán sečtením absolutních rozdílů vzdáleností matic daných dvou stromů. V tomto případě je délka větví definována počtem vnitřních uzlů rozdělených dvěma taxony na stromě. Tato metoda porovnávání se provede pro každý zdrojový strom a součet absolutních odchylek se použije jako skóre reprezentující podobnost kandidáta superstromu proti sadě zdrojových stromů. Skóre odpovídající hodnotě 0 reprezentuje situaci, ve které každý zdrojový strom je identický se superstromem. Pro nalezení superstromu s minimálním bodovací funkcí je zapotřebí otestovat více kandidátů. K vyhledávání může sloužit heuristická metoda zvaná *Nearest neighbor interchange* (NNI) nebo *sub-tree pruning and regrafting* (SPR) [4].

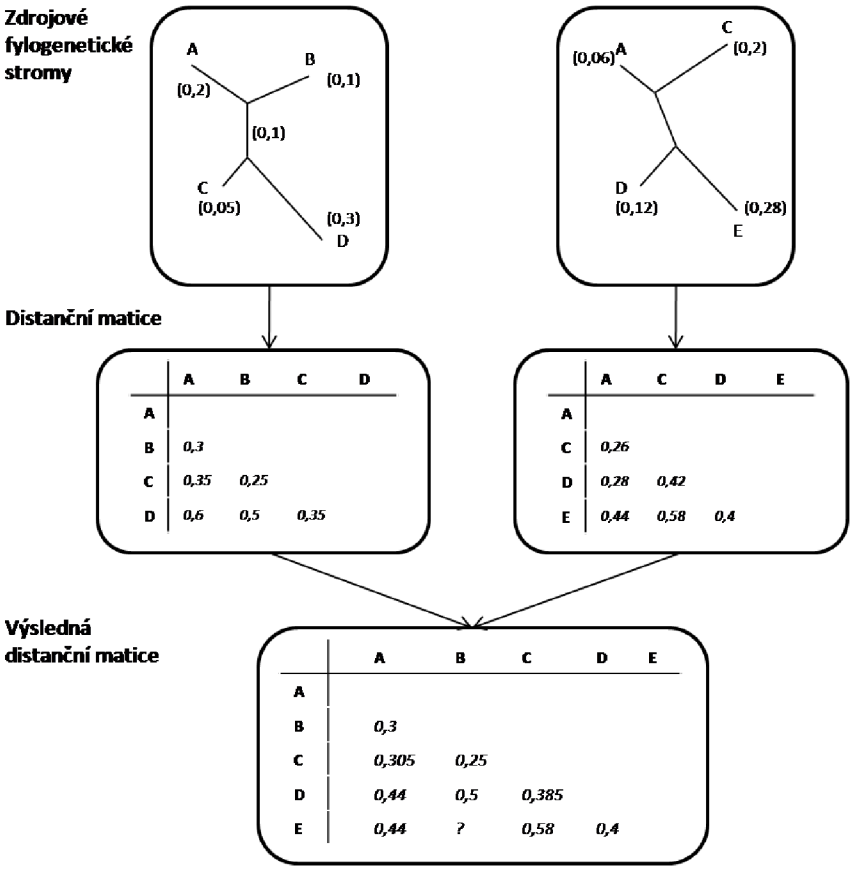


Obrázek 8: MSSA algoritmus [4]

2.1.3 Průměrný konsensus

Další možný způsob rekonstrukce superstromů zahrnuje výpočet vzdálenostní matice mezi zdrojovými stromy. Tyto metody využívají délky větví zdrojových stromů. Jednou z těchto metod je i „průměrný konsensus“. Počítá vzdálenost každého taxonu s každým skrz všechny zdrojové stromy. Průměrná vzdálenost každého taxonu je použita pro finální vzdálenostní matici, ze které je konstruován superstrom. Může se stát, že dva taxony se nikdy nevyskytnou společně v žádném ze zdrojových stromů. V tomto případě se použije průměrná vzdálenost obou taxonů k taxonům, které mají společné, a odhadne se vzdálenost, kterou by od sebe měly, kdyby se vyskytly společně na jednom stromě. Jedná se o vyplnění „prázdných míst“, pro které nejsou žádné informace o evolučních vztazích.

Jakmile je konsensuální matice kompletní, lze ji použít k rekonstrukci superstromu. Nejpoužívanější metodou je metoda nejmenších čtverců, ale je možné použít i jednodušší metodu spojování sousedů (NJ). Výhodou této metody je, že vytváří superstomy s délkami větví [4].



Obrázek 9: Metoda průměrného konsensu [4]

Prázdná místa lze dopočítat dvěma metodami: ultrametrickou a aditivní. Tyto metody využívají matematických metod k dopočítání chybějících částí ještě před fylogenetickou rekonstrukcí. Obě tyto metody tedy slouží k nalezení neznámých záznamů v evolučních datech, které se objevují, když se v pozorovaných sekvencích nukleotidů vyskytují chybějící položky nebo obsahují mezery.

Ultrametrická metoda

Vstupem ultrametrické metody je částečná matice a jejím výstupem je kompletní matice vzdáleností, pomocí které je možné konstruovat superstrom. Neznámá distanční vzdálenost *d (i, j)* je dopočítána ze dvou taxonů křížících se v tomto bodě. Výsledná hodnota poslána na výstup je maximální nalezenou hodnotou. Vyjádřeno pomocí matematického vztahu jako [1,21]:

$$d(i,j) \leq Max\{d(i,k); d(j,k)\} \forall i,j,k \in X\,, \tag{2.6}$$

kde *k* představuje taxon se známými distančními hodnotami. Hledanou hodnotou je *d (i, j)* a známými vstupními hodnotami jsou *d (i, k)* a *d (j, k).*

Chybějící hodnotu v trojúhelníku lze nalézt pouze v případě, pokud jsou dané hodnoty odlišné. V případě, že tyto dvě vzdálenosti jsou totožné, chybějící hodnotu dopočítat nelze. Ultrametrická metoda má své využití při odhadu chybějících buněk v případě, kdy vzdálenostní matice má více chybějících položek, než je přípustné [21].

Aditivní metoda

Neznámé distanční hodnoty dopočítává pomocí dvou taxonů (k, l) se známými distančními hodnotami. Součtem vertikálních hodnot taxonu *k* s horizontálními hodnotami taxonu *l* a vertikálních hodnot taxonu *l* s horizontálními hodnotami taxonu *k*, se vypočítá výsledná hodnota. Na výstup je následně poslána hodnota, která je ze získaných součtů nejvyšší, zmenšena o hodnotu vzdálenosti, ve které se tyto dva taxony protnuly. Stejně jako u předchozí metody nesmí být hodnoty součtu totožné[1,21].

Aditivní metoda j popsána vztahem [21]:

$$d(i,j) \leq Max\{d(i,k) + d(j,l); d(i,l) + d(j,k)\} - d(k,l)\,; \forall i,j,k \in X\,, \tag{2.7}$$

kde vstupem jsou neznámé distanční hodnoty *d (i, k), d (j, l), d (i, l), d (j, k)* a *d (k, l).* Hledanou hodnotou je *d (i, j)* [21].

3 POUŽITÁ DATA

Data zpracovaná v této práci jsou získána z veřejné databáze NCBI (National Centre for Biotechnology Information). Hlavní složkou této databáze je genová banka obsahující DNA sekvence, dále je zde i bibliografická databáze biomedicínské literatury a další kvalitní informace [22].

Pro porovnání získaných dat jsem zvolil mitochondriální gen označován jako 16S rRNA. Tato sekvence DNA kóduje podjednotku RNA bakteriálního ribozomu, který nese důležité informace pro zkoumání evoluční příbuznosti. Tento gen se vyskytuje u všech bakterií a jeho související forma se vyskytuje ve všech buňkách včetně eukaryotických. Při analýze tohoto genu bylo zjištěno, že v některých částech molekuly dochází k rychlým změnám, čímž se odlišují jednotlivé druhy stejného rodu, zatímco ostatní pozice se mění velmi pomalu, což naopak umožňuje rozlišit taxonomické úrovně [8].

Pro srovnání bylo zvoleno pět tříd: savci, ryby, ptáci, plazi a hmyz. První čtyři jmenované se řadí do kmene strunatců a podkmene obratlovců. Hmyz se řadí do kmene členovců. Zakomponování hmyzu má své opodstatnění a to, že fylogenetické superstromy se využívají právě pro takové případy, kdy se srovnávají i evolučně vzdálené organismy. Každou třídu reprezentuje 10 živočišných zástupců žijících na různých místech zeměkoule.

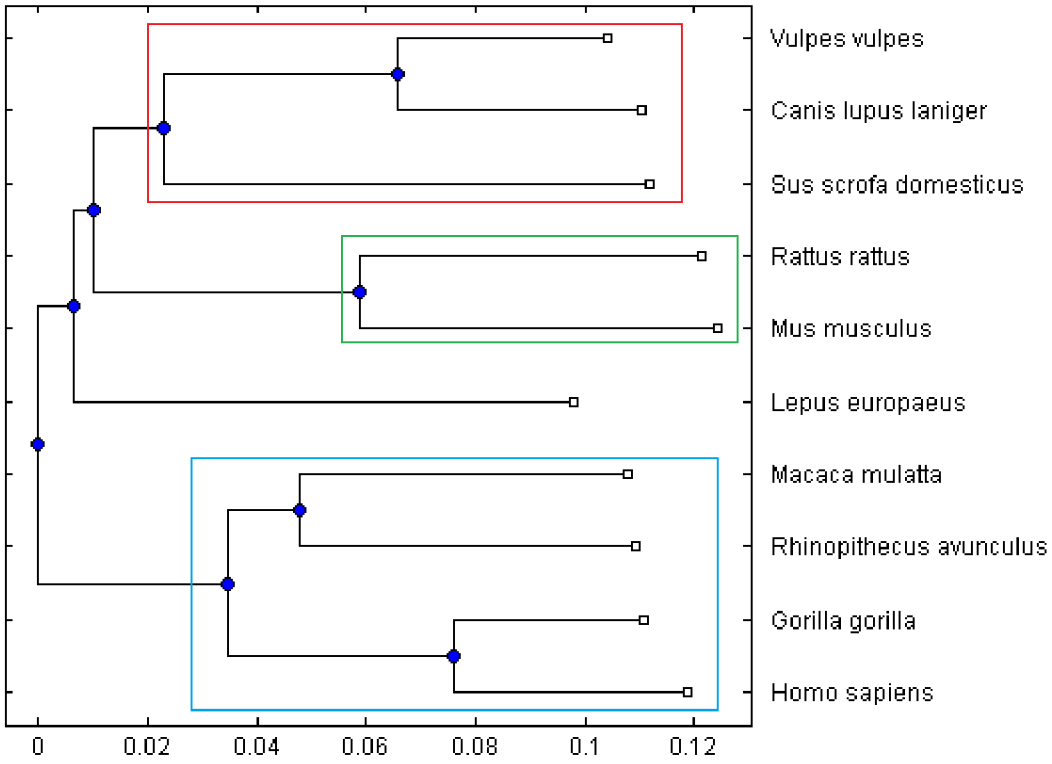
Získaná data jsou nejdříve zarovnána pomocí vícenásobného zarovnání, v programovém prostředí MatLab je již vytvořená funkce *multialign.* Následně je zvolena metoda rekonstrukce fylogenetického stromu. V tomto případě jsem zvolil metodu Neighbor-joining, v níž jednotlivé délky větví fylogenetického stromu znázorňují evoluční vzdálenost mezi porovnávanými dvojicemi.

3.1 Savci (Mammalia)

V současné době osidlují téměř celý svět. Jejich společným znakem je výživa mláďat pomoci mléčné žlázy (*mammae* – odtud tedy název *Mammalia*). V tabulce je vybráno deset zástupců, kteří reprezentují osm čeledí a pět řádů. Pro lepší přehlednost jsou tito zástupci seřazeni podle řádu, což odpovídá i shlukům ve fylogenetickém stromu.

Tabulka 6: Vybraní zástupci třídy savců

Identifikátor	Latinský název	Český název	Řád	Čeleď
353731398	Vulpes vulpes	Liška obecná	šelmy	psovití
7534303	Canis lupus	Vlk obecný		psovití
459353574	Sus scrofa	Prase domácí	sudokopytníci	prasovití
209361434	Rattus rattus	Krysa obecná	hlodavci	myšovití
34538597	Mus musculus	Myš domácí		myšovití
21492444	Lepus europaeus	Zajíc polní	zajíci	zajícovití
47156210	Macaca mulatta	Makak rhesus	primáti	kočkodanovití
297340638	Rhinopithecus avunculus	Langur indočínský		kočkodánovití
157504417	Gorilla gorilla	Gorila západní		hominidi
296100116	Homo sapiens	Člověk moudrý		hominidi



Obrázek 10: Fylogenetický strom deseti zástupců třídy savců

Popis: červená barva – řád šelem spolu s čeledí prasovitých, zelená barva – čeleď hlodavců, modrá barva – řád primátů.

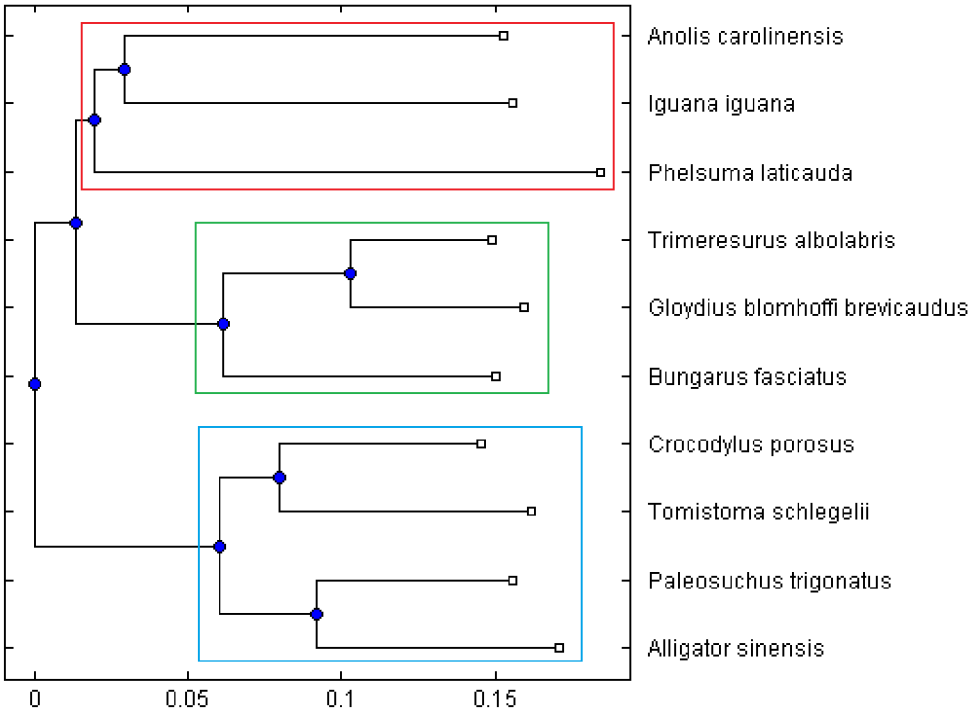
Vytvořený fylogenetický strom je možné rozdělit do tří shluků odpovídajících molekulární klasifikaci. Všichni z výše uvedených zástupců, se řadí do nadřádu placentálové (*Placentalia*). První shluk (červený) tvoří zástupci z řádu šelem, k nimž je vzdáleně příbuzný řád sudokopytníků. Druhý shluk (zelený) je tvořen dvěma zástupci řádu hlodavců. Největší shluk (modrý) reprezentuje nadřád řád primáti (*Primates*) do něhož se řadí i člověk.

3.2 Plazi (Reptilia)

Tělo plazů je pokryto šupinami nebo štíty. Jsou živočichy se studenou krví a jejich mláďata se líhnou plně vyvinutá. Existují tři řády: krokodýli (*Crocodilia*), šupinatí (*Squamata*) a želvy (*Testudines*), jejichž zástupce se v tabulce níže neobjevuje. V tabulce se nachází deset zástupců reprezentujících dva řády, krokodýli a šupinaté. Řád šupinatých je zastoupen dvěma podřády, a to ještěry (*Sauria*) a hady (*Serpentes*). Ve fylogenetickém stromě jsou naznačeny shluky, které odpovídají tomuto členění.

Tabulka 7:Vybraní zástupci třídy plazů

Identifikátor	Latinský název	Český název	Řád	Čeleď
317185859	Anolis carolinensis	Anolis rudokrký	šupinatí	leguánovití
14193045	Iguana iguana	Leguán zelený		leguánovití
325127056	Phelsuma laticauda	Felsuma zlatoocasá		gekonovití
545691539	Trimeresurus albolabris	Chřestýšovec zelený		zmijovití
194395429	Gloydius blomhoffi	Ploskovec východní		zmijovití
170774093	Bungarus fasciatus	Bungar proužkovaný		korálovcovití
82395885	Crocodylus porosus	Krokodýl mořský	krokodýli	krokodýlovití
194442184	Tomistoma schlegelii	Krokodýl úzkohlavý		krokodýlovití
154818301	Paleosuchus trigonatus	Kajmánek malý		aligátorovití
27228457	Alligator sinensis	Aligátor čínský		aligátorovití



Obrázek 11: Fylogenetický strom deseti zástupců třídy plazů

Popis: červená barva – čeleď leguánovitých spolu se zástupcem gekonovitých, zelená barva – čeleď zmijovitých spolu se zástupcem korálovcovitých, modrá barva – řád krokodýlů.

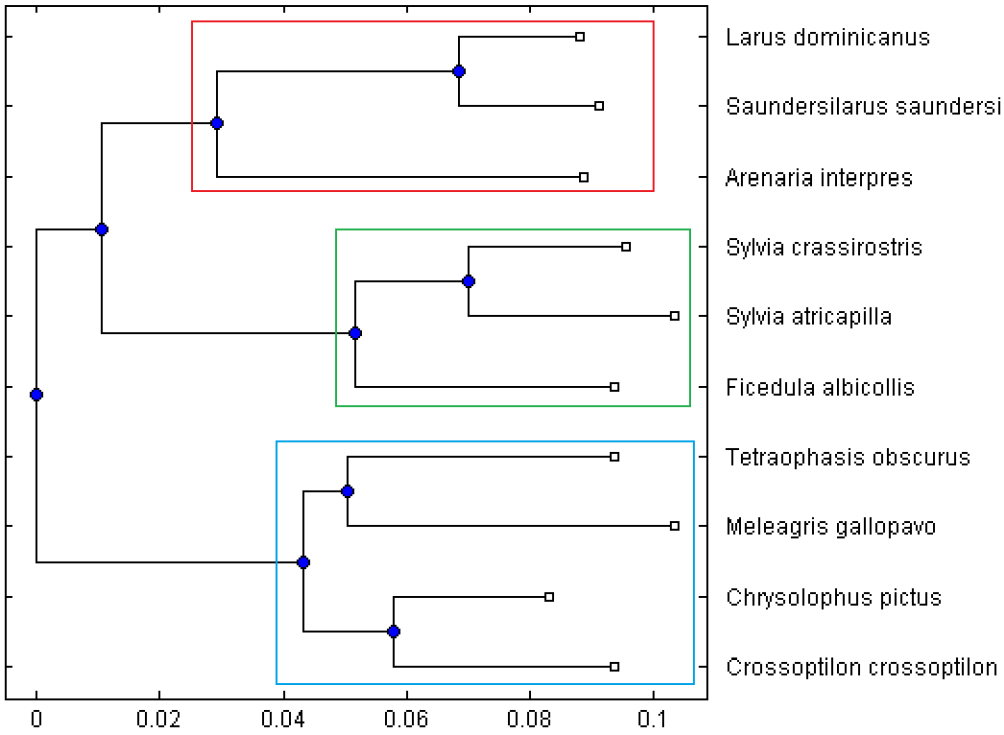
Výsledný strom je rozdělen do tří shluků. V prvním shluku (červeném) jsou zástupci řádu šupinatých (*Squamata*) z podřádu ještěrů (*Sauria*) a ve druhém shluku (zeleném) se nachází podřád hadů (*Serpentes*) ze stejného řádu jako v prvním případě. Největším shlukem (modrým) je řád krokodýlů (*Crocodilia*), který je zastoupen čtyřmi reprezentanty.

3.3 Ptáci (Aves)

Jejich hlavním znakem je přítomnost peří a trojprstých předních končetin přeměněných na křídla. Třída ptáci se dále dělí na dvě podtřídy, běžce (*Peleognathae*) a letce (*Neognathae*). V tabulce níže je uvedeno 10 zástupců, kteří spadají do třídy letců a jsou rozděleni do tří řádů a šesti čeledí.

Tabulka 8: Vybraní zástupci třídy ptáků

Identifikátor	Latinský název	Český název	Řád	Čeleď
34223665	Larus dominicanus	Racek jižní	dlouhokřídlí	rackovití
371925333	Saundersilarus saundersi	Racek čínský		rackovití
21391469	Arenaria interpres	Kameňáček pestrý		slukovití
164449777	Sylvia crassirostris	Pěnice dlouhozobá	pěvci	pěnicovití
164422147	Sylvia atricapilla	Pěnice černohlavá		pěnicovití
515021884	Ficedula albicollis	Lejsek bělokrký		lejskovití
334850229	Tetraophasis obscurus	Kur hnědohrdlý	hrabaví	bažantovití
157928635	Meleagris gallopavo	Krůta domácí		bažantovití
238550648	Chrysolophus pictus	Bažant zlatý		bažantovití
325069379	Crossoptilon crossoptilon	Bažant tibetský		bažantovití



Obrázek 12: Fylogenetický strom deseti zástupci třídy ptáků

Popis: červená barva – řád dlouhokřídlých, zelená barva – řád pěvců, modrá barva – řád hrabavých.

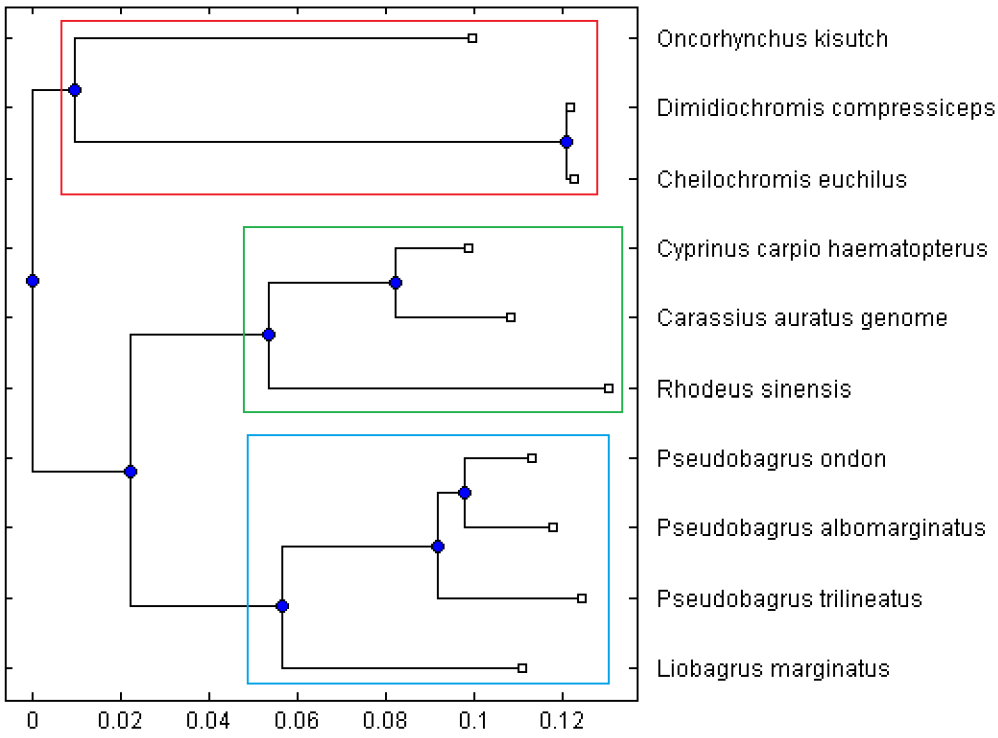
První shluk (červený) příslušného fylogenetického stromu tvoří ptáci z řádu dlouhokřídlých (*Charadriiformes*) zastoupených dvěma čeleďmi. Druhý shluk (zelený) je tvořen pěvci (*Passeriformes*) a taktéž zahrnuje tři druhy ze dvou čeledí. Poslední, třetí shluk (modrý) reprezentuje čeleď hrabavých (*Galliformes*).

3.4 Ryby (Osteichthyles)

Vyznačují se skřelemi kryjícími žábry a osifikovanou osou. Jejich přirozeným prostředím je voda, sladká i slaná. Určité druhy ryb jsou obojživelné a můžou tedy nějaký čas trávit na souši. Nadtřída ryby se dále člení na dvě podtřídy, paprskoploutvé (*Actinopterygii*) a nozdraté (*Sarcopterygii*). Podtřída nozdratí reprezentuje živé fosilie a její zástupci nejsou obsaženi v tabulce. V tabulce se tedy nachází zástupci paprskoploutvých ryb, konkrétně deset zástupců ze čtyř řádů a pěti čeledí.

Tabulka 9: Vybraní zástupci třídy ryb

Identifikátor	Latinský název	Český název	řád	čeleď
134303077	Oncorhynchus kisutch	Losos kisuč	lososotvární	lososovití
403531964	Dimidiochromis compressiceps	Cichlida plochohlavá	ostnoploutví	vrubozubcovití
470231560	Cheilochromis euchilus	Tlamovec pyskatý		vrubozuvcovití
395806662	Cyprinus carpio	Kapr obecný	máloostní	kaprovití
5835778	Carassius auratus	Karas stříbřitý		kaprovití
556506159	Rhodeus sinensis	Hořavka čínská		kaprovití
556506215	Pseudobagrus ondon	Sumíček čeťiangský	sumci	sumíčkovití
556506229	Pseudobagrus albomarginatus	Sumíček bělolemý		sumíčkovití
556505962	Pseudobagrus trilineatus	Sumíček třípruhý		sumíčkovití
560877648	Liobagrus marginatus	Bystřinovec minský		bystřinovcovití



Obrázek 13: Fylogenetický strom deseti zástupců třídy ryb

Popis: červená barva – čeleď zástupce lososotvárných spolu s řádem ostnoploutvých zelená barva – řád máloostných, modrá barva – řád sumců.

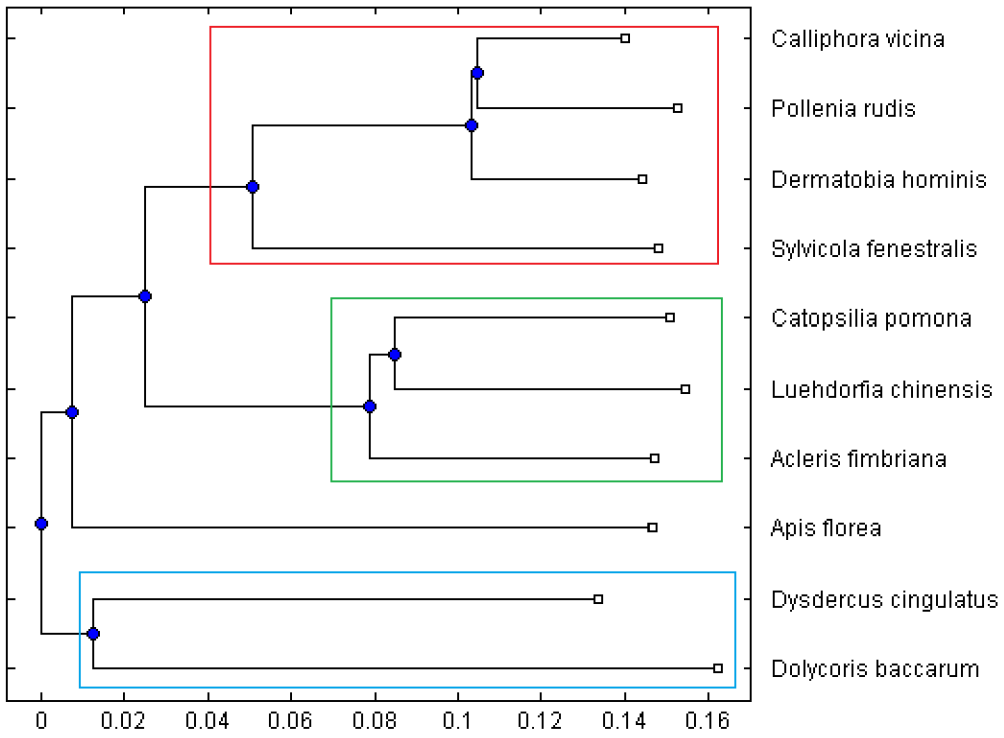
Jak lze vidět, fylogenetický strom je opět rozdělen na tři shluky. První shluk (červený) vznikl spojením řádů lososotvární (*Salmoniformes*) a ostnoploutví (*Perciformes*). Druhý shluk (zelený) představuje řád máloostní (*Cypriniformes*) a třetí (modrý) nejpočetnější představuje řád sumců (*Siluriformes*).

3.5 Hmyz (Insecta)

Tělo hmyzu je rozděleno do tří článků (hlava, hruď, zadeček). Charakteristickým znakem je, že mají tři páry nohou, tykadla a většinou složené oči. Hmyz je taky nejvíce různorodou skupinou živočichů na Zemi a vyskytuje se v každém prostředí naší planety. Třída hmyz se dělí na dvě podtřídy a to na bezkřídlé (*Apterygota*) a křídlaté (*Pterygota*). V tabulce je opět deset zástupců třídy hmyzu ze čtyř řádů a devíti čeledí.

Tabulka 10: Vybraní zástupci třídy hmyzu

Identifikátor	Latinský název	Český název	řád	čeleď
409031931	Calliphora vicina	Bzučivka obecná	dvoukřídlí	bzučivkovití
409031945	Pollenia rudis	Bzučivka zední		bzučivkovití
356489015	Sylvicola fenestralis	Stružilka okenní		stružilkovití
38324537	Dermatobia hominis	Střeček lidský		střečkovití
459284219	Catopsilia pomona	Žluťásek kasiový	motýli	běláskovití
205277633	Luehdorfia chinensis	Jasoň čínský		otakárkovití
316926259	Acleris fimbriana	Obaleč kliklový		obalečovití
428674257	Apis florea	Včela květná	blanokřídlí	včelovití
166897748	Dysdercus cingulatus	Ploštice bavlníková	polokřídlí	ruměnicovití
449785358	Dolycoris baccarum	Kněžice chlupatá		kněžicovití



Obrázek 14: Fylogenetický strom deseti zástupců třídy hmyzu

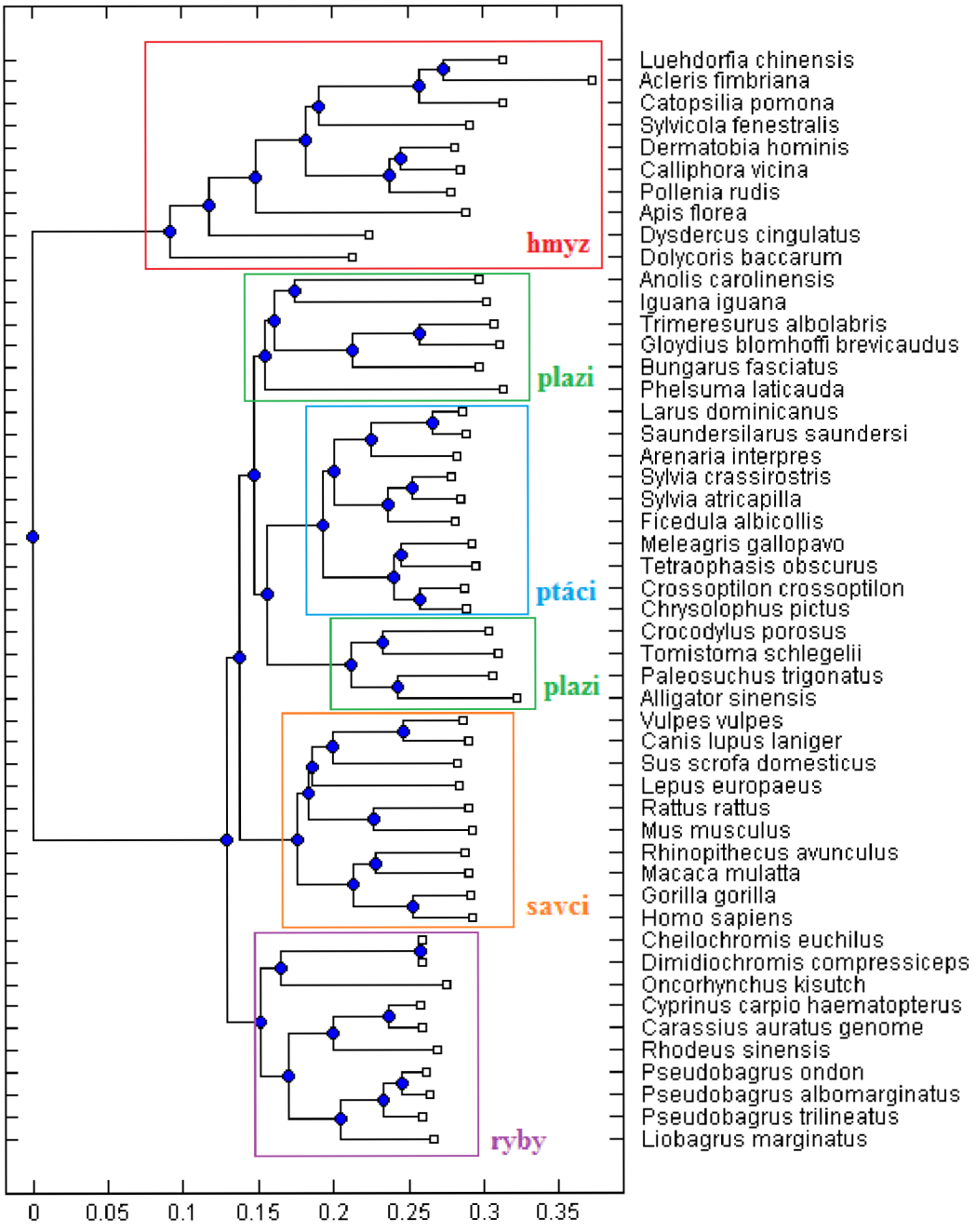
Popis: červená barva – řád dvoukřídlých, zelená barva – řád motýlů, modrá barva – řád polokřídlých.

Ze sestrojeného fylogenetického stromu jsou patrné tři shluky. První shluk (červený) obsahuje řád dvoukřídlých (*Diptera*). Druhý shluk (zelený) přestavuje řád motýlů (*Lepidoptera*) a třetí (modrý) shluk představuje řád polokřídlých (*Hemiptera*).

3.6 Fylogenetický strom

Tento fylogenetický strom je konstruován stejnou metodou, jako výše uvedené stromy. Je sestrojen z padesáti nukleotidových sekvencí, které byly použity výše.

Z níže uvedeného stromu je možné vyčíst základní informace. Je rozdělen do dvou shluků, reprezentovaných dvěma kmeny živočichů, členovci (*Arthropoda*) a strunatci (*Choradata*). Kmen členovců je zastoupen pouze jednou třídou, ta je znázorněna ve fylogenetickém stromě červeně a jedná se o třídu hmyzu (*Insecta*). U této třídy lze předpokládat, že dosahuje nejvyšších hodnot distancí, a proto je nejvíce vzdálena od ostatních druhů. Ve druhém shluku se nachází více malých shluků a obsahuje čtyři třídy. Třída plazů (*Reptilia*), znázorněna zelenou barvou je rozdělena ještě na dva shluky, řád krokodýlů (*Crocodilia*) a šupinatých (*Squamata*). Další třída (zelená) reprezentuje zástupce ptáků (*Aves*) následována oranžově označenou třídou savců (*Mammalia*) a poslední třída (fialovou) znázorňuje ryby (*Osteichthyes*).



Obrázek 15: Klasický fylogenetický strom z padesáti sekvencí

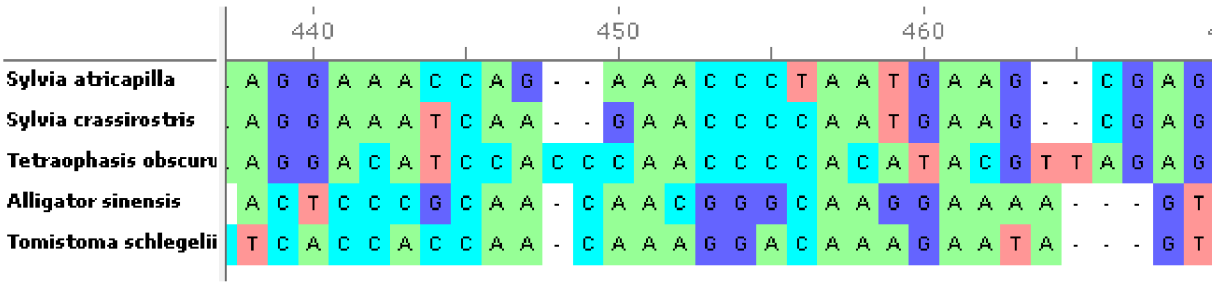
4 REALIZACE METODY PRŮMĚRNÉHO KONSENSU

Tato kapitola obsahuje popis tvorby fylogenetických stromů, jejíž součástí je práce se sekvencemi, výpočet evolučních vzdáleností a tvorba distančních matic daných fylogenetických stromů. Následně jsou popsány dvě metody dopočtu chybějících distančních vzdáleností při konstrukci fylogenetických superstromů.

4.1 Realizace fylogenetického stromu

V první části programu se zvolí, s kolika databázemi sekvencí bude pracovat, proměnná *vstup* (minimálně je možné počítat se dvěma databázemi, maximum není omezeno) a metoda, kterou bude dopočítávat neznámé distance, proměnná *metoda* (‘u‘ pro volbu ultrametrické metody, ‘a‘ pro volbu aditivní metody). Vstupní databáze sekvencí musí být ve formátu FASTA. Pro tuto práci jsou použity databáze z kapitoly tři, kde jsou podrobně představeni jednotliví zástupci daných tříd. Pro účely vývoje a pochopení jednotlivých principů metod je vybráno šest zástupců. Jmenovitě jsou vybráni následující: tři zástupci ze třídy plazů (*Alligator sinensis – Aligátor čínský*, *Tomistoma schlegelii – Krokodýl úzkohlavý* a *Crocodylus prosus – Krokodýl mořský*) a tři zástupci ze třídy ptáků (*Sylvia crassirostris – Pěnice dlouhozobá*, *Sylvia atricapilla – Pěnice černohlavá* a *Tetraophasis obscurus – Kur hnědohlavý*). Tito zástupci jsou rozděleni do dvou databází tak, že v databázi plazů jsou přidáni dva ptáci a v databázi ptáků jsou přidáni dva plazi. Vzniká tak sedmdesátiprocentní překrytí a ve výsledné distanční matici bude potřeba dopočítat jednu neznámou distanci.

Poté, co jsou požadované soubory vybrány a načteny pomocí příkazu *fastaread*, je nutné je zarovnat, kvůli jejich různým délkám. K tomu v Matlabu slouží příkaz *multialign*, který je určen k vícenásobnému zarovnání různě dlouhých sekvencí.



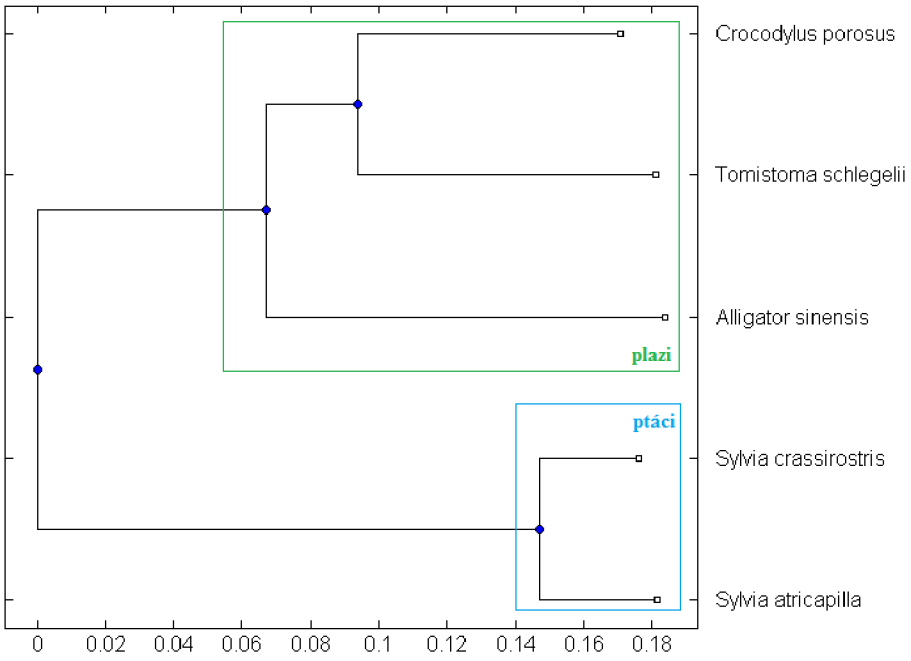
Obrázek 16: Ukázka vícenásobného zarovnání sekvencí

Následně je vypočítána (pomocí příkazu *seqpdist*) jejich evoluční vzdálenost podle modelu Jukes-Cantora, jehož princip je blíže popsán v podkapitole 1.4.1.

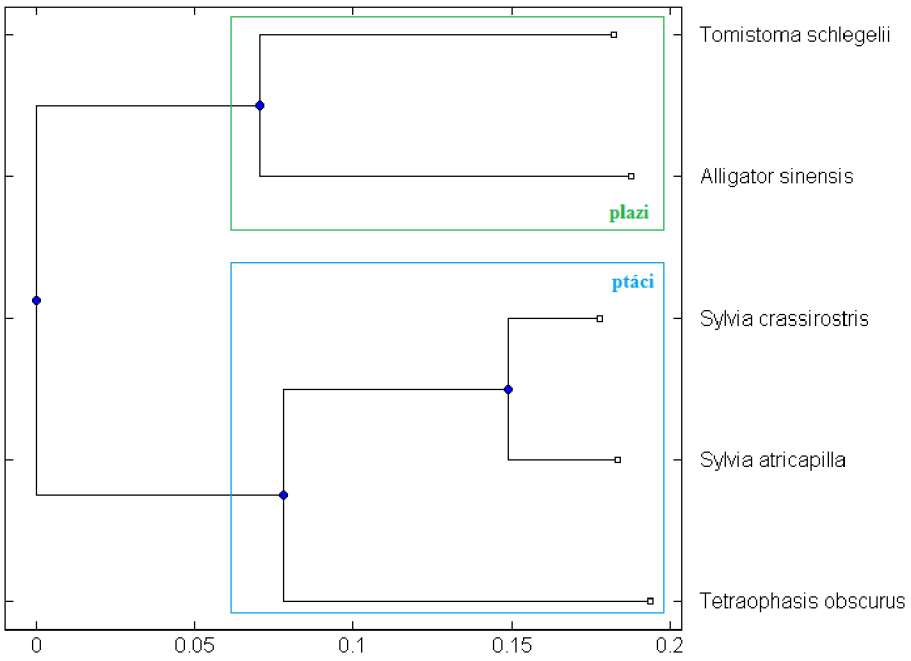
```
dist(u).distance = seqpdist(ma,'method','Jukes-Cantor','indels','pair');
```

Posledním krokem po výpočtu evoluční vzdálenosti každého jedince s každým je sestrojení fylogenetického stromu. Ten je sestrojen pomocí příkazu *seqneighjoin* a pro sestrojení stromu je zvolena metoda nejbližšího souseda, která je blíže popsána v podkapitole 1.4.4.

```
tree = seqneighjoin(final_mat_dist,'equivar',long_jvek);
```



Obrázek 17: Fylogenetický strom tří zástupců plazů a dvou zástupců ptáků



Obrázek 18: Fylogenetický strom tří zástupců ptáků a dvou zástupců plazů

4.2 Realizace fylogenetického superstromu

Druhá část programu se zaměřuje na realizaci fylogenetického superstromu ze stromů zdrojových (fylogenetických stromů). Pro jednotlivé vstupující databáze sekvencí jsou vytvořeny všechny možné kombinace jmen daných taxonů. Tato jména jsou vytvářena z taxonomických názvů tak, že jsou ukládána na místa v matici v takovém pořadí, jako jsou v maticích uloženy distance jejich křížení. V této fázi programového kódu jsou použity příkazy *strcmp* a *strcat*, které slouží k porovnávání a spojování jmen daných sekvencí. Zde jsou dvě možnosti pro uložení do výsledné distanční matice.

- Porovnávaná jména se shodují (na diagonále matice). V tomto případě je na pozici jejich křížení uložena nula.
- Porovnávaná jména se liší a uloží se kombinace obou jmen.

Na základě všech zjištěných kombinací jmen v dané databázi sekvencí se vypočítá distanční matice.

Tabulka 11: Distanční matice tří plazů a dvou ptáků

	Aligator	Tomistoma	Crocodylus	Sylvia a.	Sylvia c.
Aligator	0	0,2387	0,2127	0,3669	0,3582
Tomistoma	0,2387	0	0,1645	0,3535	0,3507
Crocodylus	0,2127	0,1645	0	0,3595	0,3547
Sylvia a.	0,3669	0,3535	0,3595	0	0,0628
Sylvia c.	0,3582	0,3507	0,3547	0,0628	0

Tabulka 12: Distanční matice tří ptáků a dvou plazů

	Sylvia a.	Sylvia c.	Tetraophasis	Aligator	Tomistoma
Sylvia a.	0	0,0628	0,2225	0,3734	0,3611
Sylvia c.	0,0628	0	0,2143	0,3636	0,3636
Tetraophasis	0,2225	0,2143	0	0,3806	0,3772
Aligator	0,3734	0,3636	0,3806	0	0,2287
Tomistoma	0,3611	0,3639	0,3772	0,2287	0

Sestrojení superstromu ze zdrojových stromů probíhá stejně jako tvorba základních fylogenetických stromů. Jsou tedy vytvořeny všechny možné kombinace jmen z obou databází. Jednotlivé kombinace jmen se porovnávají s již existujícími kombinacemi v jednotlivých fylogenetických stromech. V tomto okamžiku nastávají tři situace:

1. Kontrolovaná kombinace jmen se nachází v obou stromech, existují tedy dvě distanční hodnoty pro stejnou kombinaci taxonů. Do výsledné distanční matice jsou obě

hodnoty zprůměrovány jednoduchým aritmetickým průměrem. Výsledná hodnota je dosazena výsledné distanční matice superstromu.

2. Kontrolovaná kombinace jmen se nachází pouze v jednom stromu a existuje tak pouze jedna distanční hodnota pro daný pár jmen. Tato hodnota je následně zapsána do výsledné distanční matice superstromu.
3. Třetí možnost vznikne, když dojde ke kombinaci jmen, kdy jedno jméno je z jedné databáze sekvencí a druhé jméno z druhé databáze sekvencí. V tomto případě není známa žádná distanční hodnota a do výsledné distanční matice superstromu je uložena hodnota -1 (-1 proto, že záporná distance není možná).

Tabulka 13: Výsledná distanční matice superstromu tří plazů a tří ptáků

	Aligator	Crocodylus	Sylvia a.	Sylvia c.	Tetraophasis	Tomistoma
Aligator	0	0,2126	0,3639	0,2862	0,36360	0,2337
Crocodylus	0,2126	0	0,3594	0,3547	-1	0,3668
Sylvia a.	0,3639	0,3594	0	0,0628	0,2224	0,2628
Sylvia c.	0,2862	0,3547	0,0628	0	0,2142	0,3617
Tetraophasis	0,3636	-1	0,2224	0,2142	0	0,3771
Tomistoma	0,2337	0,3668	0,2628	0,3617	0,3771	0

Popis: červená hodnota – neznámá distance, zelené hodnoty – známé hodnoty distance z matice tří plazů a dvou ptáků, modré hodnoty – známé hodnoty distance z matice tří ptáků a dvou plazů, fialové hodnoty – zprůměrované známé hodnoty distancí z obou matic.

Ve výsledné distanční matici superstromu lze vidět evoluční vzdálenosti mezi jednotlivými jedinci. Dochází zde k výskytu záporné hodnoty, která značí, že zmíněná evoluční vzdálenost není známa a je nutné ji dopočítat. V této práci jsou sestrojeny dvě metody pro dopočet neznámých hodnot. Jedná se o metodu ultrametrickou a aditivní, které jsou popsány v podkapitole 2.1.3. V následujících kapitolách bude popsán a znázorněn princip dopočtu obou zmíněných metod.

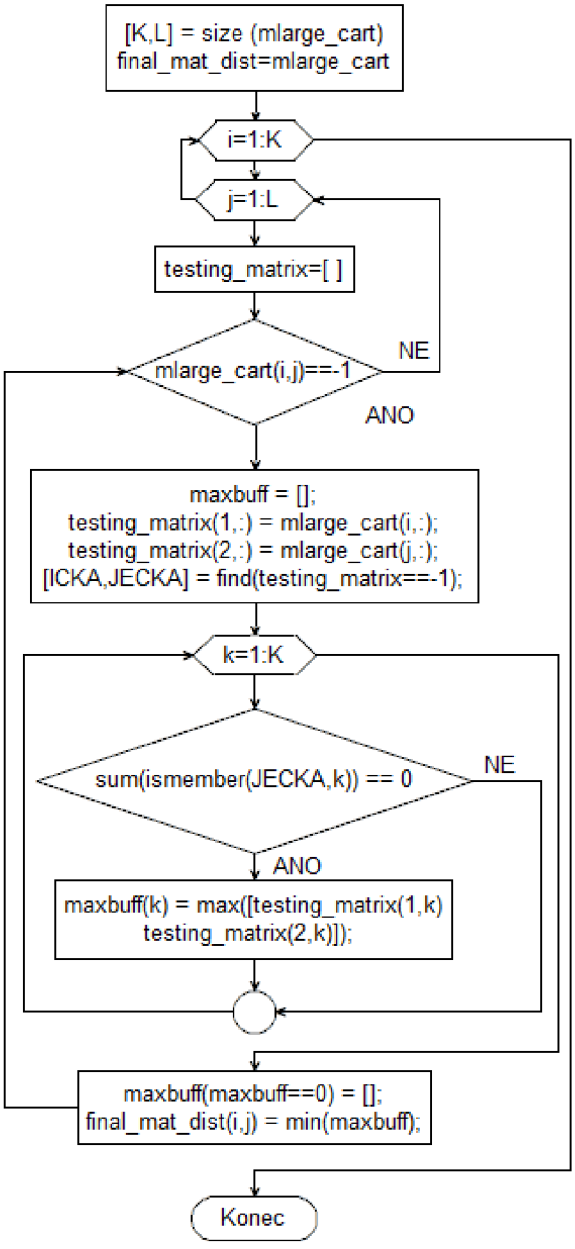
4.3 Ultrametrická metoda

Ultrametrická metoda dopočítává neznámou hodnotu distance $d(i, j)$ z taxonů, které se v dané pozici kříží a jim společných taxonů se známou distancí.

Program postupně prochází po řádcích distanční matici. V místě, kde objeví zápornou hodnotu (pozice, ve které je neznámá distanční hodnota), vybere řádky taxonů, které se na této pozici kříží. Poté tato metoda prochází po sloupcích pozice ve vybraných řádcích, ve kterých je známa distance obou taxonů. V případě, že se v jednom z řádků objeví nula nebo

záporná hodnota, je tento sloupec přeskočen, naopak pokud mají oba taxony známou distanci k taxonu jim společnému, je vybrána větší hodnota ze dvou porovnávaných. Tyto hodnoty se ukládají do vektoru *maxbuff*. Poté, co program projde všechny dvojice vybraných řádků, znovu prochází vektor *maxbuff* a hledá v něm nejmenší hodnotu. Jako původně neznámá hodnota distance dvou taxonů je vybráno minimální maximum porovnávaných hodnot distancí v řádcích, které spolu sdílejí neznámou distanci.

Na obrázku 19 je uveden vývojový diagram ultrametrické metody.



Obrázek 19: Vývojový diagram ultrametrické metody

Pro snazší pochopení principu ultrametrické metody, je zde uvedena tabulka, na které je graficky znázorněno, které řádky jsou vybrány (dvojice hodnot, ze kterých se vybírá maximum) pokud je neznámá distance mezi *x* a *y*. Tyto hodnoty jsou označeny zelenou barvou a neznámá záporná hodnota, která je aktuálně dopočítávána, má barvu červenou. Pod tabulkou je znázorněn výběr distance do proměnné *maxbuff*, ze které se na konci vybere ta nejmenší hodnota a tato hodnota je doplněna na místo neznámé distance.

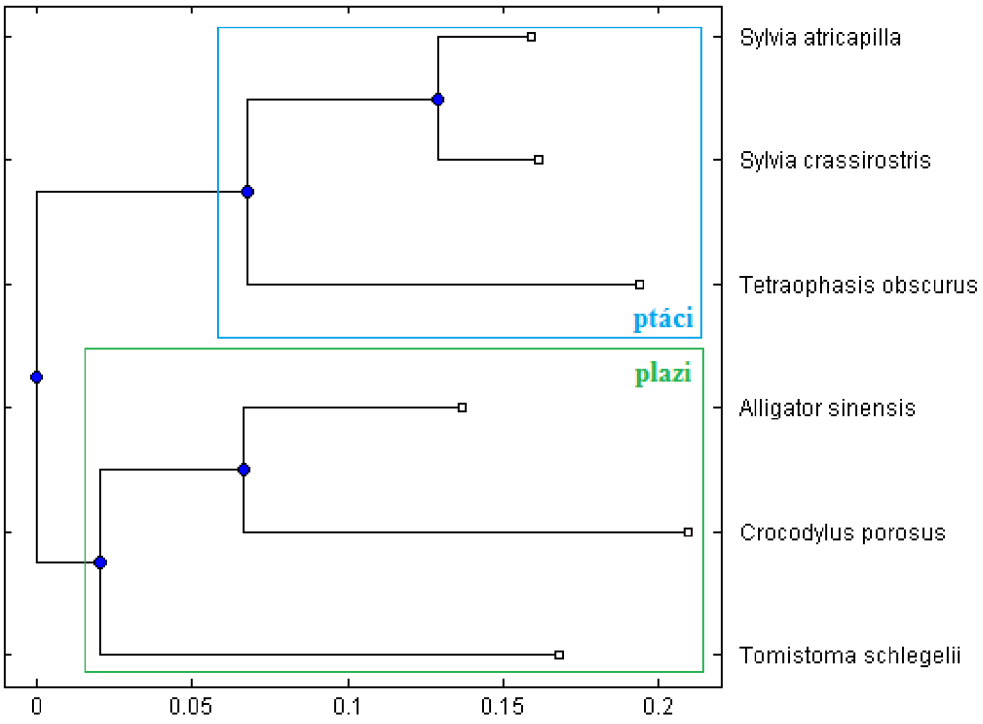
Tabulka 14: Dopočítaná distanční matice ultrametrickou metodou

	Aligator	Crocodylus	Sylvia a.	Sylvia c.	Tetraophasis	Tomistoma
Aligator	0	0,2126	0,3639	0,2862	0,36360	0,2337
Crocodylus	0,2126	0	0,3594	0,3547	0,3547	0,3668
Sylvia a.	0,3639	0,3594	0	0,0628	0,2224	0,2628
Sylvia c.	0,2862	0,3547	0,0628	0	0,2142	0,3617
Tetraophasis	0,3636	0,3547	0,2224	0,2142	0	0,3771
Tomistoma	0,2337	0,3668	0,2628	0,3617	0,3771	0

maxbuff	0,3636		0,3594	0,3547		0,3771

Popis: červená hodnota – neznámá hodnota distance, která byla vypočítána ultrametrickou hodnotou, zelené hodnoty – hodnoty distancí, ze kterých ultrametrická metoda vybírá maximum do vektoru *maxbuff*, řádek maxbuff – znázornění vybraných maxim ze dvou hodnot v daném sloupci

V původním principu této metody platí pravidlo, které v případě, kdy je vybrané minimální maximum menší než maximum v celé matici, nahradí maximální hodnotou, která se nachází v matici. Po testování velkého množství dat, a především dat s malým překrytím společných jedinců, jsem zjistil, že vznikne matice s velkým množstvím stejných distancí a data jsou tak zavádějící. Proto jsem princip v této práci upravil tak, aby po nalezení dané hodnoty již nezkoumal, zda je v celé distanční matici větší hodnota.



Obrázek 20: Ultrametrická metoda – ukázkový fylogenetický strom

4.4 Aditivní metoda

Aditivní metoda dopočítává neznámou hodnotu distance *d(i, j)* z taxonů, které se v dané pozici kříží a taxonů, které se nachází v těsné blízkosti.

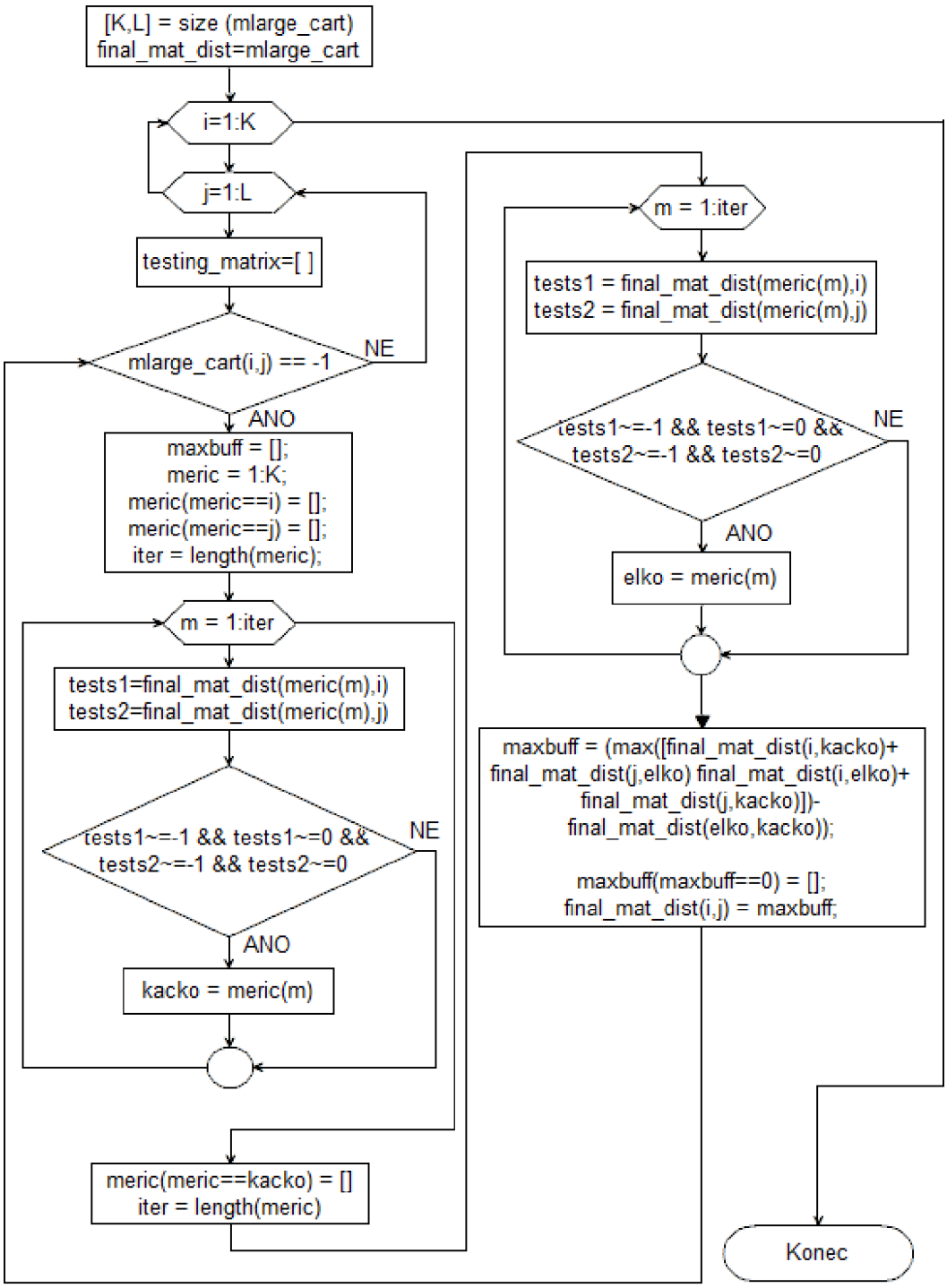
Program prochází distanční matici a v případě, že objeví zápornou hodnotu (chybějící distanční hodnota) vybere dané řádky, odpovídající taxonům, které se v dané pozici kříží. Aby bylo možné dopočítat neznámou hodnotu, je potřeba znát další dva taxony (řádky) a to *k*-tý a *l*-tý. *K*-tý taxon, který má známou distanci jak s *i*-tým tak s *j*-tým taxonem, hledá program tak, že prochází distanční matici a hledá řádek, kde se na pozici křížení s *i*-tým, a *j*-tým taxonem nenachází nula ani záporná hodnota. Pokud je daný taxon nalezen, uloží si ho a stejným způsobem hledá *l*-tý taxon. V případě, že jsou nalezeny potřebné taxony, provede se vzoreček, kterým se získá výsledná hodnota distance pro *d(i, j)*.

$$d(i,j) \leq Max\{d(i,k) + d(j,l); d(i,l) + d(j,k)\} - d(k,l)\ ; \forall i,j,k \in X\ , \qquad (2.9)$$

Jelikož základní popis funkce neřešil problém, který se vyskytoval, když bylo překrytí zdrojových stromů nízké (nebylo možné nalézt potřebné pozice pro dosazení do vzorce 2.9), navrhl jsem vlastní řešení tohoto problému. Po nalezení neznámé distance kód projde celou

matici a zjistí, ze kterých pozic je možné neznámou distanci dopočítat. Poté, co tuto hodnotu dosadí do distanční matice, hledá další hodnotu ve stejné distanční matici, do které dosadila neznámou hodnotu. Tímto způsobem pokračuje, než dopočítá všechny neznámé distance.

Na obrázku 21 je uveden vývojový diagram aditivní metody.



Obrázek 21: Vývojový diagram aditivní metody

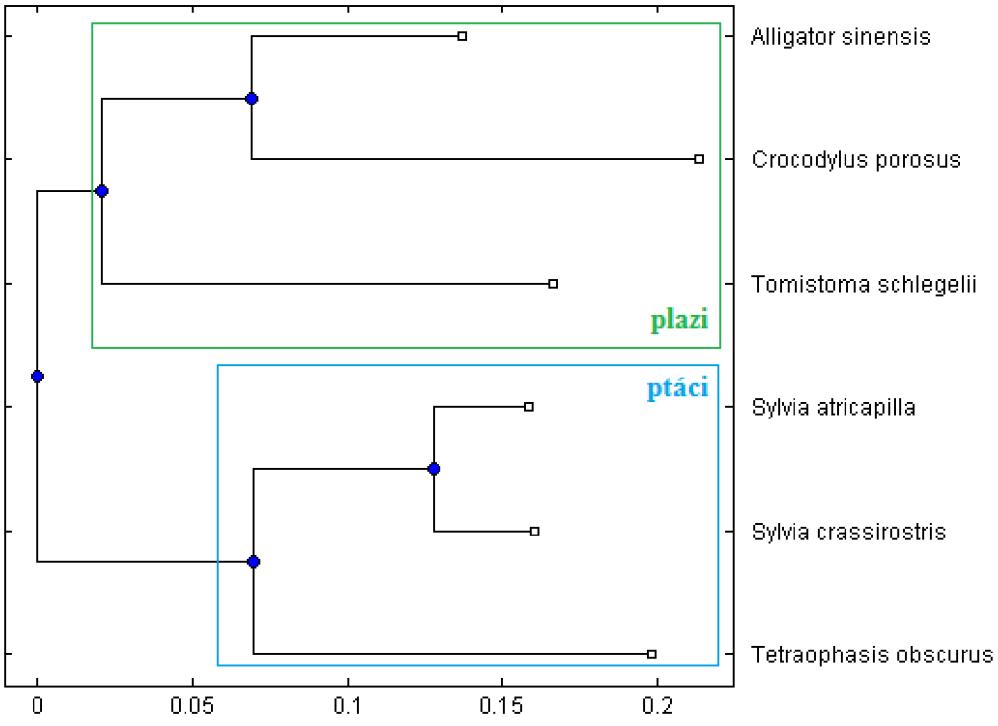
Princip výběru hodnot aditivní metodou, je graficky znázorněn v tabulce 15. Červeně je označena neznámá distance, která se aktuálně dopočítává. Zeleně jsou označeny pozice, ve kterých se kříží *i*-tý, *j*-tý, *k*-tý a *l*-tý taxon tak, aby odpovídaly vzorci 2.9. Dosazením těchto hodnot do vzorce 2.9 se získá neznámá distance.

Tabulka 15: Dopočítaná distanční matice aditivní metodou

	Aligator	Crocodylus	Sylvia a.	Sylvia c.	Tetraophasis	Tomistoma
Aligator	0	0,2126	0,3639	0,2862	0,36360	0,2337
Crocodylus	0,2126	0	0,3594	0,3547	0,3702	0,3668
Sylvia a.	0,3639	0,3594	0	0,0628	0,2224	0,2628
Sylvia c.	0,2862	0,3547	0,0628	0	0,2142	0,3617
Tetraophasis	0,3636	0,3702	0,2224	0,2142	0	0,3771
Tomistoma	0,2337	0,3668	0,2628	0,3617	0,3771	0

Popis: červená hodnota – neznámá hodnota distance vypočítaná aditivní metodou, zelené hodnoty – známé distanční hodnoty, které se dosazují do vzorce aditivní metody pro výpočet neznámé distance.

Je nutné poznamenat, že obě metody pracují s celou distanční maticí. V případě, kdyby pracovaly pouze s její trojúhelníkovou částí, nebylo by možné dopočítat jedinou hodnotu, protože obě metody vyžadují pracovat s celými řádky. Což plyne i z tabulek, které jsou přiloženy k popisu konstrukce obou metod.



Obrázek 22: Aditivní metoda - ukázkový fylogenetický superstrom

Vzhledem k tomu, že se dopočítává pouze jedna hodnota z distanční matice, výsledný rekonstruovaný superstrom metodou aditivní je totožný se stromem rekonstruovaným ultrametrickou metodou. Touto cestou jsme si ověřili, že metody dokážou správně určit dva shluky a přiřadit k sobě nejbližší taxony.

5 TESTOVÁNÍ A ANALÝZA DAT

V předchozích podkapitolách byly představeny dvě metody dopočtu neznámých distancí v distanční matici. Různý překryv zdrojových stromů a použitá konstrukční metoda vede k rozdílným výsledkům pro stejná vstupní data. Je tedy nezbytné tato data objektivně vyhodnotit a rozlišit do jaké míry odpovídá topologie superstromů referenční topologii. K tomu poslouží Robinson-Fouldova vzdálenost a Pearsnův korelační koeficient.

Robinson-Fouldova vzdálenost

Principem této metody je srovnání dvou stromů na základě výskytu stejných shluků. Pro vyhodnocení nejlepšího stromu je nutné určit si strom referenční. S tímto stromem budou ostatní stromy srovnávány pomocí následujícího vzorce [25]:

$$dRF = (n_{c1-2} + n_{c2-1})/n, \tag{3.0}$$

kde $n_{c1\text{-}2}$ je počet shluků v prvním stromu, které se nevyskytují ve druhém stromu a $n_{c2\text{-}1}$ je počet shluků ve druhém stromu, které se nevyskytují v prvním stromu, n je počet shluků obou stromů [25].

Nejlepší strom je takový, jehož Robinson-Fouldova vzdálenost k referenčnímu stromu je co nejblíže nule.

Pearsonův korelační vzorec

Slouží ke stanovení zkreslení vstupních dat, v našem případě dvou stromů. Stromy srovnává na základě jejich distancí a na základě toho určí, jak těsný vztah k sobě stromy mají.

Koeficient korelace je bezrozměrné číslo a může nabývat hodnot o mínus jedna do plus jedna. Kladné znaménko naznačuje přímou závislost a naopak záporné znaménko závislost nepřímou [29].

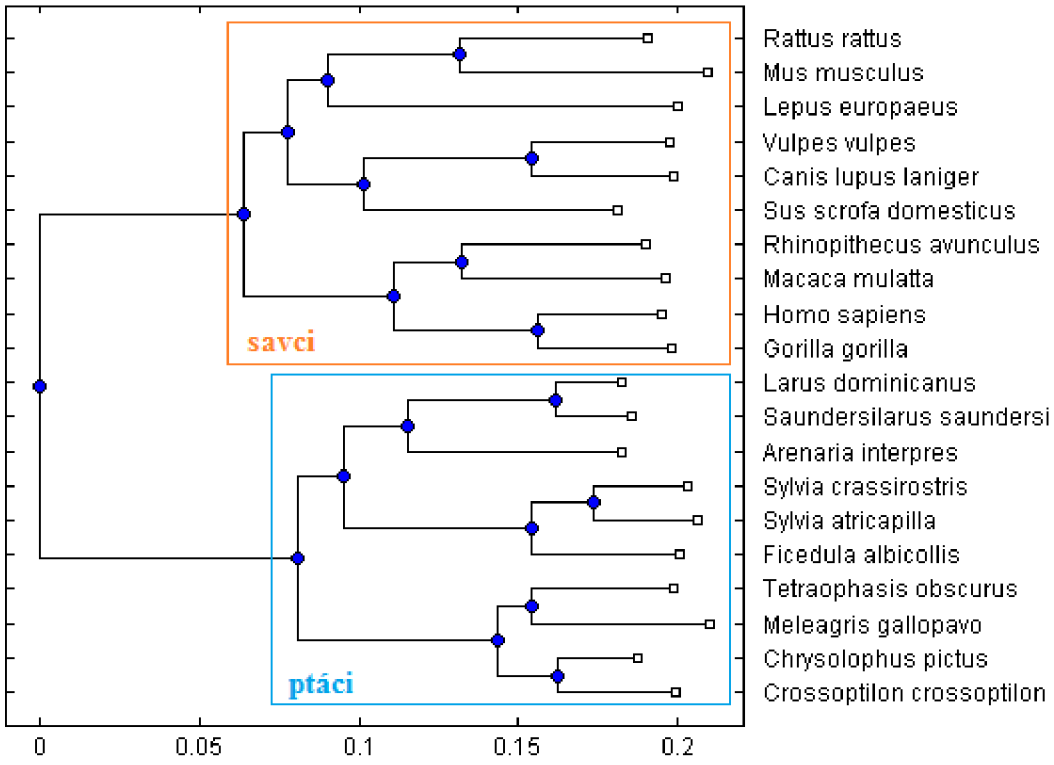
Tabulka 16: korelační koeficient [29]

0	Žádný vztah
0 – 0,3	Zanedbatelný vztah
0,3 – 0,5	Nepříliš těsný vztah
0,5 – 0,7	Středně těsný vztah
0,7 – 0,9	Velmi těsný vztah
0,9 – 1	Extrémně těsný vztah
1	pevná závislost

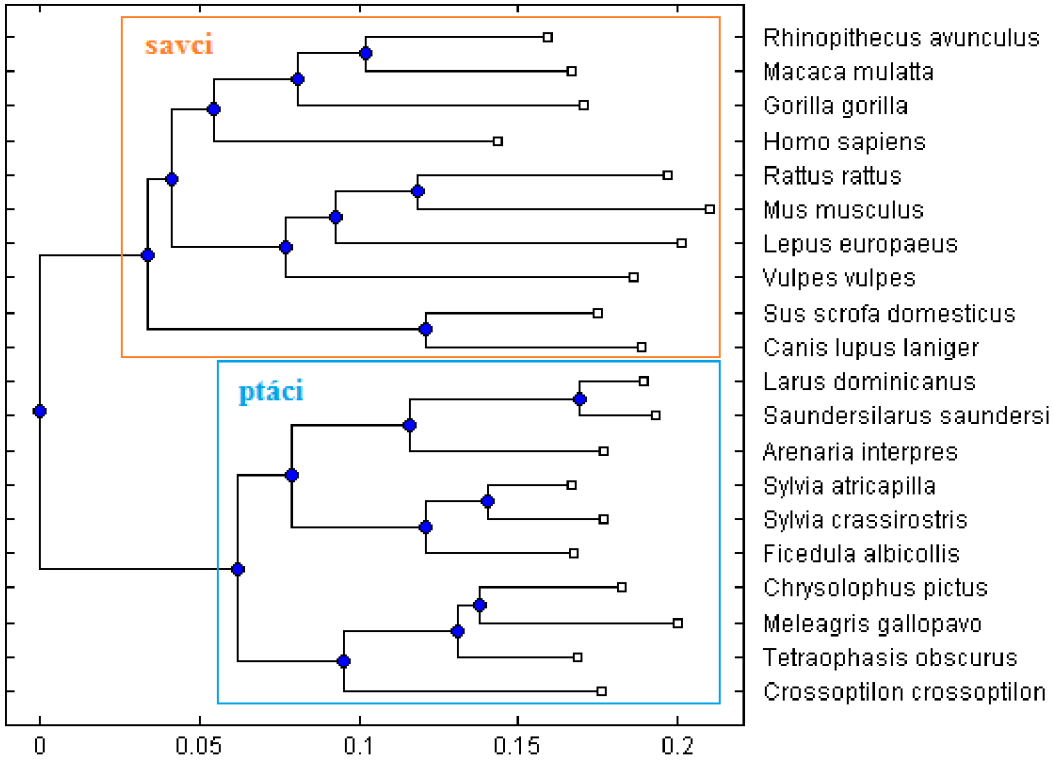
5.1 Ultrametrická metoda

Přesnost a kvalita ultrametrické metody je testována na náhodně vybraných dvojicích databází z kapitoly 3, které jsou sestaveny tak, že spolu sdílejí jeden až devět taxonů z deseti (deseti až devadesáti procentní překrytí stromů). Referenční strom je sestrojen jako běžný fylogenetický strom a obsahuje všechny zástupce porovnávaných databází. K referenčnímu stromu jsou následně srovnávány superstromy s různým procentem stejných taxonů.

V referenčním stromu (obrázek 23) jsou správně zachovány třídy savců (*Mammalia*) a ptáků (*Aves*) a jednotlivé řády jsou správně odděleny. Tedy shluk savců je rozdělen správně na řád hlodavců, ke kterým je přiřazen řád zajíců, řád šelem a řád primátů. Ve shluku ptáků jsou to pak správně zařazení jedinci do řádu dlouhokřídlých, řádu pěvců a řádu hrabavých.

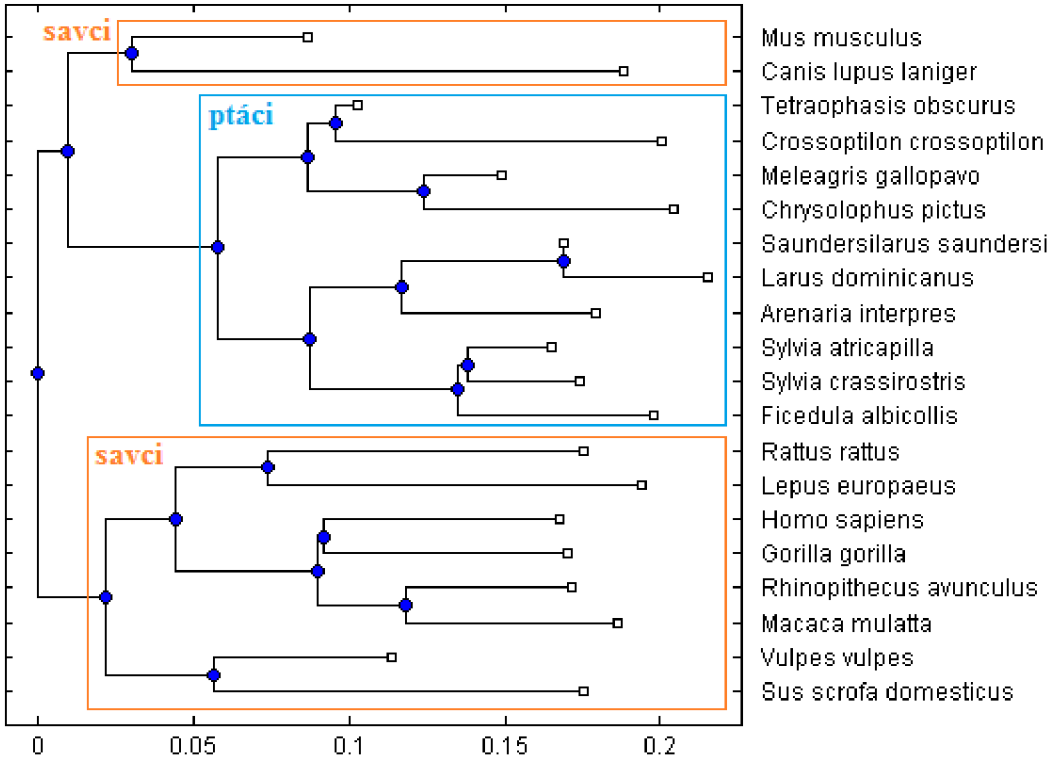


Obrázek 23: Referenční fylogenetický superstrom tříd savců a ptáků



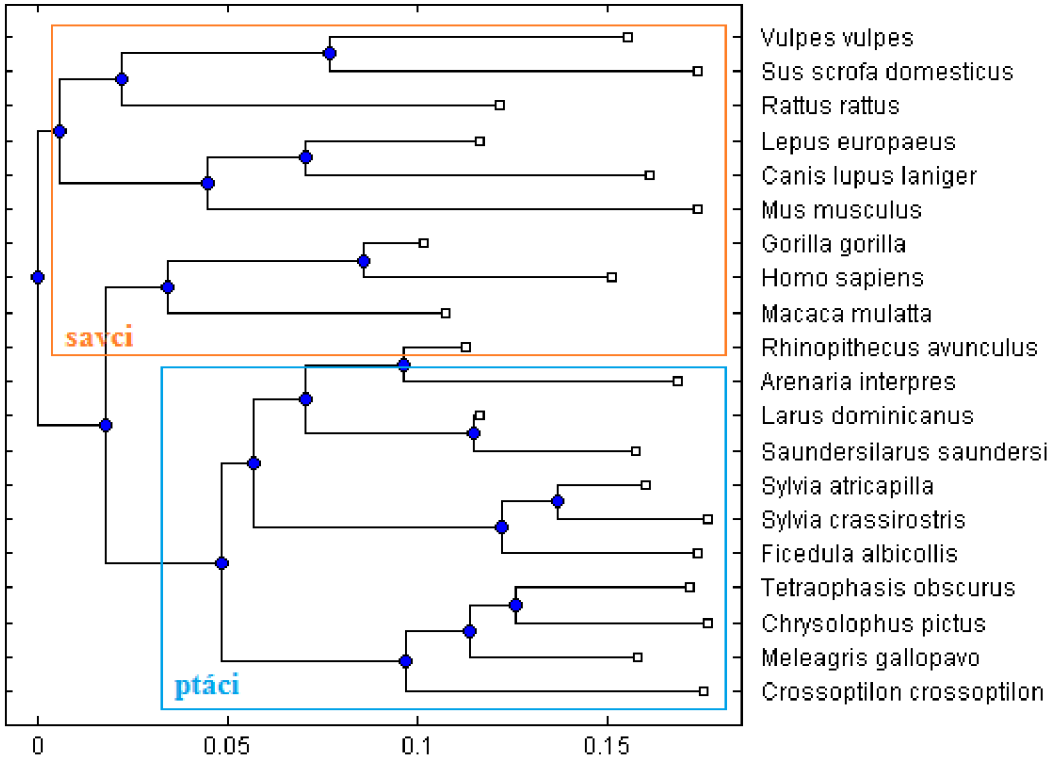
Obrázek 24: Ultrametrická metoda: fylogenetický superstrom tříd savců a ptáku s 90 % překrytím

Superstrom s devadesátiprocentním překrytím má správně vytvořeny shluky tříd stejně jako referenční strom. Jednotlivé uspořádání se již liší, což můžeme vidět na shluku primátů, který již není v jednom uzlu, k řádu hlodavců společně se zajícem je chybně přiřazen *Vulpes vulpes* (šelma). Shluk třídy ptáků se téměř nezměnil, až na shluk reprezentující hrabavé, který má jiné uspořádání, ale jedinci jsou zařazeni správně.



Obrázek 25: Ultrametrická metoda: fylogenetický superstrom tříd savců a ptáku s 60 % překrytím

Superstrom s šedesátiprocentním překrytím lze ještě klasifikovat jako strom s velmi těsným vztahem podle korelačního koeficientu (viz tabulka 16). Superstrom má správně vytvořeny dva shluky obsahující třídu ptáků (*Aves*) a třídu savců (*Mammalia*). Ve shluku třídy ptáků se nachází špatně zařazený uzel savců, který spojuje hlodavce s šelmou. Správně zůstávají zařazeny a spojeny do shluku řády hrabavých, dlouhokřídlých a pěvců. Ve shluku savců však zůstal správně pouze shluk řádu primátů, ostatní řády se rozpadly a jsou spojeny s jedinci jiných řádů.



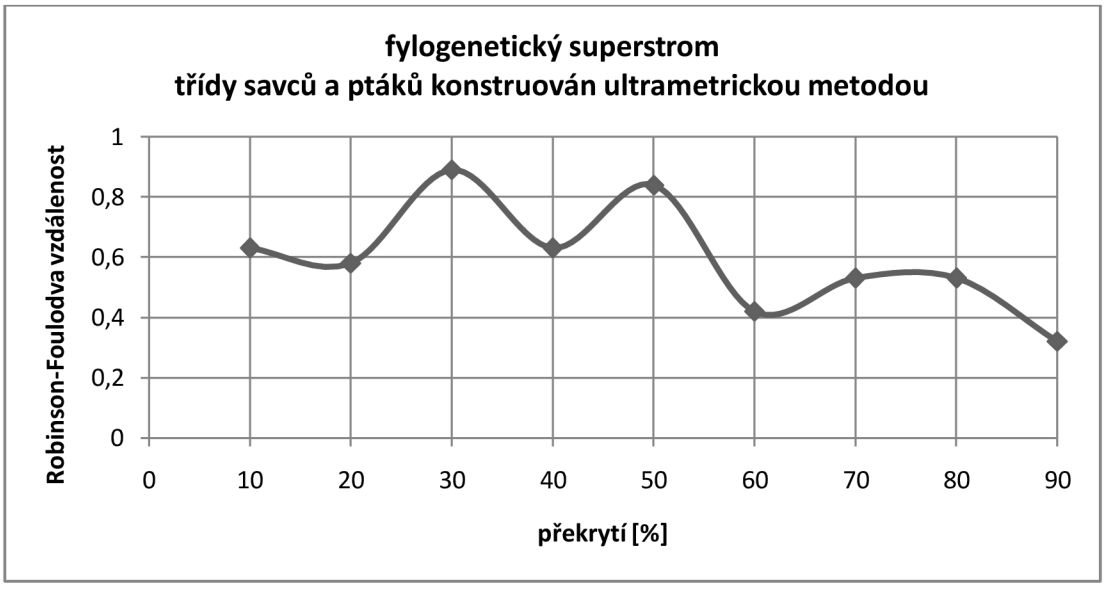
Obrázek 26: Ultrametrická metoda: fylogenetický superstrom tříd savců a ptáku s 10 % překrytím

Pro porovnání je zde přidán i superstrom s desetiprocentním překrytím, který korelační koeficient hodnotí jako středně těsný ke stromu referenčnímu. Shluk savců (*Mammalia*) se rozpadl na dva a neúplný shluk primátů je přiřazen blíže k shluku ptáků (*Aves*). Jeden z primátů je nesprávně přiřazen k řádu dlouhokřídlých. V druhém shluku savců již není zachována soudržnost řádů. Superstrom správně zachoval shluk ptáků se správným rozdělením na řády.

Z tabulky 17 vytvořené pro vyhodnocení testů ultrametrické metody na výše uvedených datech je patrné, že od šedesáti procent překrytí jsou vztahy mezi stromy klasifikovány jako velmi těsné a můžeme tedy konstatovat, že dané stromy se velmi podobají stromu referenčnímu. Graf 1 ukazuje závislost Robinson-Fouldovi vzdálenosti na procentuálním překrytí testovaného stromu se stromem referenčním, kde můžeme vidět kmitající hodnoty, což značí v tomto případě nezávislost obou sledovaných hodnot. Poslední tři hodnoty by se však dali klasifikovat jako závislé a naznačují větší shodu vytvořených shluků obou stromů s větším procentuálním překrytím.

Tabulka 17: Vyhodnocení ultrametrické metody superstromu savců a ptáků

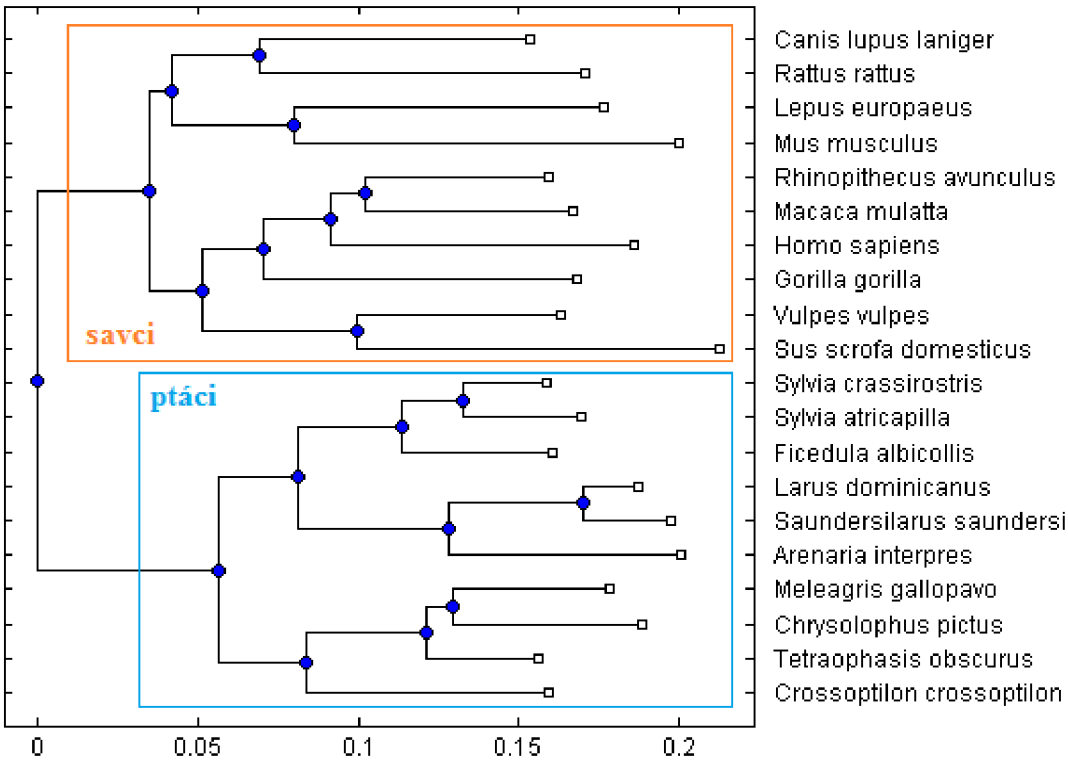
překrytí [%]	10	20	30	40	50	60	70	80	90
RF	0,63	0,58	0,89	0,63	0,84	0,42	0,53	0,53	0,32
Korel	0,58	0,54	0,31	0,54	0,31	0,74	0,79	0,73	0,92



Graf 1: Robinson-Fouldova vzdálenost pro konstrukci fylogenetického superstromu třídy savců a ptáků ultrametrickou metodou

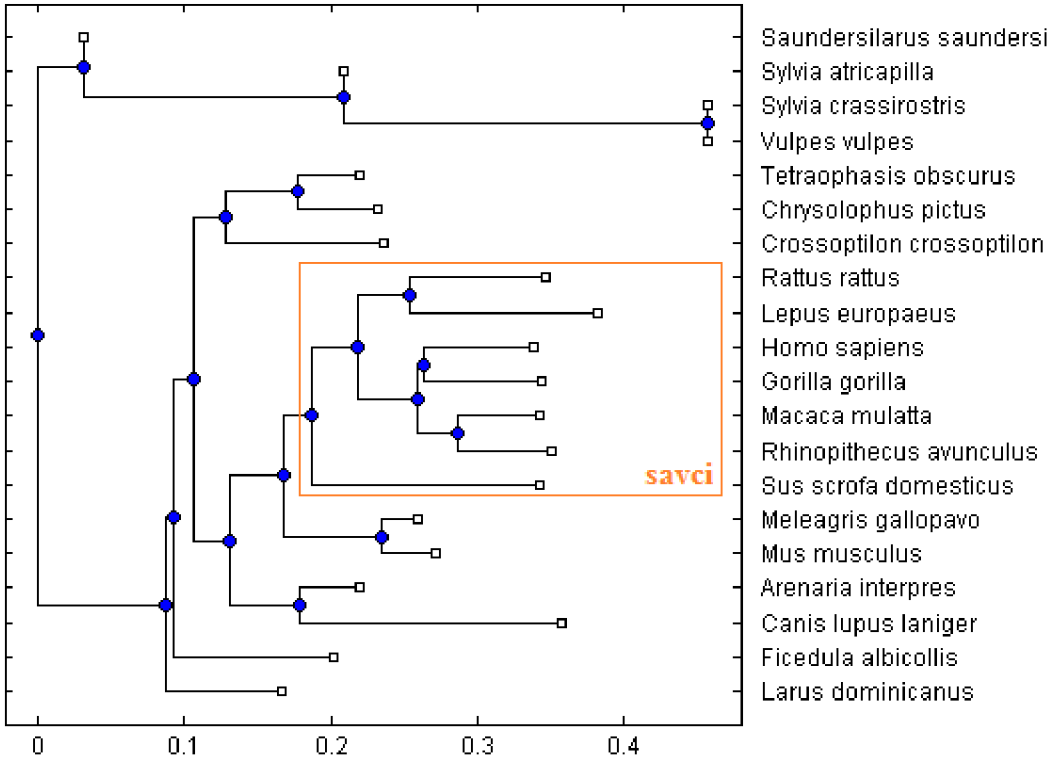
5.2 Aditivní metoda

Aditivní metoda byla podrobena stejnému testování jako metoda ultrametrická a byla tak získána data z pěti kombinací databází. Výsledné superstromy s různým procentem překrytí byly srovnávány se stejnými referenčními stromy, jako předešlá metoda. Je tedy možné obě metody na základě získaných dat porovnat výsledky obou metod.



Obrázek 27: Aditivní metoda: fylogenetický superstrom tříd savců a ptáku s 90 % překrytím

Referenční strom zůstal stejný jako v případě ultrametrické metody, jelikož jsou testovány stejné databáze. Superstrom s devadesátiprocentním překrytím konstruovaný aditivní metodou vychází téměř stejně jako metodou ultrametrickou, což je dáno i tím, že jsou neznámé pouze dvě hodnoty distance ve stromě. Strom je rozdělen do dvou shluků. Shluk ptáků (*Aves*) obsahuje správně rozřazené řády hrabavých, dlouhokřídlých a pěvců. Druhý shluk představující třídu savců (*Mammalia*) správně určil řád primátů, ovšem nesprávně rozdělil řád šelem mezi dva shluky.



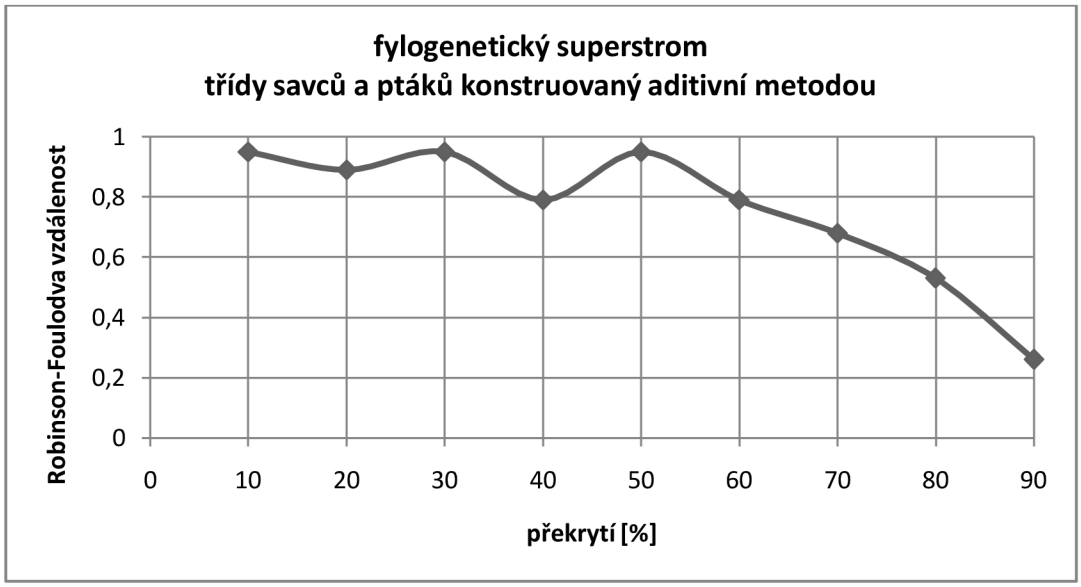
Obrázek 28: Aditivní metoda: fylogenetický superstrom tříd savců a ptáku s 60 % překrytím

Superstrom s šedesátiprocentním překrytím je viditelně špatně sestrojen a obsahuje mnoho artefaktů. Hodnota korelačního koeficientu vůči referenčnímu stromu je 0.21, což odpovídá zanedbatelnému vztahu. Není zde dodrženo rozdělení shluků podle tříd a jednotlivé řády jsou taky narušeny. I přesto zůstaly některé řády správně pospolu, jako například řád primátů.

Jak můžeme vidět v tabulce 18, tak superstromy vytvořené aditivní metodou začínají být významné až od osmdesátiprocentního překrytí. To dokazuje i tabulka 20, která obsahuje další testované databáze. Z grafu 2, který ukazuje závislost Robinson-Fouldovy vzdálenosti na procentech překrytí lze vypozorovat, že čím větší je překrytí zdrojových stromů, tím více shluků se ve stromech shoduje.

Tabulka 18: Vyhodnocení aditivní metody superstromu savců a ptáků

překrytí [%]	10	20	30	40	50	60	70	80	90
RF	0,95	0,89	0,95	0,79	0,95	0,79	0,68	0,53	0,26
Korel	-0,07	-0,01	0,03	0,29	0,12	0,21	0,25	0,76	0,92



Graf 2: Robinson-Fouldovia vzdálenost pro konstrukci fylogenetického superstromu třídy savců a ptáků aditivní metodou

5.3 Vyhodnocení

K vyhodnocení poslouží data získaná testováním pěti náhodně vytvořených dvojic s překrytím od deseti do devadesáti procent. Každý superstrom vytvořený s určitým překrytím zdrojových stromů byl porovnán se stromem referenčním pro danou dvojici taxonů. Ke srovnání těchto stromů posloužilo stanovení zkreslení vstupních dat na základě korelace distancí a rozdílu obsahu shluků. Pro obě metody byly vytvořeny tabulky obsahující informace a na jejichž základu je vyhodnocena úspěšnost obou metod.

Tabulka 19: Výstupní tabulka testování ultrametrické metody

ultrametrická metoda										
	hmyz x plazi		hmyz x savci		plazi x ptáci		savci x ptáci		savci x ryby	
překrytí [%]	RF	Kor.koef	RF	Kor.koef	RF	Kor.koef	RF	Kor.koef	RF	Kor.koef
10	0,58	0,45	0,89	0,33	0,11	0,87	0,63	0,58	0,63	0,46
20	0,63	0,39	0,84	0,36	0,42	0,79	0,58	0,54	0,68	0,37
30	0,79	0,21	0,95	0,32	0,53	0,85	0,89	0,31	0,84	0,26
40	0,84	0,46	0,89	0,41	0,63	0,71	0,63	0,54	0,53	0,65
50	0,84	0,51	0,89	0,42	0,63	0,73	0,84	0,31	0,53	0,56
60	0,63	0,8	0,84	0,74	0,53	0,78	0,42	0,74	0,26	0,84
70	0,32	0,78	0,84	0,69	0,42	0,84	0,53	0,79	0,68	0,47
80	0,53	0,85	0,84	0,81	0,53	0,86	0,53	0,73	0,68	0,65
90	0,52	0,87	0,68	0,86	0,11	0,95	0,32	0,92	0,42	0,87

RF…Robinson-Fouldova vzdálenost, Kor.koef… Pearsnův korelační koeficient

Výstupem testování ultrametrické metody na výše uvedených datech je tabulka 19. Podíváme-li se na Robinson-Fouldovu vzdálenost, můžeme vidět, že ve většině případů odpovídá hodnotě korelačního koeficientu, tedy čím vyšší je korelační koeficient, tím nižší je RF. U hodnot korelačního koeficientu můžeme pozorovat růst s vyšším překrytím u většiny dvojic, jsou ale i případy, kdy ultrametrická metoda vytvoří lepší superstrom z dat s menším překrytím a naopak. Hranicí, kdy ultrametrická metoda dokáže vytvořit superstromy s velmi těsným vztahem ke stromu referenčnímu, je šedesátiprocentní překrytí zdrojových stromů.

Tabulka 20: Výstupní tabulka testování aditivní metody

aditivní metoda										
	hmyz x plazi		hmyz x savci		plazi x ptáci		savci x ptáci		savci x ryby	
překrytí [%]	RF	Kor.koef	RF	Kor.koef	RF	Kor.koef	RF	Kor.koef	RF	Kor.koef
10	0,95	-0,06	0,95	-0,02	0,95	0,07	0,95	-0,07	0,95	0
20	0,95	-0,04	0,89	0,11	0,95	0,02	0,89	-0,01	0,89	0,03
30	0,84	-0,02	0,84	0,26	0,95	0,01	0,95	0,03	0,79	0,04
40	0,95	0,18	0,84	0,45	0,84	0,24	0,79	0,29	0,53	0,78
50	0,84	0,19	0,79	0,53	0,79	0,21	0,95	0,12	0,53	0,72
60	0,68	0,53	0,95	0,52	0,63	0,36	0,79	0,21	0,32	0,86
70	0,32	0,82	0,79	0,57	0,63	0,31	0,68	0,25	0,68	0,65
80	0,42	0,88	0,84	0,79	0,11	0,96	0,53	0,76	0,53	0,83
90	0,52	0,88	0,68	0,86	0,11	0,95	0,26	0,92	0,47	0,88

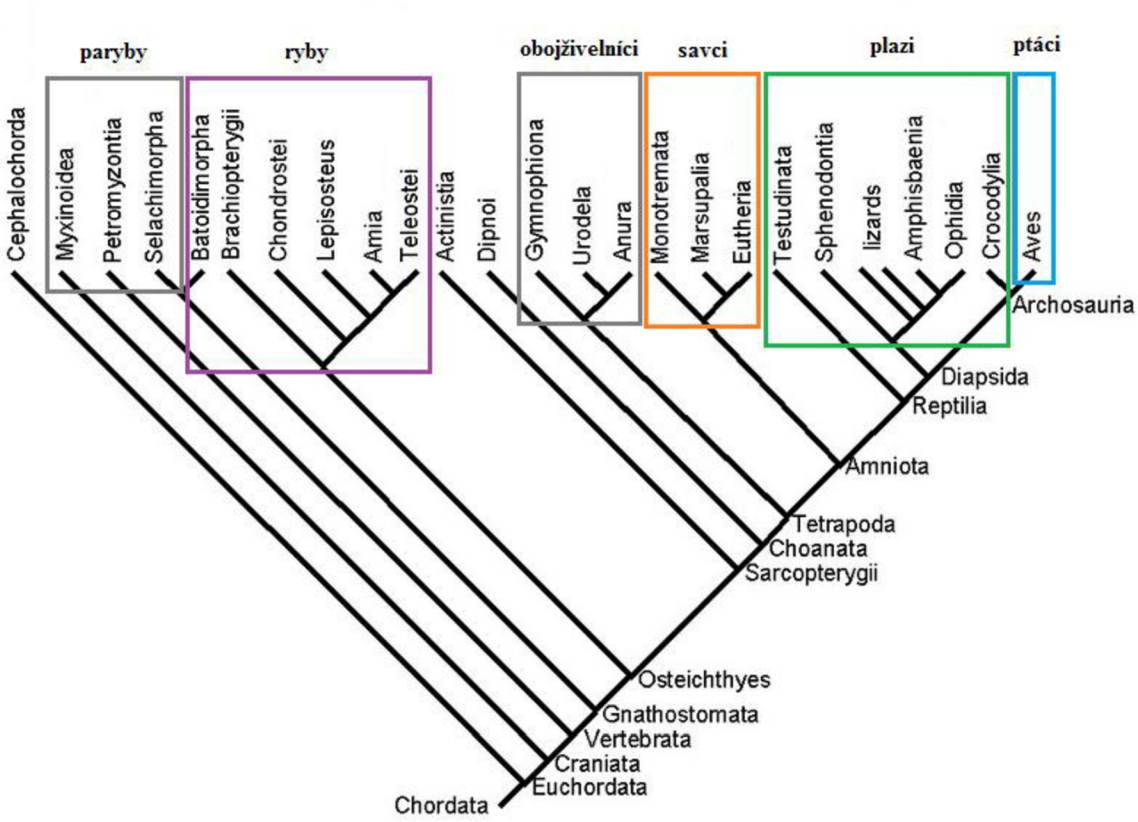
RF…Robinson-Fouldova vzdálenost, Kor.koef… Pearsnův korelační koeficient

Stejně jako předchozí metody byla testována i metoda aditivní a hodnoty testování jsou zaneseny do tabulky 20. Kvalita superstromů konstruovaných aditivní metodou se zdá být mnohem více závislá na procentuálním překrytí zdrojových stromů, nežli je tomu u metody ultrametrické. Lze si všimnout téměř u všech dvojic (vyjma dvojice savci x ptáci) rostoucí hodnoty korelačního koeficientu s rostoucím překrytím sloučených zdrojových stromů. Hodnoty RF mají tendenci klesat přibývajícím překrytím. Použití RF jako parametr pro srovnání dvou stromů se zdá být nespolehlivý a spíše jako doplňující informace ke korelačnímu koeficientu. Z výsledků srovnání superstromů konstruovaných aditivní metodou vyplívá, že hranicí, kdy tato metoda dokáže vytvořit superstrom s velmi těsným vztahem, je osmdesáti procentní překrytí zdrojových stromů.

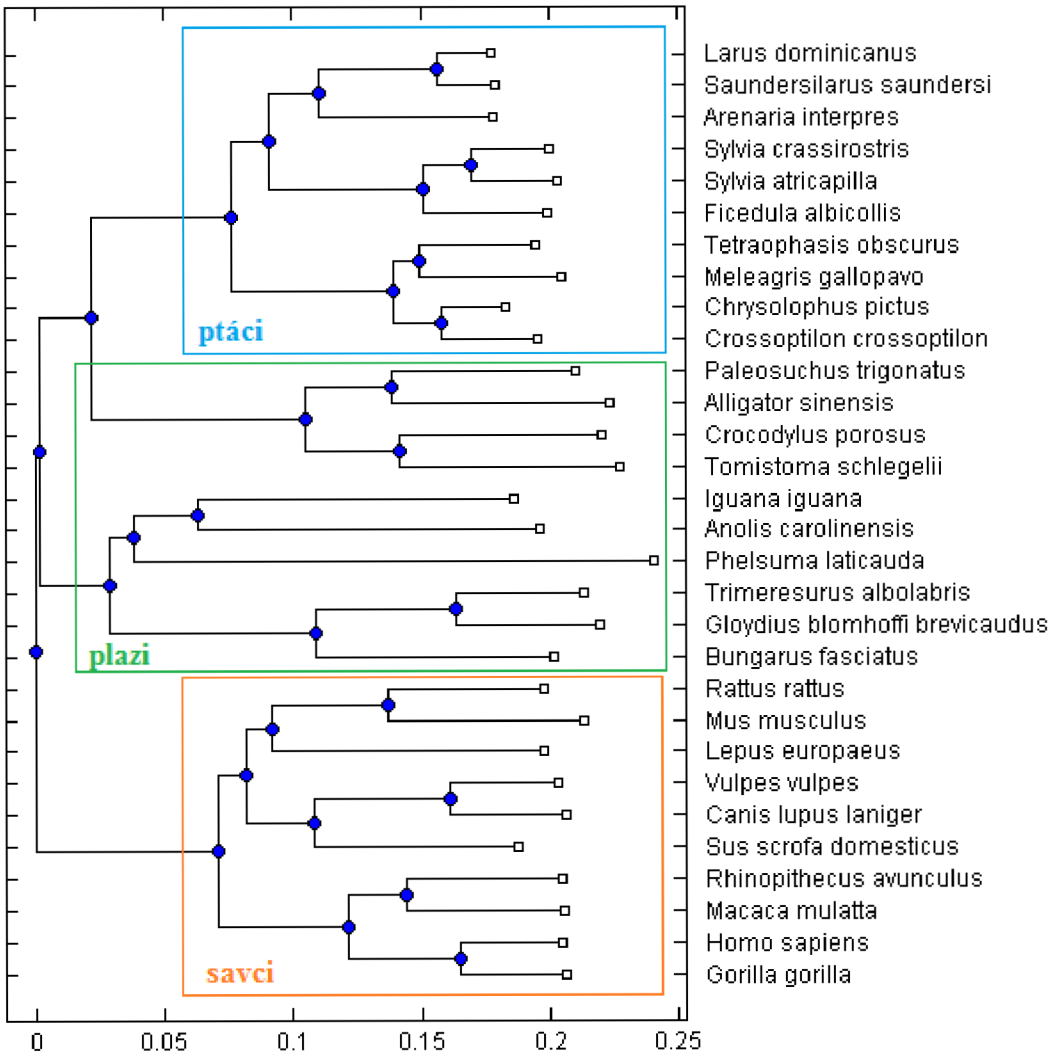
6 DISKUZE

Ultrametrická metoda stejně jako metoda aditivní dokážou vytvořit superstrom i z minimálního procentuálního překrytí ovšem za cenu toho, že tyto stromy neodpovídají ani zdaleka stromu referenčnímu a jsou v rozporu s tradičními hypotézami fylogenetických stromů. Z provedeného testování plyne, že obě vytvořené metody dopočtu neznámých hodnot dávají významné výsledky až od určitého procentuálního překrytí, jehož hranice se pro každou metodu liší.

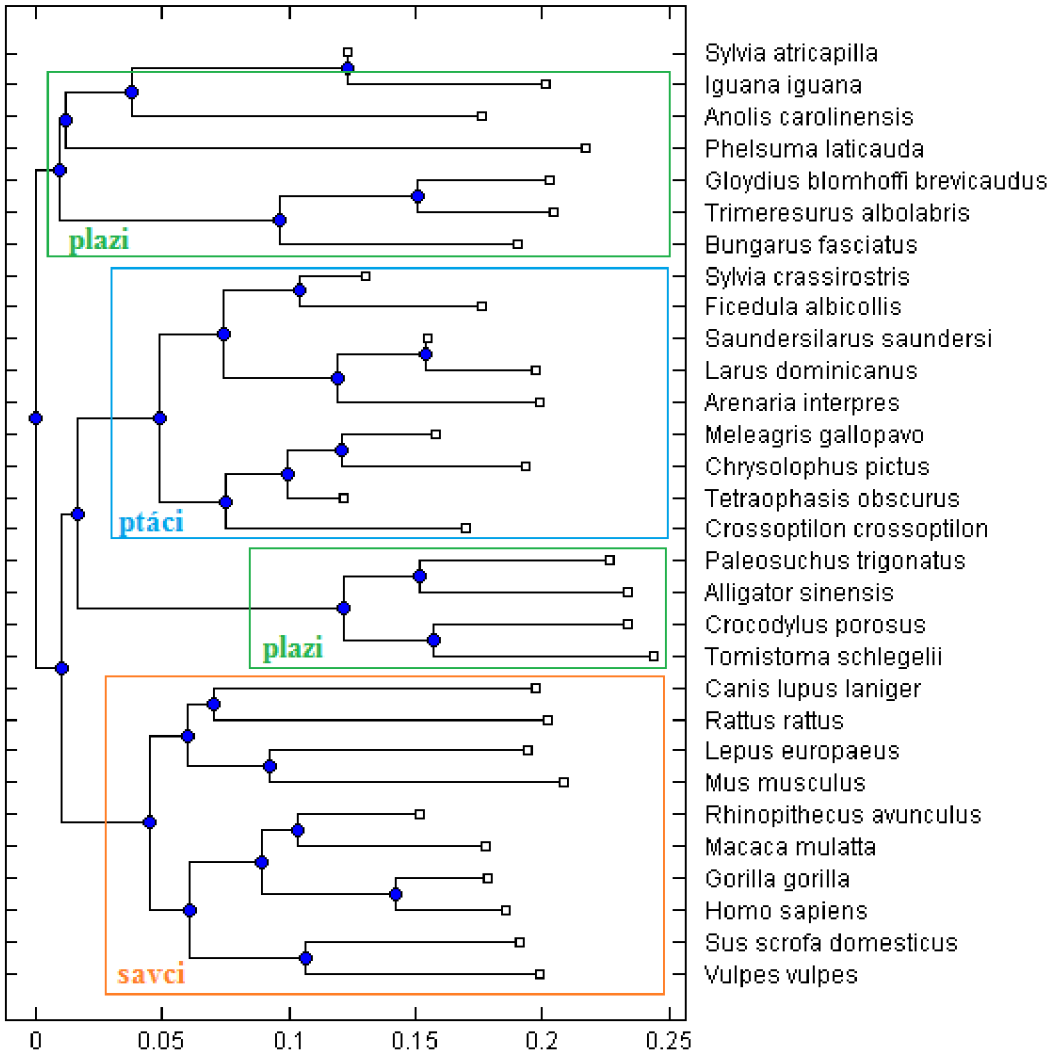
Na obrázcích 30 a 31 je vidět konstrukce superstromu ze tří tříd s padesátiprocentním překrytím. V referenčním stromě si můžeme všimnout tří shluků (třídy *Reptilia*, *Mammalia, Aves*) přičemž třída plazů, konkrétně řád krokodýlů, je oddělen a přiřazen blíže třídě ptáků, stejně jako v kapitole 3 ve výsledném fylogenetickém stromě. Správnost rozdělení třídy plazů je podložena i přiloženým obrázkem 29, na kterém je vyobrazeno základní schéma evoluce obratlovců (*Vertebrata*).



Obrázek 29: Fylogenetický strom obratlovců [32]

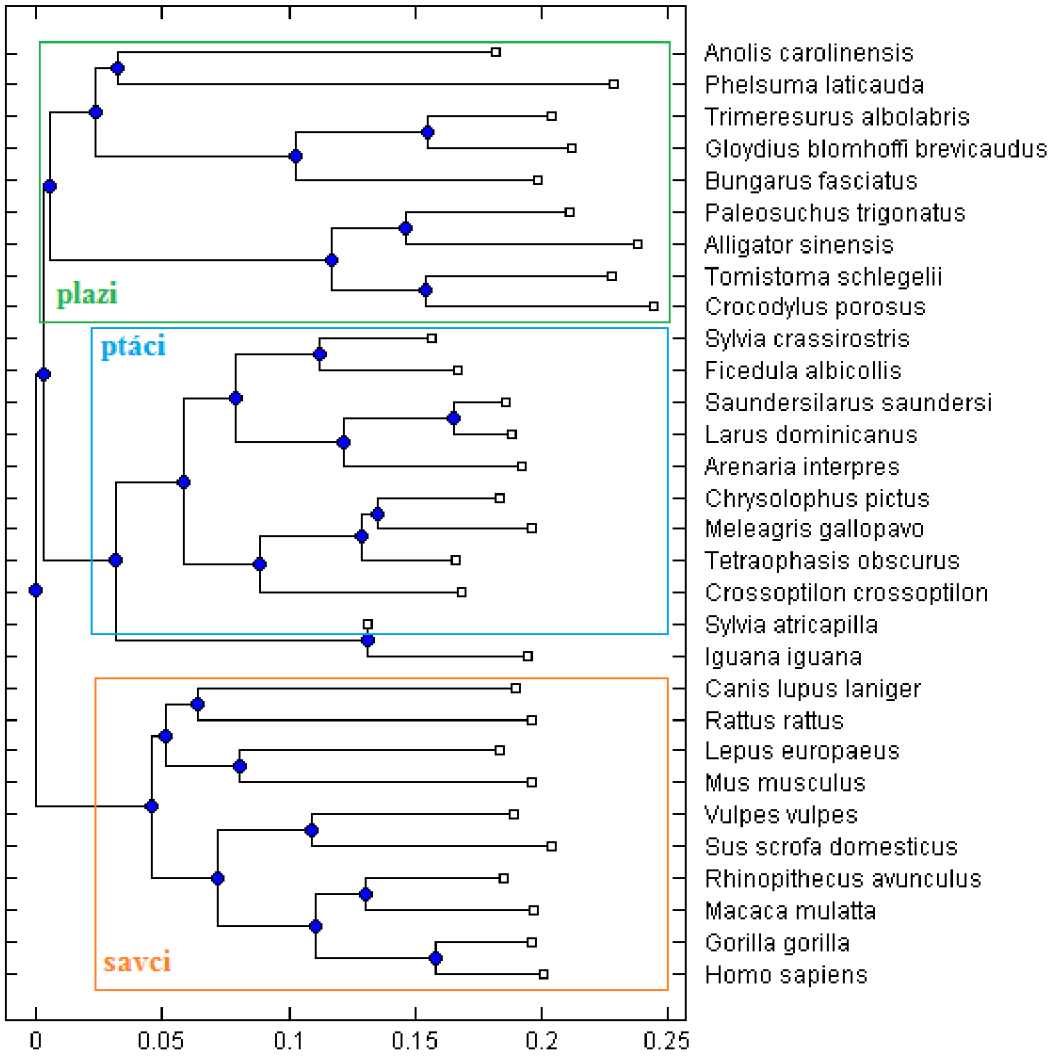


Obrázek 30: Referenční fylogenetický superstrom tříd plazů, ptáků a savců



Obrázek 31: Referenční fylogenetický superstrom tříd plazů, ptáků a savců s padesátiprocentním překrytím konstruovaný ultrametrickou metodou

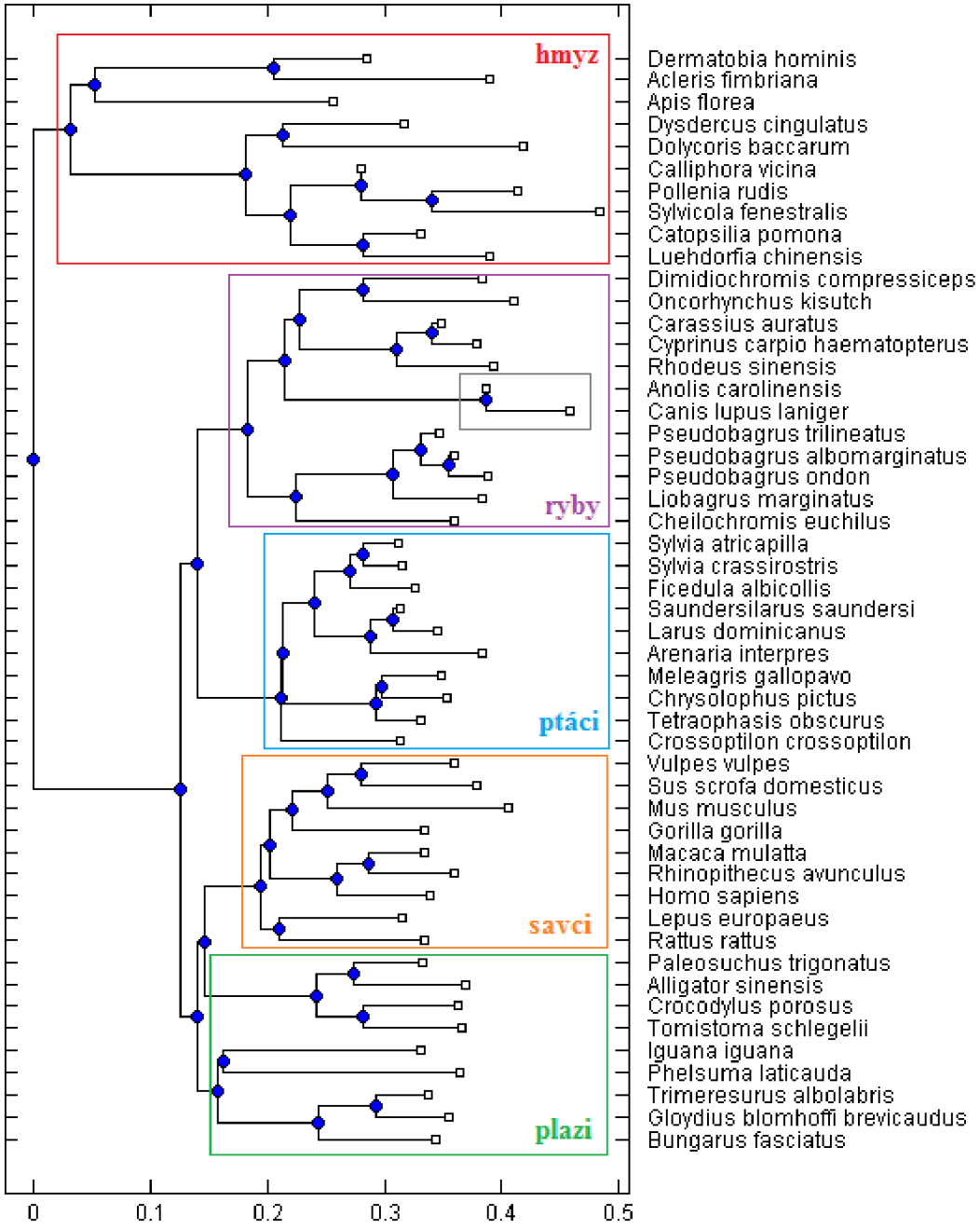
Superstrom konstruovaný ultrametrickou metodou obsahuje čtyři oddělené shluky. Správně je tedy rozdělena třída plazů (*Reptilia*), od které je odtržen řád krokodýlů, přiřazený ke třídě ptáků (*Aves*). Nesprávně, podle výše uvedených referenčních dat, je blíže k tomuto shluku přiřazena třída savců (*Mammalia*), čímž je druhá část třídy plazů klasifikována jako nejvzdálenější, přestože by tomu tak být nemělo. Ve stromě se objevují chybně zařazení jedinci a jednotlivé řády nejsou vždy pospolu, jako například *Sylvia atricapilla* tvořící pár s *Iguana iguana* nebo rozpad řádu hlodavců (*Rattus rattus, Mus musculus*) a jejich spárování s jedinci jiných řádů. Na základě těchto znalostí lze konstatovat, že fylogenetický superstrom je v rozporu s tradičními hypotézami fylogenetických vztahů a padesátiprocentní překrytí se zdá být nedostatečné.



Obrázek 32: Referenční fylogenetický superstrom tříd plazů, ptáků a savců konstruovaný aditivní metodou

Superstrom konstruovaný aditivní metodou obsahuje čtyři oddělené shluky. Třída plazů (*Reptilia*), je rozdělena na dva shluky, kdy shluk krokodýlů je evolučně blíže třídě ptáků (*Aves*). Nejvzdálenějším shlukem je třída savců (*Mammalia*). Tento superstrom obsahuje stejné chyby spojování jedinců různých tříd a nesprávné rozdělení některých řádů jako ultrametrická metoda. I tento fylogenetický superstrom je v rozporu s tradičními hypotézami fylogenetických vztahů a padesátiprocentní překrytí není dostatečné pro správnou klasifikaci.

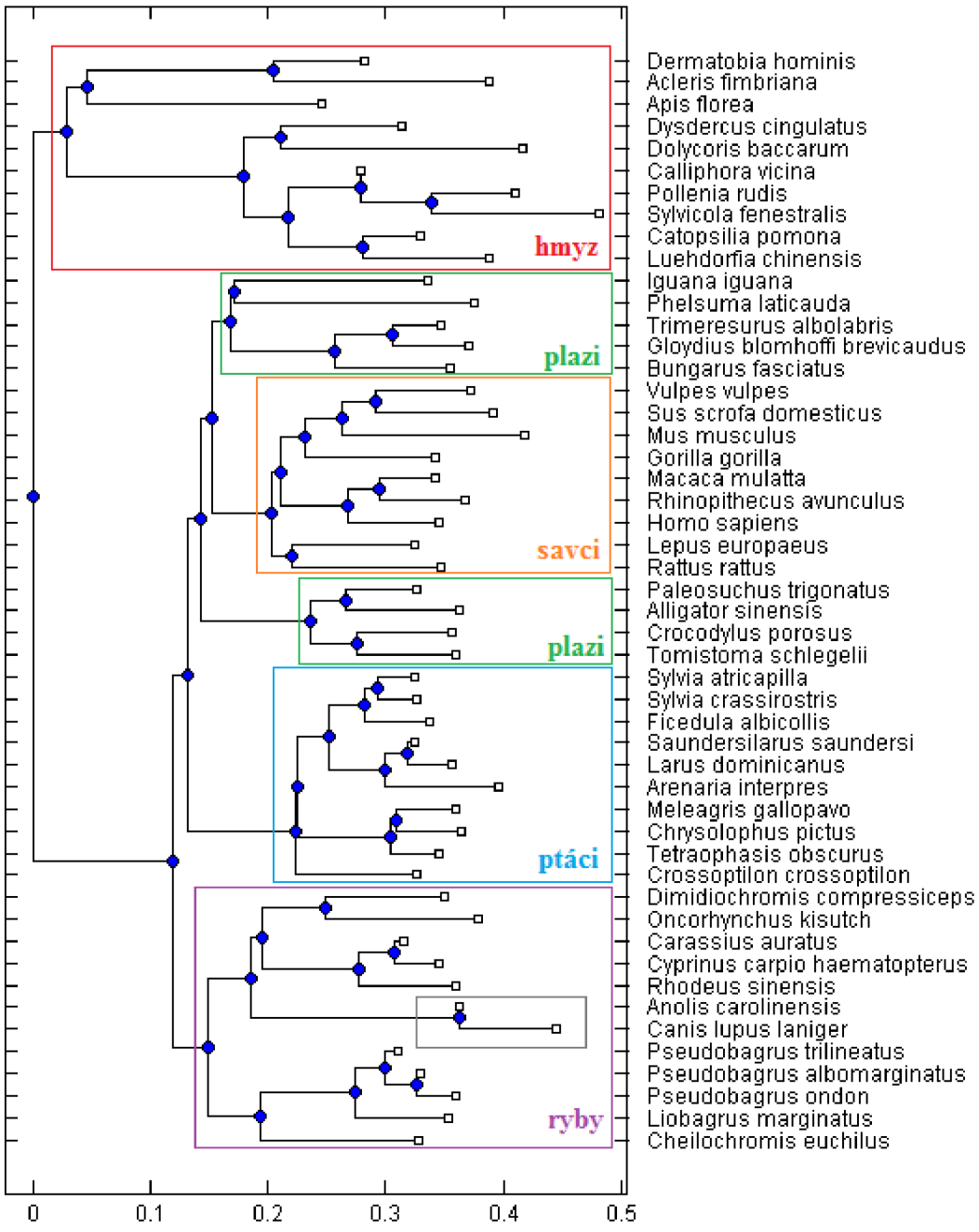
Na závěr testování obou vytvořených metod jsem vytvořil zdrojové stromy všech pěti tříd (padesát taxonů) s devadesátiprocentním překrytím, abych otestoval, jak si s tím jednotlivé metody poradí a zároveň demonstroval, že lze slučovat i větší množství zdrojových stromů. Získané superstromy jsou srovnány k fylogenetickému stromu na obrázku 15, který je brán, jako referenční.



Obrázek 33: Fylogenetický superstrom z padesáti sekvencí konstruovaný ultrametrickou metodou

Ultrametrická metoda správně oddělila shluky všech pěti tříd. Jasně jsou zde rozlišeny dva velké shluky, horní shluk reprezentující kmen členovců (*Arthropoda*) a druhý shluk

reprezentující kmen strunatců (*Chordata*). Vyskytují se zde ale nedokonalosti, jako je správné rozdělení shluku plazů (*Reptilia*), kdy krokodýli mají vlastní uzel, ale jsou nesprávně zařazeni blíže savcům (*Mammalia*) místo k ptákům (*Aves*). Umístění třídy ptáků blíže třídě ryb a celkovým špatným uspořádáním evoluce v tomto velkém shluku. Drobnou chybou pak je i spojení *Canis lupus* (savec) s *Anolis carolinensis* (plaz) a přiřazení této dvojice do třídy ryb. Z výše uvedených dat je tedy možné vyčíst, že ultrametrická metoda dokáže správně analyzovat na úrovni tříd, ale výsledný superstrom má nedostatky a je tak v mírném rozporu s tradičními hypotézami fylogenetických vztahů.



Obrázek 34: Fylogenetický superstrom z padesáti sekvencí konstruovaný aditivní metodou

Fylogenetický superstrom na obrázku 34, vytvořený aditivní metodou, obsahuje taktéž dva velké shluky kmene členovců (*Arthropoda*) a kmene strunatců (*Chordata*). Druhý zmiňovaný shluk je tvořen savci (*Mammalia*), plazi (*Reptilia*), ptáky (*Aves*) a rybami (*Osteichthyles*). Za nedostatky tohoto superstromu jsou považovány drobné chyby na úrovni řádů, které jsou zapříčiněny i špatným spojením dvojice *Canis lupus* (savec) s *Anolis carolinensis* (plaz), a přiřazení této dvojice do třídy ryb, stejně jako u předešlé metody, a prohození třídy savců s třídou ptáků, čímž by byla dodržena správná evoluce tříd. I přesto je možné konstatovat, že aditivní metoda si se stejnými daty poradila lépe a výsledný superstrom není ve výrazném rozporu s tradičními hypotézami fylogenetických vztahů.

Z vytvořených superstromů a výsledků testování obou metod odvozuji, že obě metody mají tendenci dávat lepší výsledky ze zdrojových stromů s vyšším procentem překrytí, přesto se ale stává, že je vytvořen kvalitnější superstrom ze zdrojových stromů s nižším stupněm překrytí nebo naopak z vyššího stupně překrytí je vytvořen méně kvalitní superstrom. Tato skutečnost je dána i konstrukční metodou, kterou jsou superstromy vytvořeny. Jedná se o metodu spojování nejbližších sousedů (Neighbor joining) a jak už sám název této metody napovídá, metoda spojuje dva sousedy ve finální distanční matici s nejmenší vzdáleností. Výsledný strom je tak závislý na pořadí, v jakém jsou ve finální matici uspořádány taxony a vzhledem k nemožnosti „správného seřazení“ (neexistuje pravidlo, kde v matici má ležet cizí taxon) je výsledkem pokaždé jiný fylogenetický superstrom. Lze tedy konstatovat, že metoda NJ je závislá na prostorovém uspořádání taxonů v matici s nemožností dosáhnout vždy správného uspořádání. Může se tak stát, že superstrom, který vznikl ze zdrojových stromů s nízkým procentem překrytí, bude více odpovídat referenčnímu stromu než superstrom vzniklý ze stromů s vyšším procentuálním překrytím.

Jednou z příčin některých chybných výsledků fylogenetické analýzy mohly být i velmi vzdálené sekvence, které ve výsledném fylogenetickém superstromu vytváří takzvané artefakty dlouhých větví. Tyto artefakty v topologii stromu jsou způsobeny výrazně odlišnou substituční rychlostí a praktický může dojít ke třem situacím:

- Přitahování dlouhých větví (dvě velmi rozdílné sekvence jsou přitahovány k sobě a směrem ke kořeni.
- Odpuzování dlouhých větví – dva příbuzné geny, výrazně pozměněné, jsou odpuzovány od sebe.
- Vyrušování dlouhých větví – jedna dlouhá větev ovlivňuje topologii celého stromu.

Mezi kladné stránky fylogenetických superstromů bezesporu patří možnost využití dat z více zdrojů, čímž je umožněno získat lepší výsledky prováděných fylogenetických analýz a řešit situace, pro které do této doby nebyly jiné možnosti. Další výhodou a zároveň i

nevýhodou je možnost vytvoření fylogenetického superstromu ze stromů, mezi nimiž nejsou všechny distance známy. Metoda průměrného konsensu spolu s metodami dopočtu neznámých distancí si s takovými stromy dokáže poradit a neznámé distance dopočítává ze známých distančních vzdáleností k taxonům vyskytujícím se ve stromě. Zde se dostáváme k nevýhodě a to proto, že tento dopočet nemusí vždy odpovídat skutečné vzdálenosti mezi danými taxony a výsledný superstrom tak může být výrazně zkreslený.

Realizované metody dopočtu neznámých distancí dokážou vytvořit superstrom z libovolného množství zdrojových stromů. Výsledný superstrom je však závislý na procentu překrytí dat ve zdrojových stromech. Jak plyne z testování, tak ultrametrická metoda dává klasifikovatelné výsledky už při šedesátiprocentním překrytí, zatímco aditivní metoda vyžaduje až osmdesátiprocentní překrytí. Obě realizované metody se tedy hodí spíše pro dopočet méně neznámých vzdáleností v distanční matici a nejsou tak vhodné pro spojování stromů s velkým množstvím neznámých taxonů. Zmíněné hranice pro vytvořené metody jsem určil na základě testování, které nebylo tak rozsáhlé a je možné, že bude dosaženo jiných výsledků na mnohonásobně větších datech. Čímž se dostávám k další nevýhodě, kdy s rostoucím objemem vstupních dat se zvyšuje časová náročnost prováděných algoritmů a je potřeba mít kvalitní počítačovou techniku.

7 ZÁVĚR

Tato diplomová práce se zabývá metodami rekonstrukce fylogenetických superstromů. Úvod práce obsahuje teoretické seznámení s molekulární fylogenetikou, která je důležitá pro pochopení dalšího textu. Tato kapitola zahrnuje zpracování molekulárních dat, popis fylogenetických stromů a metody jejich rekonstrukce.

Další kapitola je věnována fylogenetickým superstromům. Jsou zde probrány důvody, pro které jsou superstromy používány a popsány metody pro jejich sestrojení. Práce popisuje tři metody pro konstrukci fylogenetických superstromů založených na principu kompletace částečných vzdálenostních matic.

Třetí část je věnována vytvoření zdrojových stromů pro následnou srovnávací studii rekonstrukčních algoritmů superstromů. Data použita v této práci jsou získána z veřejně dostupné databáze NCBI. Pro porovnání získaných dat jsem zvolil mitochondriální gen označován jako 16S rRNA. Tato sekvence DNA kóduje podjednotku RNA bakteriálního ribozomu, který nese důležité informace pro zkoumání evoluční příbuznosti. Následně jsem sestrojil zdrojové stromy z vytvořené databáze sekvencí pro každou z pěti tříd živočichů: savce (*Mammalia*), plazy (*Reptilia*), ptáky (*Aves*), ryby (*Osteichthyes*) a hmyz (*Insecta*).

Ve čtvrté kapitole je popsán program vytvořený v programovém prostředí Matlab, který slouží k realizaci fylogenetického superstromu. Je zde popsáno, v jakém formátu mají být vstupní data, jejich zpracování a popis výpočtu distanční matice. Následně je zde graficky znázorněn a popsán princip, jak ultrametrické metody, tak i metody aditivní, které slouží k dopočtu neznámých hodnot v distanční matici vzdáleností.

Pátá kapitola slouží k otestování vytvořených metod pro sestrojení superstromů a srovnání jejich výstupů s referenčními stromy. Je zde vytvořeno několik superstromů na kterých jsou předvedeny dobré vlastnosti obou metod, ale i jejich nedostatky. Výsledky testů jsou uvedeny v tabulkách, ze kterých jsem určil, že hranicí, kdy ultrametrická metoda dokáže vytvořit superstromy s velmi těsným vztahem ke stromu referenčnímu, je šedesátiprocentní překrytí zdrojových stromů. Zatímco u aditivní metody je hranicí osmdesátiprocentní překrytí zdrojových stromů. Při vysokém stupni překrytí zdrojových stromů tak dává lepší výsledky metoda aditivní, zatímco při nízkém stupni překrytí je lepší variantou metoda ultrametrická.

Ve zbylé části této diplomové práce jsou diskutovány získané výsledky. Vytvořené fylogenetické superstromy jsou konfrontovány s veřejně dostupným fylogenetickým stromem popisující evoluci strunatců a taky s klasickým fylogenetickým stromem z kapitoly tři. Jsou zde diskutovány i výhody a nevýhody fylogenetických superstromů.

Seznam zkratek

DNA	deoxyribonukleová kyselina
IUPAC	International Union of Pure and Apllied Chemistry
J-C model	Jukes-Cantorův model (metoda pro výpočet evoluční vzdálenosti)
ME	Minimum evolution (metoda pro rekonstrukci fylogenetického stromu)
ML	Maximum Likelihood (metoda pro rekonstrukci fylogenetického stromu)
MPR	Matrix Representation with Parsimony (metoda pro konstrukci fylogenetického superstromu)
MSSA	The most Simile Supertree Algorithm (metoda pro konstrukci fylogenetického superstromu)
NJ	Neighbor-joining method
OTU	Operational Taxonomic Unit
RNA	ribonukleová kyselina
UPGMA	Unweighted Pair Group Method with Arithmetic Mean (metoda pro rekonstrukci fylogenetického stromu)

Seznam příloh

Literatura

[1] BININDA-EMONDS, O. *Phylogenetic supertrees: combining information to reveal the tree of life*. Boston: Kluwer Academic Publishers, 2004, xiv, 550 p. ISBN 14-020-2330-8.

[2] BININDA-EDMONDS, O. R. P. The Evolution of supertrees: Trends in ecology and evolution (Personal Edition). 2004, 19(6):315-322.

[3] CAVALLI-SFROZA, L. L. a A. W. F. EDWARDS. Phylogenetics analysis: Models and estimation procedures. *Am.J.Hum.Genet*. 1967, 19:233-257.

[4] CREEVEY, C. J. a J. O. MCINERNEY. Trees from trees: construction of phylogenetic supertrees using clann. *Methods In Molecular Biology Clifton Nj*. 2009, vol. 537, pp. 139-161.

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

[6] DOWELL, Karen. *Molecular Phylogenetics: An introduction to computional methods and tools for analyzing evolutionary relationships*. Math 500, 2008, 18 s.

[7] ECK, R. V. a M.O. DAYHOFF. *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation: Silver Springs, Maryland, 1966.

[8] Encyklopaedia Britannica. [online]. 2011 [cit. 2013-11-30]. Dostupné z: http://www.britannica.com/EBchecked/topic/502158/ribosomal-RNA-rRNA#ref1116894

[9] FELSTEIN, J. Evolutionary trees from DNA sequences: A maximum likelihood approach.*Molecular Evolution*. 1981, 17:368-376.

[10] FITCH, W. M. Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.* 1971, 20:406-416.

[11] FLÉGR, Jaroslav. *Evoluční Biologie*. 2. vyd. Praha: ACADEMIA, 2009. 572 s. ISBN 978-80-200-1767-3.

[12] FLÉGR, Jaroslav. *Molekulární Fylogenetika*. Elektronické skripta. Praha: PřF UK v Praze, 2007. 43s.

[13] FLEGR, Jaroslav. *Úvod do evoluční biologie*. Vyd. 1. Praha: Academia, 2007, 544 s. Galileo, sv. 13. ISBN 978-802-0015-396.

[14] GASCUEL, O. *BIONJ: An Improved Version of the NJ Algorithm Based on a Simple Model of Sequence Data*. Molecular Biology evolution, 1977, 14(7):685-695.

[15] HARTIGAN, J. A. Minimum Mutation Fits to a Given Tree. *Biometrics*. 1973, vol. 29, issue 1.

[16] HARTWELL, LH, L. HOOD, ML. GOLDBERG, AE. REYNOLDS, LM. SILVER a RC. VERES. Genetics: From Genese to Genomes. 2008, 3rd Ed. McGraw-Hill: New York.

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

[18] JUKES, TH. a CR. CANTOR. Evolution of protein molecules. *Mammalian Protein Metabolism*. New York: Academic Press, 1969, pp. 21-132.

[19] KIMURA, M. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*. 1980, 16:111-120.

[20] LIO, P. a N. GOLDMAN. Models of Molecular and Phylogeny. *Genome Research*. 1998, 8:1233-1244.

[21] MAKARENKOV, V. a F. J. LAPOINTE. Weighted least-squares approach for inferring phylogenies from imcomplete distance matrices. *Bioinformatic*. 2004, 20:2113-2121.

[22] National Center for Biotechnology Information (NCBI). *Gen bank* [online]. 2011 [cit. 2013-12-15]. Dostupné z: http://www.ncbi.nlm.nih.gov/

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

[24] NEI, Masatoshi a Sudhir KUMAR. *Molecular evolution and phylogenetics*. New York: Oxford University Press, 2000, xiv, 333 p. ISBN 01-951-3585-7.

[25] ROBINSON, D.F., L.R. FOULDS. *Mathematical biosciences: Comparison of phylogenetic trees*. New York, 1981, 53:131-147. ISBN 0025-5564/81/01013117S02.50.

[26] RZHETSKY, Andre a Masatoshi NEI. Theoretical Foundation of the Minimum-Evolution Method of Phylogenetic Inference. *Molecular Biology Evolution*. 1993, 10:1073-1095.

[27] SAITOU, Naruya a Masatoshi NEI. The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology Evolution*. 1987, 4:406-425.

[28] Sequence Manipulation Suite: IUPAC codes. [online]. 2012 [cit. 2013-12-02]. Dostupné z: http://www.bioinformatics.org/sms2/iupac.html

[29] STÁVKOVÁ, Jana a Jaroslav DUFEK. *Biometrika*. Vyd. 2., nezměn. V Brně: Mendelova univerzita, 2012, 178, [13] s. ISBN 978-80-7375-634-5.

[30] TAMURA, K. a M. NEI. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*. 1993, 10: 512-526.

[31] ZIMA, Jan, Miloš MACHOLÁN, Pavel MUNCLINGER a Jaroslav PIÁLEK. *Genetické metody v zoologii*. 1. vyd. Praha: Karolinum, 2004, 239 s. Učební texty Univerzity Karlovy v Praze. ISBN 80-246-0795-6.

[32] Zoology: Comparative Vertebrate Anatomy [online]. 2008 [cit. 2014-04-15]. Dostupné z: http://www.ou.edu/class/zoo2204/CVAhome.html