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. KRISTÝNA JIRÁSKOVÁ
AUTHOR

VEDOUCÍ PRÁCE — Ing. HELENA ŠKUTKOVÁ
SUPERVISOR

BRNO 2012

VYSOKÉ UČENÍ
TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

Ústav biomedicínského inženýrství

Diplomová práce

magisterský navazující studijní obor
Biomedicínské inženýrství a bioinformatika

Studentka: Bc. Kristýna Jiráskové ***ID:*** 106148
Ročník: 2 ***Akademický rok:*** 2011/2012

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: 6.2.2012 ***Termín odevzdání:*** 18.5.2012

Vedoucí práce: Ing. Helena Škutková
Konzultanti diplomové práce:

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

UPOZORNĚNÍ:

Abstrakt

Fylogenetika v posledních desetiletích zaznamenala velký rozvoj. Vývojem počítačů a přístrojů pro sekvenování biopolymerů bylo získáno ohromné množství fylogenetických dat z různých zdrojů a odlišných typů. Snahou vědců je zrekonstruovat z těchto dat kompletní strom života. Fylogenetické superstromy tuto možnost teoreticky představují, jelikož na rozdíl od fylogenetických stromů umožňují kombinaci všech doposud získaných informací. Tato práce představuje metodu konstrukce superstromů metodou průměrného konsensu.

Klíčová slova

Fylogram, fylogenetický strom, OTU, fylogenetický superstrom

Abstract

The phylogenetic reconstruction has noted great development in recent decades. The development of computers and device for sequencing biopolymers have been an enormous amount od phylogenetic data from different sources and different types. The scientists are trying to reconstruct a comlet tree of life from these data. The phylogenetic supertree are theoretically this option because a supertree alow a combination of all information gathered so far – in contras to the phylogenetic trees. This thesis present the method of reconstruction supertrees using average konsensus method.

Key words

Phylogenetics, phylogenetic tree, OTU, phylogenetic supertree

JIRÁSKOVÁ, K. *Metody rekonstrukce fylogenetických superstromů: diplomová práce.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2011. 71 s. Vedoucí práce Ing. Helena Škutková.

Prohlášení

Prohlašuji, že svou diplomovou práci na téma Metody rekonstrukce fylogenetických superstromů jsem vypracovala 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 autorka uvedeného semestrálního projektu dále prohlašuji, že v souvislosti s vytvořením tohoto projektu jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

V Brně dne …… .....

podpis autora (autorky)

Poděkování

Mé poděkování patří v první řadě vedoucí diplomové práce Ing. Heleně Škutkové za trpělivost, věcné připomínky k práci a praktické rady. Dále bych ráda poděkovala Karolovi Mikulášovi za pomoc při tvorbě praktické částí a Mirce Onderkové za trpělivost. V neposlední řadě patří velké díky rodině, která mi zajistila studium jak z pohledu finanční stránky, tak byla po celou dobu studia velkou psychickou oporou. Poděkování si zaslouží také mí nadřízení, kteří mi umožnili věnovat se studiu.

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Fylogenetika je obor systémové biologie, který hledá skutečnou příbuznost na základě představy, že všechny organismy mají svého univerzálního společného předka. Historii evolučních vztahů mezi organismy zobrazuje za pomoci fylogenetických stromů. Zpočátku byly tyto stromy sestavovány biology pouze subjektivně na základě jejich vlastních zkušeností. Zlomem ve fylogenetice bylo objevení struktury DNA a rozluštění molekulárních dat, které vedlo ke vzniku molekulární fylogenetiky.

Rozvoj technik sekvenování vyústil v založení nezávislých databází, do kterých jsou přijímána data ze všech projektů sekvenování od institucí i jednotlivců z celého světa. Tyto databáze obsahují informace z různých zdrojů, data se překrývají a jsou odlišných typů, což komplikuje sestrojení fylogenetických stromů. Problém fylogenetické analýzy z nekompletních nebo nejistých vstupních dat je nejpodstatnější otázka v systematické biologii. Řešení tohoto problému představují fylogenetické superstromy.

Metody superstromů kombinují více fylogenetických stromů k vytvoření celkově lepších superstromů, které obsahují kompletní sadu všech listů nacházejících se ve vstupních stromech. Pro provedení analýzy je nezbytné, aby zdroje fylogenetických stromů byly spojeny se soubory sdílenými taxony.

V dnešní době je známo nejméně šestnáct metod, které se konstrukcí fylogenetických superstromů zabývají. Každá z těchto metod přistupuje k problému kombinování informací z různých zdrojů odlišnými způsoby. V této diplomové práci je představena metoda průměrného konsensu, která počítá vzdálenosti mezi všemi zdrojovými stromy. Průměrná vzdálenost každého taxonu je použita v konečné matici vzdáleností, ze které je superstrom sestrojen. Tato metoda je univerzálně použitelná, kombinuje všechny zdrojové stromy a produkuje dobře řešené a zřejmě přesné superstromy, které nejsou ve výrazném rozporu s tradičními hypotézami fylogenetických vztahů.

1 Molekulární fylogenetika

Země je jediná známá planeta, na které existuje život. Počet druhů se odhaduje na šest až sto miliónů, pouze jeden a půl milionu druhů bylo již objeveno, popsáno a zařazeno. Rozmanitost a příbuznost druhů protíná všechny kontinenty. Dnes již není pochyb o tom, že život na Zemi se vyvinul postupnými přeměnami druhů, které v boji za životaschopnými mutacemi nabyli až dnešních forem.

Významnou roli ve fylogenetice zaujímá britský přírodovědec Charles Darwin, který s publikací *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (O vzniku druhů přírodním výběrem, neboli uchováním prospěšných plemen v boji o život - většinou zkracovaná na: O původu druhů, roku 1859 způsobil revoluci. Od něho pochází představa Stromu života, který v souhlase s geologickými epochami rašil stále dalšími větvemi nových druhů a vedl ke vzniku k Homo sapiens. [13]

Věda zabývající se rozpoznáváním vzájemných příbuzenských vztahů organismů se nazývá fylogenetika, které je věnována první kapitola práce.

1.1 Fylogenetika

Pojem fylogenetika v biologii označuje studium evoluční příbuznosti mezi skupinami organismů, např. mezi druhy, populacemi apod. Termín fylogenetika je odvozen z řeckých termínů phyle v překladu kmen a termínu genesis v překladu původ, zrození. [17]

Fylogenetika jako věda má za úkol analyzovat jednotlivé vývojové linie, které jsou odborně označovány jako taxony. Vychází z předpokladu, že organismy vytváří hierarchicky uspořádaný systém, ve kterém se všechny organismy vyvinuli ze společného předka (LUCA – *last univerzal common asesor*). [9]

Fylogeneze se zabývá popisem vzniku, štěpení, proměnami a vymíráním vývojových linií. Fylogenetika rekonstruuje průběh kladogeneze (větvení), ale všímá si i anageneze, tedy vývoje vlastností organismů v rámci linie. Proces fylogeneze bývá znázorňován fylogenetickými stromy, které zachycují posloupnosti dějů, které vedly k formování současné množiny druhů.

Klasická fylogenetika pracovala s morfologickými znaky, např. počet nohou, barva očí apod. V současné době se opírá o molekulární znaky, mezi které patří nukleotidové a proteinové sekvence. Tento samostatný obor systematické biologie se nazývá molekulární fylogenetika. [9,21]

1.2 Molekulární fylogenetika

V současné době jsou ve fylogenetice využívány molekulární znaky. Jedná se především o pořadí monomerů v řetězcích biopolymerů, případně chemické a fyzikální vlastnosti biopolymerů, které jsou zjišťovány pomocí molekulárně biologických metod.

1.2.1 Výhody molekulárních znaků

1. Jsou genetické. Je znám proces dělení, nejsou závislé na prostředí ani generickém pozadí. Právě na této úrovni vznikají evoluční mutace v DNA.
2. Je jich obrovské množství. Velikost genomu je odhadována mezi od $0{,}5.10^6 - 600.10^9$. Lidský genom obsahuje přes tři miliardy párů bází. Vědci odhadují, že se lidé mezi sebou liší v 0,1%, tj. třech milionech párů bází.
3. Jsou použitelné od nejvzdálenějších srovnání až po porovnávání jedinců téhož druhu (např. srovnání sekvence tygra bengálského s medúzou a s tygrem ussurijským).
4. Jsou selekčně neutrální, nepodléhají vlivům prostředí, ve kterém se organismus vyskytuje.
5. Jsou jednoznačně popsatelné. Genetická informace je v DNA zapsána v digitální formě, takže je možnost ji překódovat například do alfabetických znaků a předat ji v této podobě bez jakékoliv ztráty či zkreslení informace.
6. Jsou nezávislé. [10, 17]

Základními nevýhodami molekulárních znaků je, že neposkytují informace o anagenezi, někdy mohou mít destruktivní charakter.

1.2.2 Zpracování molekulárních dat

Většina molekulárněbiologických dat je získávána sekvenováním DNA. Jedná se o určení sekvence nukleotidů (resp. bází) v jednom z řetězců DNA (druhý řetězec je komplementární). Pro určení přesné sekvence nukleotidů v úseku DNA byly vynalezeny dvě metody – Sangerova a Maxam & Gilbertova metoda. V dnešní době je sekvenováním velmi rychlé a poměrně levné.

V praxi je pro vytvoření fylogenetického stromu výhodnější použití dat protetických. Pro účely fylogenetiky se však metody pro sekvenování proteomu vzhledem ke své technické náročnosti nepoužívají. Řešením je jednoduchý převod genomických dat na data proteomická. [7, 20]

Molekulární znaky získané sekvenováním (nejčastěji pořadí nukleotidů v DNA) je nutné srozumitelně popsat tak, aby umožnila počítačové zpracování. Používá se jednoduchý

zápis ve formě IUPAC (*International Union of Pure and Apllied Chemistry*) kódů, znázorněný v tabulce 1. [1]

Tabulka 1 IUPAC kódy pro nukleové kyseliny

IUPAC kód	význam
A	adenin
C	cytosin
G	guanin
T	thymin
U	uracil
N	cokoliv (any)

Pro zápis DNA monomerů jsou používány čtyři znaky A pro adenin, C pro cytosin, G pro guanin a T pro tymin. Pořadí v sekvenci odpovídá lineárnímu zápisu sledu monomerů v molekule DNA a vychází z biosyntézy makromolekul tj. pro DNA od 5' ke 3' konci. Ze čtyř druhů bází je možné uspořádat 64 kombinací trojic (tripletů), které mohou být potencionálními kodóny. Aminokyseliny kóduje pouze 61 tripletů, zbylé tři jsou terminační. Při přechodu ze zápisu nukleotidů na zápis aminokyselin dochází k nenávratné ztrátě informací – jedná se tedy o degenerovanou transformaci. Zpětná transformace je nejednoznačná, jelikož několik kodónů kóduje jednu aminokyselinu. [7]

Jak již bylo zmíněno, ve fylogenetice se častěji používají sekvence proteinů namísto sekvencí genomických. Tento zápis je znázorněn v tabulce 2. [1] Hlavním důvodem použití je, že sekvence proteinů jsou mnohem vice konzervované než sekvence nukleotidů, ve kterých se snadněji hromadí mutace. Jedna aminokyselina je kódována více triplety bází což znamená, že mutace v jednom nukleotidu nemusí nutně znamenat změnu v expresi genu, protože translací může vzniknout stejný protein jako v původním genu. [20]

Tabulka 2 IUPAC kódy pro aminokyseliny

IUPAC kód	zkratka AK	název AK
A	Ala	alanin
C	Cys	cytosin
D	Asp	asparát
E	Glu	glutamát
F	Phe	fenylalanin
G	Gly	glycin
H	His	histidin
I	Ile	izoleucin
K	Lys	lysin
L	Leu	leucin
M	Met	metionin
N	Asn	asparagin
P	Pro	prolin
Q	Gln	glutamin
R	Arg	arginin
S	Ser	serin
T	Thr	threonin
V	Val	valin
W	Tpr	tryptofan
Y	Tyr	tyrosin

1.2.3 Formát datového souboru

Při práci se sekvenčními daty je nutné uložit do paměti nejen sekvence, ale i další informace týkající se zisku těchto dat. Používá se proto několik různých formátů, které lze mezi sebou vzájemně převádět. Mezi nejpoužívanější se řadí formát FASTA, který je vhodný jak pro zápis proteinů, tak nukleotidů. Sekvence zapsaná ve formátu FASTA začíná jednořádkovým popisem, který je odlišen znaménkem větší než („>“) umístěným na začátku řádku, dále následuje „hlavička“, ve které je zapsán název sekvence a popis. Písmena zapsaná v hlavičce a ukončena mezerou představují identifikátor sekvence. Další řádky obsahují samotnou posloupnost sekvenčních znaků.

Výhodou formátu FASTA je možnost slučování a importování několika sekvencí do jednoho souboru a následné zpracovávání v programu, což bylo využito také v této diplomové práci. Sekvence lze pro přehlednost oddělit prázdným řádkem, který ale není nezbytný. [7]

1.2.4 Primární databáze sekvencí

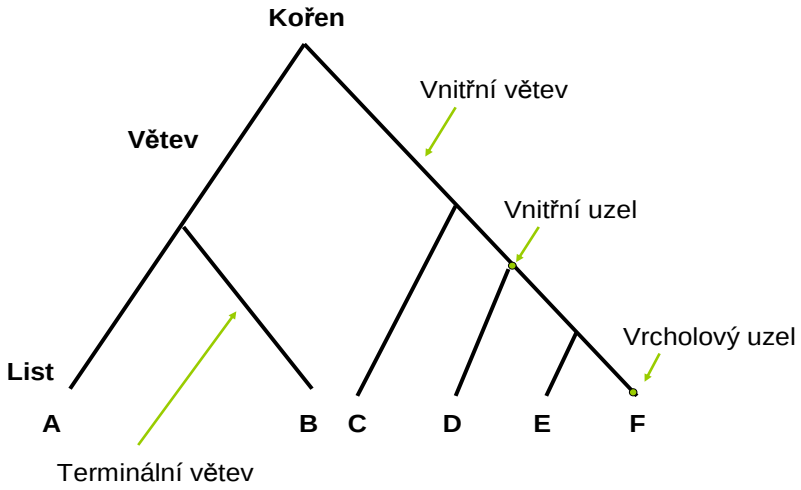
Rozvoj technik sekvenování a rostoucí počet získaných dat vedl k založení trojice primárních databází. Patří sem americká databáze GenBank, japonská DDBJ a evropská EMBL. Všechny tři jsou propojeny a vzájemně se vyměňují informace. Data jsou přijímána

z projektů sekvenování od institucí i jednotlivců a lze je získat prostřednictvím webových rozhraní nebo lze bezplatně získat celou databázi prostřednictvím FTP. Nevýhodou těchto databází je nulová záruka získání kvalitních a správných dat.

1.3 Fylogenetické stromy

Grafickým zobrazením průběhu fylogeneze je fylogenetický strom. Pokud je ve stromu ukázána souvislost mezi organismy a je vztažena k časové ose, jedná se o zakořeněný strom (Obrázek 1). V tomto případě jsou z kořene stromu (*root*), který představuje posledního společného předka, postupně odvětvovány větve (*branches*). Větve ve stromu tedy představují určitý druh organismu. Pořadí a místo větvení odráží časovou posloupnost vzájemně odvětvených vývojových linií. Tento děj je ve fylogenetickém stromu znázorněn pomocí uzlů (*nodes*). Terminální (koncové) větve znázorňují jednotlivé druhy, které byly zahrnuty do analýzy. Listem se rozumí konkrétní organismus, který byl do analýzy zahrnut. [9, 21]

Druhy jedné vývojové linie vytvářejí pro účely fylogenetických studií tzv. operační taxonomickou jednotku OTU (*Operation Taxonomic Unit*).



Obrázek 1 Anatomie fylogenetického stromu [5]

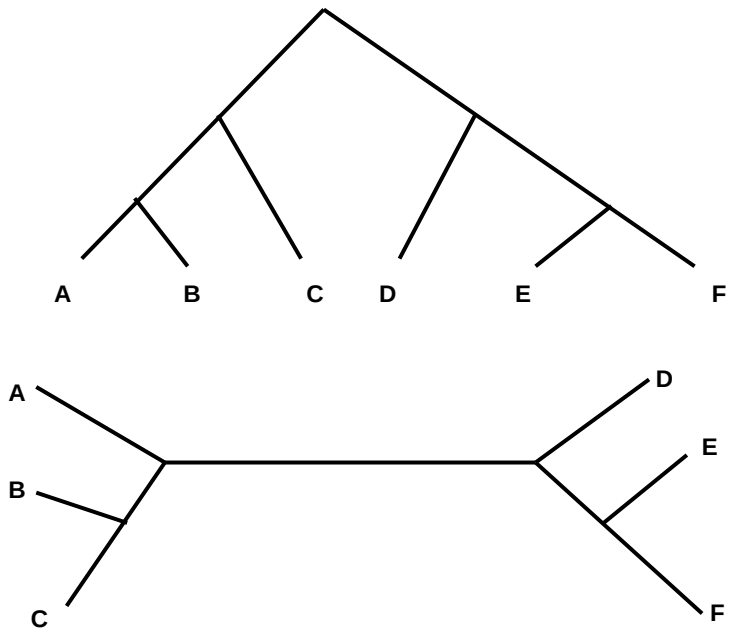
V případě, že délka větví vyjadřuje dobu trvání jednotlivých druhů či množství evolučních změn, ke kterým v jednotlivých větvích došlo, jedná se o fylogenetický strom, tj. fylogram (*phylogenetic tree*). Druhy stromů jsou probrány v následující podkapitole.

1.3.1 Typy stromů

Fylogenetické stromy můžeme dělit z několika hledisek. Podle toho jestli mají kořen na kořenové a nekořenové, podle obsahu znázorněných informací na kladogram, fenogram nebo fylogram.

Kořenový strom je přirozeným modelem a vykazuje pro všechny druhy společného předka všech OTU. Hrany stromu získaly přirozenou orientaci ve směru od kořene k listům. Podrobněji byl tento typ stromu popsán v předchozím odstavci.

V případě nekořenového stromu není nejstarší společný předek přesně identifikován. Umístění kořene v nekořenovém stromu však lze odhadnout vložením vnější skupiny, která je dostatečně vzdálená od ostatních druhů. Následující obrázek (Obrázek 2) znázorňuje oba zmiňované druhy stromů.



Obrázek 2 Nahoře zakořeněný strom, dole strom nezakořeněný

Kladogram neboli schéma kladogeneze je nejjednodušší znázornění stromu. Vyjadřuje pouze pořadí odvětvování jednotlivých druhů (OTU), nikoli míru jejich vzájemné příbuznosti. Délky větví nemají žádný význam.

Fenogram, neboli fenetický strom, vyjadřuje podobnost studovaných taxonů. V analýze je brán zřetel na fenetickou příbuznost jednotlivých OTU. Délka větví vyjadřuje počet změn nebo dobu trvání vzniku nového druhu.

Strom nazývaný fylogram vyjadřuje vzájemnou příbuznost porovnaných organismů (pořadí odvětvování od společného předka). Může být buď kladogramem vyjadřující pouze

pořadí odvětvování jednotlivých taxonů od společných předků, nebo zahrnuje i informace o anagenezi jednotlivých organismů. [9]

1.3.2 Metody konstrukce dendrogramů

Pojem dendrogram označuje speciální typ stromu, při kterém je jako vstupní formát pro konstrukci použita matice genetických vzdáleností (*distance matrix methods*) nebo se využívají znakové metody konstrukce (*cluster analysis*), tedy souboru metod zabývajících se podobností vícerozměrných objektů.

Vlastní konstrukce dendrogramu je založena na jednom z následujících dvou principů. V prvním případě je snaha o sestrojení jediného „nejlepšího možného“ stromu. Druhou možností je sestrojení celé množiny všech možných dendrogramů, které lze sestrojit na základě daných vstupních dat. Výstupem tohoto vyhledávacího postupu je několik nejlepších stromů, se kterými jde dále pracovat. Vyhledávací algoritmy jsou s rostoucím počtem OTU výpočetně náročné.

Následující vzorec (1.1) představuje vztah mezi počtem možných vytvořených zakořeněných stromů (N) a počtem sekvencí (n), ze kterých je tento strom vytvořen.

$$N = \frac{(2n-3)!}{2^{n-2}(n-2)!}. \tag{1.1}$$

Na nezakořeněný strom lze pohlížet jako na kořenový strom, kterému chybí jedna větev (kořen). Počet takto vytvořených stromů je tedy roven celkovému počtu sekvencí, od které jednu sekvenci odečteme, což vyjadřuje vztah 1.2.

$$N = \frac{(2n-5)!}{2^{n-3}(n-3)!}. \tag{1.2}$$

Pro přehlednost následuje tabulka s možným počtem dendrogramů pro daný počet sekvencí.

Tabulka 3 Počet možných dendrogramů pro daný počet sekvencí

počet OTU	počet kořenových stromů	počet nekořenových stromů
1	1	1
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10 395	945
8	135 135	10 395
9	2 027 025	135 135
10	34 459 425	2 027 025
15	2×10^{14}	8×10^{12}
20	8×10^{21}	2×10^{20}
30	5×10^{38}	9×10^{36}
50	3×10^{76}	3×10^{74}

Řešením výpočetní náročnosti je použití heuristických metod místo metod deterministických, které z analýzy předem vyloučí možnosti, u kterých není předpokládán správný výsledek. Bohužel tento předpoklad může do analýzy zanést chybu.

1.4 Rekonstrukční techniky

Ke konstrukci fylogenetických stromů jsou využívány dva druhy metod. První je metoda konstrukce na základě znakových (kvalitativních) dat, kdy je použita přímo sekvence znaků. Na každé pozici je digitální znak, který nabývá přesně vymezených hodnot. Tyto metody pracují s pravděpodobností změny hodnoty, která je rovna substituci aminokyseliny nebo nukleotidu. Do této skupiny patří metoda maximální parsimonie, která funguje na principu vyhledání všech stromů a následném zjišťování, který strom obsahuje nejmenší počet evolučních změn (mutací). Druhou metodou pracující se znakovými daty je metoda maximální věrohodnosti (*Maximum Likelihood*), která je velmi spolehlivá, jelikož prohledává všechna dostupná data. U každého sestrojeného stromu je uváděna pravděpodobnost, s jakou mohl vytvářet soubory znaků odpovídající vstupním datům. Výhoda této metody je však vykoupena náročným algoritmem, který je prakticky nepoužitelný pro delší sekvence. [21]

Tato diplomová práce se znakovými metodami nezaobírá, proto byly tyto metody zmíněny pouze v úvodu podkapitoly.

1.4.1 Distanční metody

V následující kapitole bude probrán druhý zmiňovaný přístup – konstrukce fylogenetických stromů na základě distančních (kvantitativních) dat. Jedná se o metody založené na výpočtu podobností mezi jednotlivými sekvencemi. Data vzdáleností nejsou sekvence znaků, nýbrž matice vzdáleností. V prvém kroku je nutné sekvence zarovnat a následně mezi sebou porovnat, jelikož vzdálenost mezi dvěma OTU je dána počtem bodových mutací. Počet bodových mutací mezi dvěma sekvencemi vztažený na délku sekvence je označován jako tzv. proporcionální vzdálenost (distance) sekvencí neboli p-distance.

Distanční vzdálenost neodpovídá skutečné evoluční vzdálenosti. Aby bylo možné zkonstruovat fylogram, který bude znázorňovat evoluční vzdálenost, je nutné provést korekci na vícenásobné a zpětné mutace ve stejné pozici a na paralelní vzájemně nezávislé mutace, případně výskyt konvergentních mutací. Mezi distanční metody patří shluková analýza UPGMA (*unweighted pair group method with arithmetic averages),* metoda nejbližšího souseda (*Neighbor-joining metoda*) značená NJ a metoda minimální evoluce (*minimum evolution*).

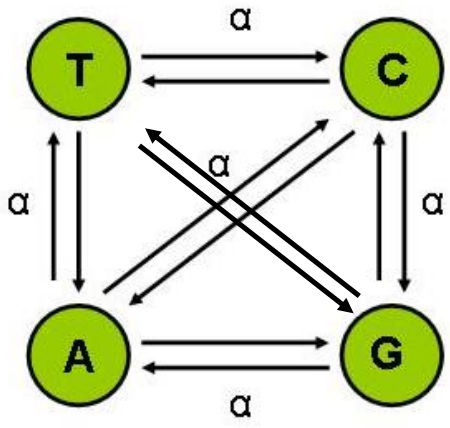
Je třeba také zmínit modely evolučního vývoje v biologické sekvenci. Tyto modely se snaží vytvořit matematický popis bodových mutací, ke kterým mohlo v průběhu evoluce

v sekvenci dojít. Matematické modely lze použít k výpočtu evoluční vzdálenosti mezi dvěma a více sekvencemi. Na základě tohoto popisu sekvencí (organismů) lze rekonstruovat jejich evoluční vývoj. Bodové mutace se v sekvenci DNA dělí na inzerci, deleci a substituci. [20, 21]

1.4.2 Jukes-Cantorův evoluční model

Nejjednodušším evolučním modelem umožňující korekci je jedno parametrický Jukes-Cantorův model, označovaný jako J-C model. Parametr α popisuje pravděpodobnost změny jednoho nukleotidu na jiný. Tento model předpokládá, že každý nukleotid má stejnou pravděpodobnost změny na jiný nukleotid a stejnou frekvenci výskytu. Nukleotidové sekvence si jsou tedy ekvivalentní.

Následující schéma (Obrázek 3) jasně popisuje, že změna frekvence přechodu jednoho nukleotidu na jiný je rovna α, frekvence přeměny báze na jinou je 3α. Průměrný počet změn je dán součinem frekvence změn a času.



Obrázek 3 Jukes – Cantorův jednoparametrický model

Parametr α je pro genomické sekvence roven $\alpha = \frac{3}{4}$, jelikož máme čtyři druhy nukleotidů. Pro protetické sekvence je $\alpha = \frac{19}{20}$, jelikož je známo dvacet aminokyselin.

Vztah 1.3 umožňuje vypočítat evoluční vzdálenost d bez znalosti času a frekvence výskytu změn. Hodnota p vyjadřuje relativní počet změn mezi sekvencemi.

$$d = -\frac{3}{4}\ln(1-\frac{4}{3}p) \tag{1.3}$$

Rovnice 1.4 představuje matematický zápis výpočtu pravděpodobného počtu substitucí p, ke kterým došlo u dvou druhů od okamžiku větvení ze společného předka. Parametr D popisuje množství pozorovaných rozdílů mezi dvěma danými sekvencemi.

$$d = -\frac{3}{4}\ln(1-\frac{4}{3}D) \tag{1.4}$$

Jukes-Cantorův model je nejjednodušší evoluční model, který dává přesné výsledky pouze, pokud není množství rozdílů mezi porovnávanými sekvencemi příliš velké. Pokud je procento rozdílů velké a genetická vzdálenost vychází vyšší jak jedna, je vhodnější použít složitější a přesnější model, kterým je například Kimurův dvou parametrový model. Tento model předpokládá, že pravděpodobnost translací a transverzí se navzájem liší.

1.4.3 UPGMA

Základní a nejjednodušší metodou konstrukce fylogenetického stromu je shluková metoda UPGMA (*Unweighted Pair Group Method with Aritmetic mean*). Anglická slova, ze jejichž začátečních písmen byl název metody odvozen, nesou základní vlastnosti metody. Unweighted, v překladu neváženy, udává, že všechny párové vzdálenosti mají stejný vliv na tvorbu stromu. Pair group, dvojice, značí, že shluky jsou vytvářeny kombinací dvou hodnot. Arithmetic mean, aritmetický průměr, je statistická metoda, kterou jsou dopočítávány párové vzdálenosti ke každému shluku, které jsou střední hodnotou vzdálenosti ke všem členům. [12,14]

Tato metoda předpokládá, že substituční rychlost je konstantní, takže distance je přímo úměrná času a dále pak, že všechny dnešní taxony „domutovaly“ stejně daleko. Tyto předpoklady jsou však téměř vždy porušeny, je tedy možné, že touto metodou vytvořený strom (fenogram) je nesprávný.

Následuje podrobně popsaný postup pro vytvoření fylogenetického stromu metodou UPGMA [12, 19].

1. Sestrojit trojúhelníkovou matici vzdáleností každého druhu s každým.
2. Nalézt v tabulce dvě sekvence s nejmenší vzdáleností, které se spojí pomocí dvou větví a uzlů.
3. V matici obě sekvence nahradit nově spojenou dvojici.
4. Vypočítat vzdálenost od ostatních sekvencí. Uzel je umístěn do těžiště mezi oběma sekvencemi – je tedy aritmetickým průměrem.
5. Postup opakovat, dokud nedojde ke spojení všech OTU.
6. Sestrojit dendrogram postupným spojováním jednotlivých dvojic ve stejném pořadí, v jakém jsme je spojovali v tabulce.
7. Kořen se umisťuje do středu poslední větve.

1.4.4 Minimální evoluce (ME)

Metoda minimální evoluce (*minimum evolution*) zpracovává distanční data pomocí kritéria optimality. Algoritmus této metody je velice náročný, jelikož počítá sumu délek všech

větví pro všechny možné stromy. Z těchto stromů je vybrán ten, jehož součet délek větví je nejmenší. Délky větví musí být vypočítány metodou nejmenších čtverců. Výsledný strom se většinou příliš neliší od stromu získaného metodou NJ. Určitého zjednodušení lze dosáhnout využitím heuristických algoritmů, které z analýzy vyloučí některé stromy.

V následujících krocích je metoda nejmenších čtverců (*least squares*) bodově popsána [21].

1. Je známa genetická vzdálenost všech párů sekvencí.
2. Vezme se první topologie a zkouší se, jak dobře do ní distance pasují. Délky větví se mění tak, aby pasovaly co nejlépe. Nejlepší skóre vycházející z rovnice 1.5 je uchováno v paměti.

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(D_{ij} - d_{ij})^2. \tag{1.5}$$

3. Skóre se určí i u dalších topologií.
4. Projít všechny topologie a vybrat tu s celkově nejlepším skóre.

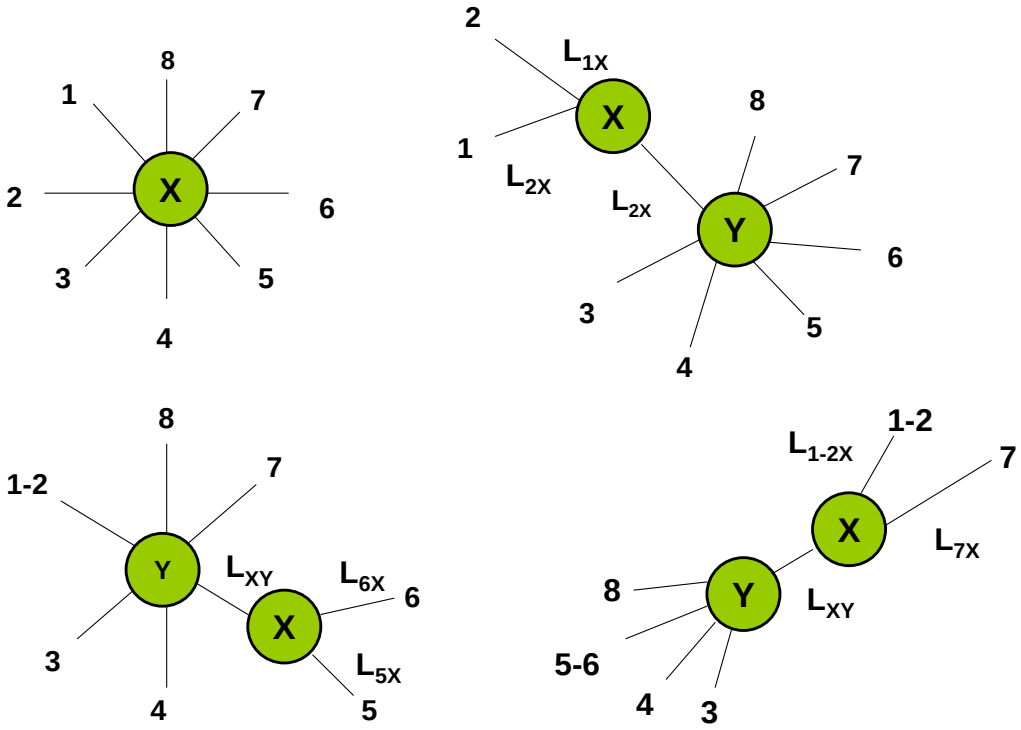
Nejmenší čtverce garantují nalezení správného stromu, pokud jsou dobře spočítané distance. U metody minimální evoluce jsou délky větví optimalizovány úplně stejně, jako v případě nejmenších čtverců, ale topologie vzájemně porovnáváme podle součtu délek všech větví a vybereme tu s nejmenším součtem podle rovnice 1.6.

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij} \tag{1.6}$$

1.4.5 Neighbor-joining (NJ)

Metoda s názvem neighbor-joining, neboli metoda spojování sousedů, je aproximací metody minimální evoluce. Je to metoda konstrukce dendrogramu na základě distančních dat, která na rozdíl od UPGMA nepředpokládá konstantní substituční rychlost v jednotlivých větvích. Metoda vychází z hvězdicového uspořádání všech OTU s centrálním uzlem (obr. 4 vlevo), kdy je v každém kroku spojena ta dvojice OTU, jejíž spojení nejvíce zmenší délku stromu (součet délek všech jeho větví). Základem je hvězdicový strom. Postupně dochází ke spojování větví, které jsou nahrazovány novými uzly. V každém kroku je vytvořen nový strom pro menší množinu objektů. Celý postup je opakován, dokud není dosaženo binárního větvení. Výsledkem této metody je fylogram (Obrázek 4 – vpravo dole) – metoda tedy zohledňuje nestejnou evoluční rychlost. [19, 27]

Jedná se o úsporný algoritmus, je tedy možné zpracovávat jak velké množství sekvencí, tak velmi dlouhé sekvence.



Obrázek 4 Konstrukce dendrogramu metodou neighbor-joining

2 Fylogenetické superstromy

Metoda superstromů kombinuje více fylogenetických stromů k vypracování celkově lepších superstromů. Lze ji použít pro kombinaci fylogenetických informací z databází, kde se datové soubory částečně překrývají a kde jsou informace získány z různých zdrojů získaných pomocí různých typů dat (např. DNA, proteiny, morfologie) nebo algoritmů (např. metodou parsimonie, pravděpodobnosti, vzdálenostní). [2, 6]

První zmínka o superstromech je datována na začátek roku 1980. Obecně zde byly popsány soubory pravidel popisu fylogenetických informací ze zdrojových stromů, které mohou být kombinovány. Různé metody používají odlišná pravidla. Konečný výsledek by měl umožňovat nejen kombinaci informací obsažených ve zdrojovém stromě, ale také závěry o vztazích, které jsou ve zdrojovém stromě. Vzniklý superstrom by neměl obsahovat žádné rozporuplné vztahy ke zdrojovému stromu.

Vzhledem k rostoucí popularitě metody superstromů, bylo nutné vytvořit hodnocení výkonnosti těchto algoritmů, které by odhalili nejlepší možné řešení. Superstromy lze hodnotit podle několika kritérií. Základní kritériem je hodnocení podobnosti superstromů se vstupními stromy, dále pak zkoumání podobnosti mezi velkými stromy a celkovým stromem nebo dle úrovně řešení a výpočetního času algoritmu. [19, 27]

2.1 Superstromy

Superstrom kombinuje informace ze souboru taxonomicky se překrývajících fylogenetických stromů a vytvoří z nich superstrom nebo sadu superstromů na stejně dobré úrovni. Superstrom obsahuje kompletní sadu všech listů nacházejících se ve vstupním stromu. Analýza vyžaduje, aby zdroje fylogenetických stromů byly spojeny se soubory sdílenými taxony. Zdroje stromů, které nemají společné taxony, nelze kombinovat. Případné řešení by nabízel třetí strom, který by sdílen taxony s oběma stromy. [2, 19]

O superstromu mluvíme právě tehdy, když dílčí stromy neobsahují stejnou množinu OTU. Pokud by obsahovaly stejnou množinu OTU, jednalo by se o konsenzuální strom, který má odlišný přístup ke konstrukci. Konsensuální strom nesmíme se superstromem zaměňovat.

2.1.1 Výhody superstromů

1. **Souhrnné informace**

Je zde mnoho důvodů, proč využít právě superstromy. Hlavním důvodem je, že k vyřešení daného problému lze použít více dat, což zvyšuje pravděpodobnost dosažení lepšího výsledku. Metoda superstromů umožňuje pracovat s informacemi z více různých

zdrojů, lze tedy vyřešit případy, které by nebyly možné odvodit pouze z jednoho zdroje. Je možné kombinovat genetická data s morfologickými daty. Tato kombinace umožnila vytvoření databáze, která obsahuje jak makroevoluční, tak mikroevoluční informace.

Z praktického hlediska málo datových sad obsahuje naprosto stejné informace o druhu, kmenu či proteinech, takže jejich kombinování bývá obtížné. U superstromů je teoretická možnost využití 100% informace pro rekonstrukci stromu života. Bohužel je velmi pravděpodobné, že chybí stále ještě mnoho dat. Jediným omezením pro použití metody superstromů je, že musí být možné reprezentovat data jako stromy.

2. **Určení podobných stromů**

V reálných databázích není největší problém překrývání dat, ale odlišnost typů dat. K chybě může dojít při generování zdrojových stromů, chybné homologii, záměně řádků, horizontálnímu přenosu genů apod. Pokud se jedná o malé rozdíly, jsou v genu náhodně rozděleny a kombinací mnoha genů bývá druh stromu odhalen. Pokud jsou však velké rozdíly, informace se nekombinují, ale zjistí se míra kompatibility v rámci jednotlivých zdrojových stromů.

3. **Možnost sestrojení velkého stromu z několika malých**

Biologické databáze se den ode dne zvětšují a pro uživatele se stávají nepřehledné a výpočetně obtížné. Zde se uplatňuje metoda „Rozdělej a panuj“, kde je jeden velký problém rozdělen na řadu menších. S každou podskupinou se dá snadno manipulovat a v rámci podskupiny samostatně řešit. V závěru jsou tyto podskupiny pospojovány a je nalezeno souhrnné řešení. Metoda superstromů je pro tento přístup zcela vhodná – je jednodušší pospojovat několik malých stromů než vytvořit jeden velký. [2, 6]

2.2 Metody konstrukce superstromů

Existuje několik různých metod, každá přistupuje ke kombinaci informací z více stromů různými způsoby. V této podkapitole budou zmíněny tři nejpoužívanější – průměrný konsensus, matrix reprezentace a MSSA typ. Největší část bude věnována metodě průměrný konsensus, jelikož právě tato metoda byla použita k vypracování diplomové práce.

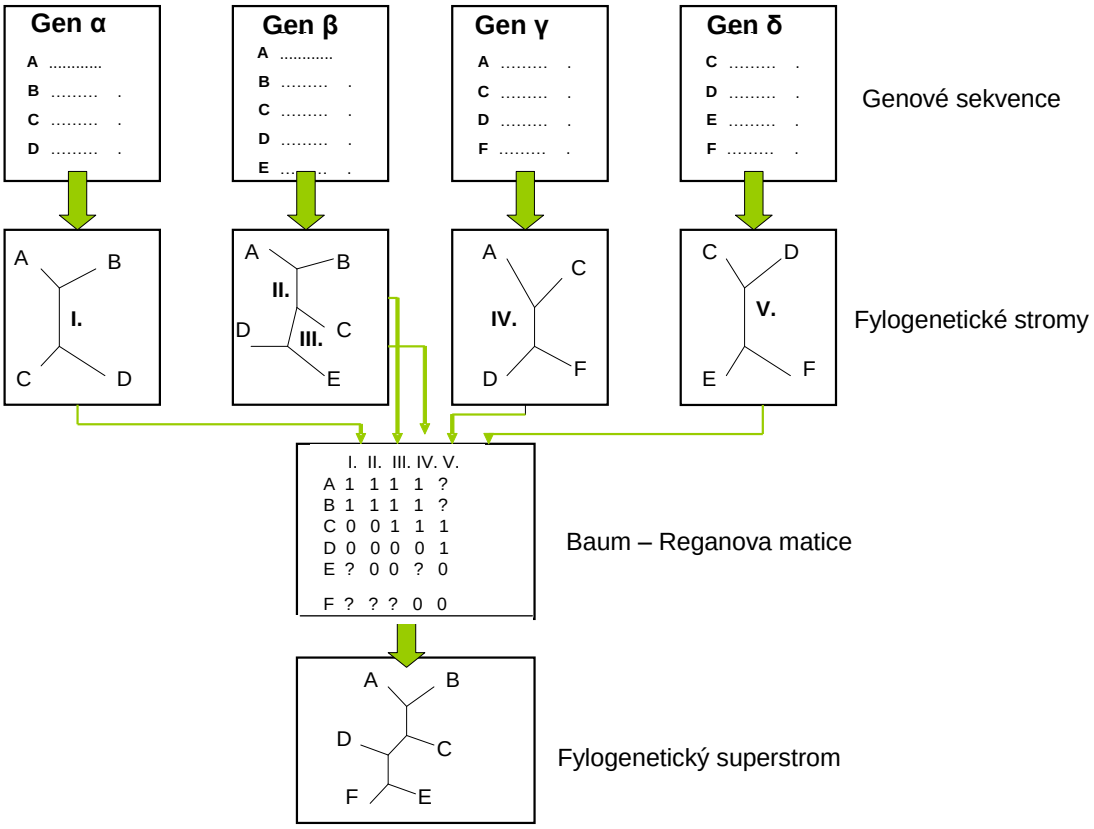
2.2.1 Maticová reprezentace úspornosti (MPR)

Jedná se o nejpoužívanější metodu nezávisle popsanou pány B. R. Baumem a M. A Reganem. Je nazývaná maticovou reprezentací úspornosti (*Matrix representation with parsimony* - MPR). Využívá kódovací schéma pro konstrukci matice, která reprezentuje

vztahy mezi zdrojovými stromy. Obvykle je tento algoritmus využit pro rekonstrukci superstromu z této matice.

Tato metoda určuje vnitřní větve v rámci každého ze zdrojových stromů a aplikuje se jednoduché kódovací schéma 0 a 1, které se využívá k určení taxonů na obou stranách větve (Obrázek 5). Všechny taxony na jedné straně jsou označeny 1, taxony na straně druhé jsou označeny 0. Pokud není na určité pozici taxon, využívá se značení „?“. Pro nekořenové zdrojové stromy (jedná se o nejběžnější data) nezáleží na označení větve 1 nebo 0. Zakódování všech vnitřních větví je poté spojeno do jedné matice, která obsahuje sloupec pro každou vnitřní větev.

Metoda maticové reprezentace šetrnosti dosahovala v doposud provedených analýzách dobrých výsledků a vysokou úroveň kvality i při použití jednoduchého heuristického hledání. Tato matice je využívána ke konstrukci superstromu. [2, 19]

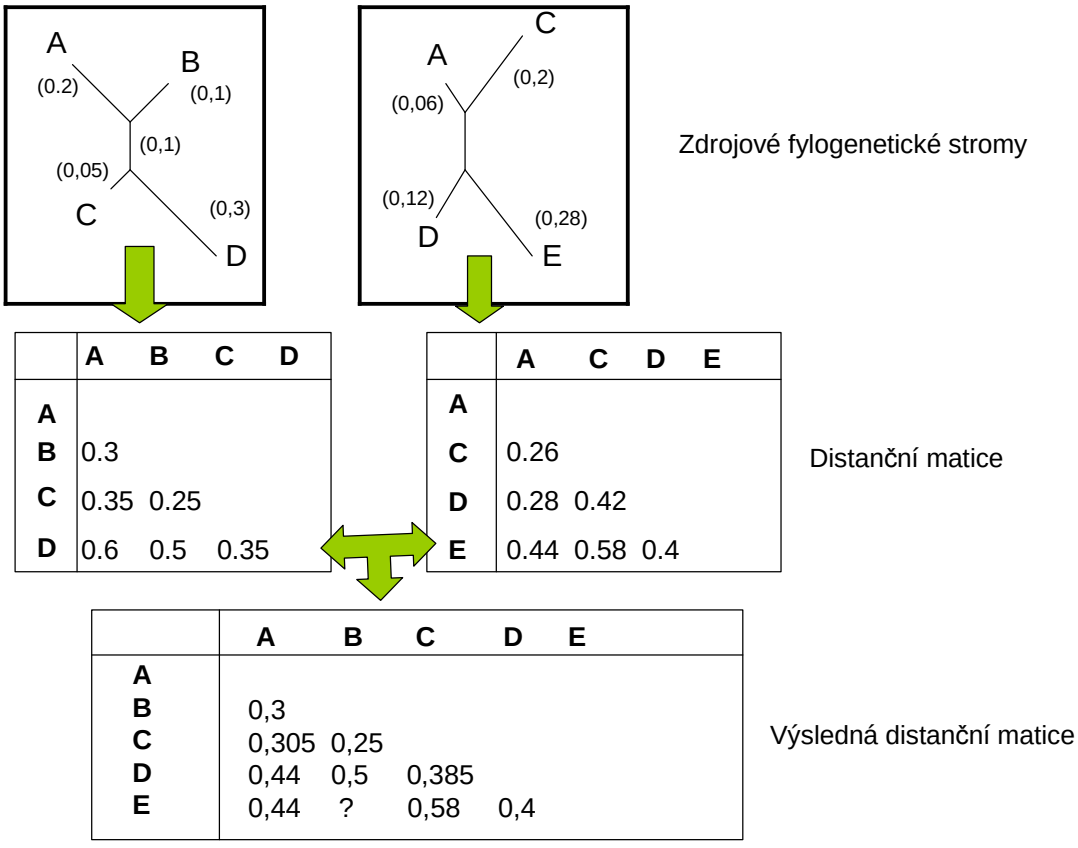


Obrázek 5 Maticová reprezentace pomocí MPR [6]

2.2.2 Průměrný konsensus

Druhý přístup k rekonstrukci superstromu zahrnuje výpočet vzdáleností mezi zdrojovými stromy. Tyto metody využívají délky větví zdrojových stromů. Jedna z těchto metod se nazývá „průměrný konsenzus“. Tento přístup počítá vzdálenosti mezi všemi zdrojovými stromy. Průměrná vzdálenost každého taxonu je použita v konečné matici vzdáleností, ze které je sestrojen superstrom. Někdy však může nastat situace, kdy se dva taxony nikde nevyskytují společně. V tomto případě je využíván odhad vzdálenosti. V podstatě se jedná o vyplňování „prázdných míst“, kde neexistují žádné informace o evolučních vztazích. K tomu byly vyvinuty speciální metody, které tyto chybějící hodnoty vypočítají. Jakmile je konsensuální matice kompletní, lze rekonstruovat superstrom. Nejčastěji je používaná metoda nejmenších čtverců, ale lze využít i metodu spojování sousedu (NJ). Výpočet průměrného konsensu je znázorněn na obrázku 6. [6]

Výhodou této metody je, že je v superstromu zahrnuta (znázorněna) i délka větví.



Obrázek 6 Metoda průměrného konsensu [6]

Pro dopočítávání prázdných míst se používají dvě metody – ultraparametrická a aditivní, které odhadují chybějící části před fylogenetickou rekonstrukcí za použití matematických metod. V praktické části jsou obě použity, proto jim budou věnovány následující řádky.

Obě metody slouží k nalezení chybějících záznamů v evolučních datech. Tento případ nastává, pokud pozorované sekvence nukleotidů nebo proteinů obsahují mezery nebo chybějící položky. Tyto neúplné soubory dat mohou vznikat za různých situací. Mohou být způsobeny nedostatkem biologického materiálu, nepřesností experimentálních metod nebo kombinací nepředvídatelných faktorů. Kromě toho experimentální techniky jako DNA-hybridizace, komparativní sérologie nebo mikroarray hybridizace jsou limitovány a často vedou ke vzniku neúplné distanční matice. [15]

V této diplomové práci došlo ke kombinaci částečně se překrývajících fylogenetických stromů, které jsou odvozeny z různých zdrojů. Vytvořena byla nekompletní distanční matice, která má sloužit pro výpočet superstromů, kde používané metody Neighbor-Joining a UPGMA vyžadují plné matice vzdáleností mezi všemi druhy.

Existují různé přístupy navržené k řešení náročných problémů při odvozování fylogeneze z parciálních distančních matic. Přímý přístup umožňuje sestavit fylogenetický strom z částečné matice vzdáleností pomocí specifického stromu (*building algorithm*). V této práci byly využity nepřímé metody spoléhající na odhad chybějící části před fylogenetickou rekonstrukcí za použití matematických metod. Mezi základní nepřímé metody patří ultrametrická metoda a aditivní metoda. [15]

1. Ultrametrická metoda

Ultrametrická metoda dopočítává neznámou distanční vzdálenost *d(i,j)* ze dvou taxonů, které se v tomto bodě protínají. Na výstup je zaslána maximální nalezená hodnota. Tento vztah je matematicky zapsán ve vztahu 2.1, ve kterém index *k* představuje taxon, ve kterém jsou distanční hodnoty známé. Chybějící hodnota v trojúhelníku je nalezena právě tehdy, pokud jsou odlišné. Pokud jsou dvě dostupné vzdálenosti stejné, nelze chybějící hodnotu odhadnout

$$d(i, j) \leq \text{Max}\{d(i, k); d(j, k)\} \ \forall\, i, j,k \in X \qquad (2.1)$$

. Vstupem u ultrametrické metody je částečná (parciální) matice, výstupem kompletní matice vzdáleností, ze které lze konstruovat superstrom. V rovnici (2.1) jsou známé vstupní distance označeny *d(i, k)*; *d(j, k)*, hledaná distanční hodnota je indexována *d(i, j)*.

Ultrametrická procedura se používá pro odhad chybějících buněk v případě, že matice vzdáleností má více než předem stanovené procento chybějících položek (toto procento se volí v závislosti na rozměru matice). V opačném případě se použijí aditivní procedury.

2. Aditivní metoda

Aditivní metoda je náročnější obdobou metody ultrametrické. Při dopočítávání neznámé distanční hodnoty jsou do výpočtu zahrnuty dva taxony (*k, l*), pro které jsou distanční hodnoty známé. Výsledná hodnota je maximální součet vertikálních hodnot taxonu *k* s horizontálními hodnotami taxonu *l* a vertikálních hodnot taxonu *l* s horizontálním taxonem *k*. Na výstup je odeslána nejvyšší hodnota z těchto součtů zmenšená o hodnotu distance, ve které se oba taxony protnuly. Opět platí podmínka, že hodnoty součtu nesmí být stejné. Rovnice 2.2 představuje aditivní metodu. [7]

$$d(i,j)+d(k,l) \le Max\{d(i,k)+d(j,l);d(i,l)+d(j,k)\} \forall i,j,k \in X \qquad (2.2)$$

Pro zjednodušení výpočtů lze rovnici zjednodušit (2.3)

$$d(i,j) \le Max\{(d(i,k)+d(j,l);d(i,l)+d(j,k))-d(k,l)\} \forall \mathrm{i,j,k,l} \in \mathrm{X} \qquad (2.3)$$

I u této metody jsou vstupem známé distanční hodnoty *d(i, k)*, *d(j, k)*, *d(i, l)*, *d(j, l)*, *d(k, l)*. Po výpočtu je na výstup zaslána dopočítaná neznámá hodnota *d(i,k)*.

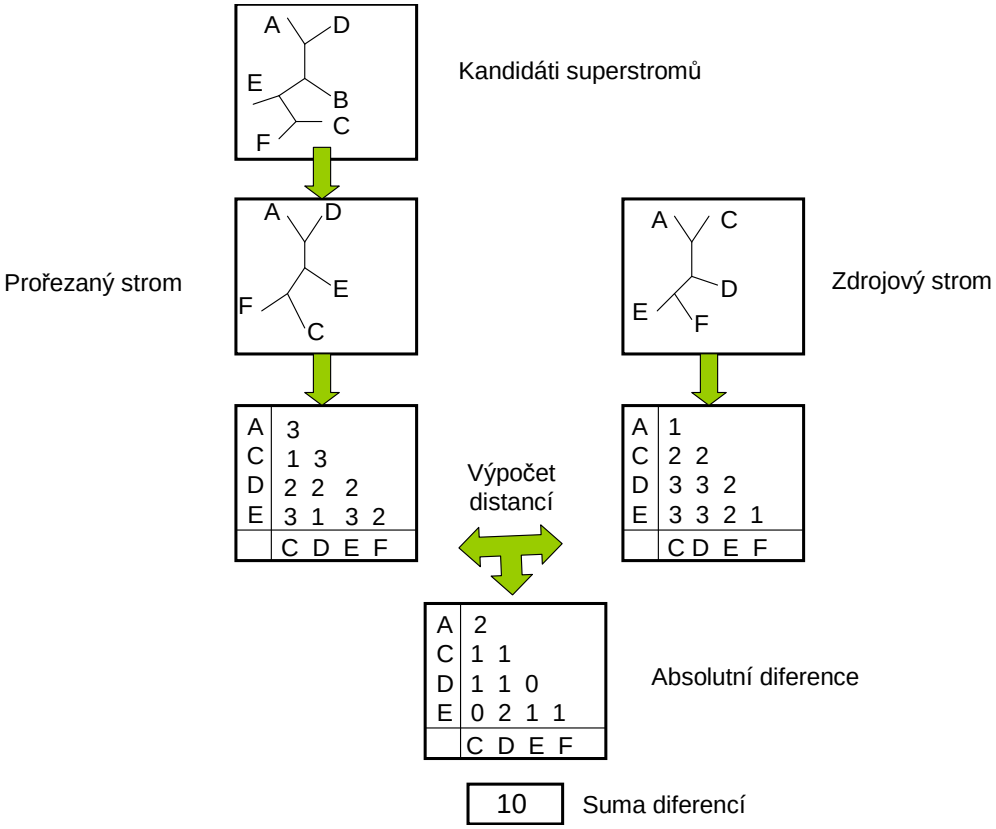
2.2.3 MSSA

Metoda MSSA (*The most similar supertree algortihm*) hledá nejlepší superstrom bez průměrování informací ze zdrojových stromů. Metoda využívá funkci, která slouží k posouzení kandidáta superstromu. [6]

Jedná se o heuristické hledání superstromu s minimální funkcí, který je nejvíce podobný některému superstromu ze souboru zdrojových stromů. Rozdíl mezi kandidátem superstromu a každým zdrojovým stromem se vypočítá samostatně. Celkové skóre získáme sečtením jednotlivých skóre.

Při rekonstrukci fylogenetického superstromu touto metod se využívá hledání superstromu s využitím bodových funkcí, které jsou minimalizovány a vrací strom, který je nejvíce podobný některému ze zdrojových stromů. Tato skórovací funkce pracuje na principu individuálního porovnávání daného stromu se všemi stromy, které jsou obsaženy ve zdrojových stromech. Superstrom obsahuje všechny taxony a všechny zdrojové stromy, je tedy pravděpodobné, že obsahuje také námi hledanou podskupinu (kandidáta), která se odvětvila ze stejného souboru taxonů. Rozdíl mezi dvěma stromy se vypočítá jako suma absolutních diferencí (rozdílů) vzdáleností matic mezi dvěma stromy. V tomto případě je délka cesty definována počtem vnitřních uzlů oddělených dvěma taxony na stromě. Tato srovnávací metoda se provádí pro každý zdrojový strom. Součet absolutních odchylek je výsledek popisující podobnost kandidáta superstromu ku souboru zdrojových stromů. Pokud je výsledná hodnota nula, každý zdrojový strom je identický se superstromem. Pro nalezení

minimální skórovací funkce musí být testováno více kandidátů. Pro vyhledávání je možné použít metodu nejbližšího zvanou *Nearest neigbor interchange* – NNI. Schéma metody MSSA je znázorněno na obrázku 7.



Obrázek 7 MSSA algoritmus [6]

3 Použitá data

Data zpracovávaná v této diplomové práci jsou získána z centrální databáze proteinových sekvencí – UniProt, která obsahuje komplexní a vysoce kvalitní informace, které byly získány sekvenováním. Jako výchozí byl zvolen protein MT - ND1, známého také pod jmény MTND1, NADH dehydrogenace subunit 1, NADH ubiquinone oxidoreductace chain 1. Tento gen je součástí velkého komplexu enzymů známého jako komplex I., který je umístěn ve vnitřní mitochondriální membráně, a je nezbytný pro oxidační fosforylaci, při které přenáší elektrony z NADH do respiračního řetězce. [25]

Byly vybrány čtyři třídy z živočišné říše, které se ve vědecké klasifikaci řadí do kmene strunatců a podkmene obratlovců. Jedná se o třídu savců, plazů, ptáků a ryb. Vzhledem k širokému záběru využití fylogenetických superstromů, jsou tyto čtyři třídy doplněny o třídu hmyzu, která patří do kmene členovců. Z každé třídy bylo účelně vybráno deset zástupců živočichů, které reprezentují jednotlivé rody, obydlené kontinenty a životní prostředí. Speciální místo zaujímá člověk, který bude při rekonstrukci fylogenetického superstromu představovat sdílený taxon, bez které by nebylo možné superstrom sestrojit.

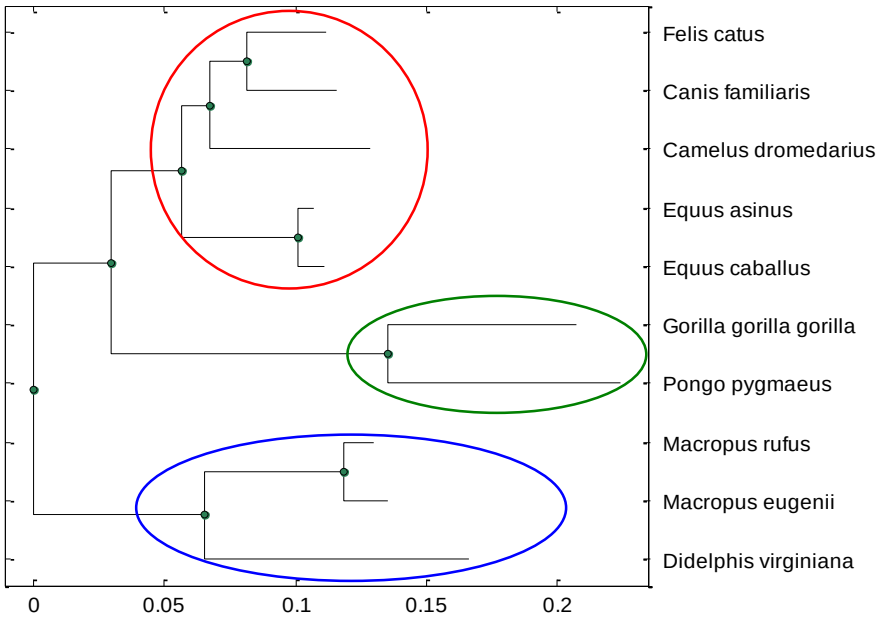
V následující podkapitole budou použitá data představena podrobněji.

3.1 Savci – Mammalia

Vybraní zástupci savců mají ve své sekvenci shodně 318 aminokyselin. Při provedené předběžné analýze bylo zjištěno, že se shodují přibližně v 60%, což odpovídá 188 identickým pozicím. Maximální distance v této skupině dosahuje hodnoty 0,25. Z tabulky 4 je patrné, že zástupci vybráni z třídy savců reprezentují celkem osm čeledí. Data v tabulce jsou seřazeny podle řádu, který umožňuje snadnou orientaci ve fylogenetickém stromu na obrázku 8. Skórovací matice použitá při sestrojení stromu byla z důvodu téměř šedesáti procentní podobnosti nastavena na hodnotu PAM80.

Tabulka 4 Zástupci třídy savců

Identifikátor	Český název	Latinský název	Řád	Čeleď
P92475	Osel africký	Equus asinus	Lichokopytníci	Koňovití
Q6J6X5	Kůň domácí	Equus caballus	Lichokopytníci	Koňovití
A8DIM4	Velbloud jednohrbý	Camelus dromedarius	Sudokopytníci	Velbloudovití
Q1HKI1	Pes domácí	Canis familiaris	Šelmy	Psovití
P48900	Kočka domácí	Felis catus	Šelmy	Kočkovití
Q9T9Z0	Gorilla nížinná	Gorilla gorilla	Primáti	Hominidi
Q9T9X9	Orangutan bornejský	Pongo pygmaeus	Primáti	Hominidi
P41304	Vačice americká	Didelphis virginiana	Vačice	Vačicovití
O78705	Klokan rudý	Macropus rufus	Dvojitozubci	Klokanovití
O78704	Klokan dama	Macropus eugenii	Dvojitozubci	Klokanovití



Obrázek 8 Fylogenetický strom savců

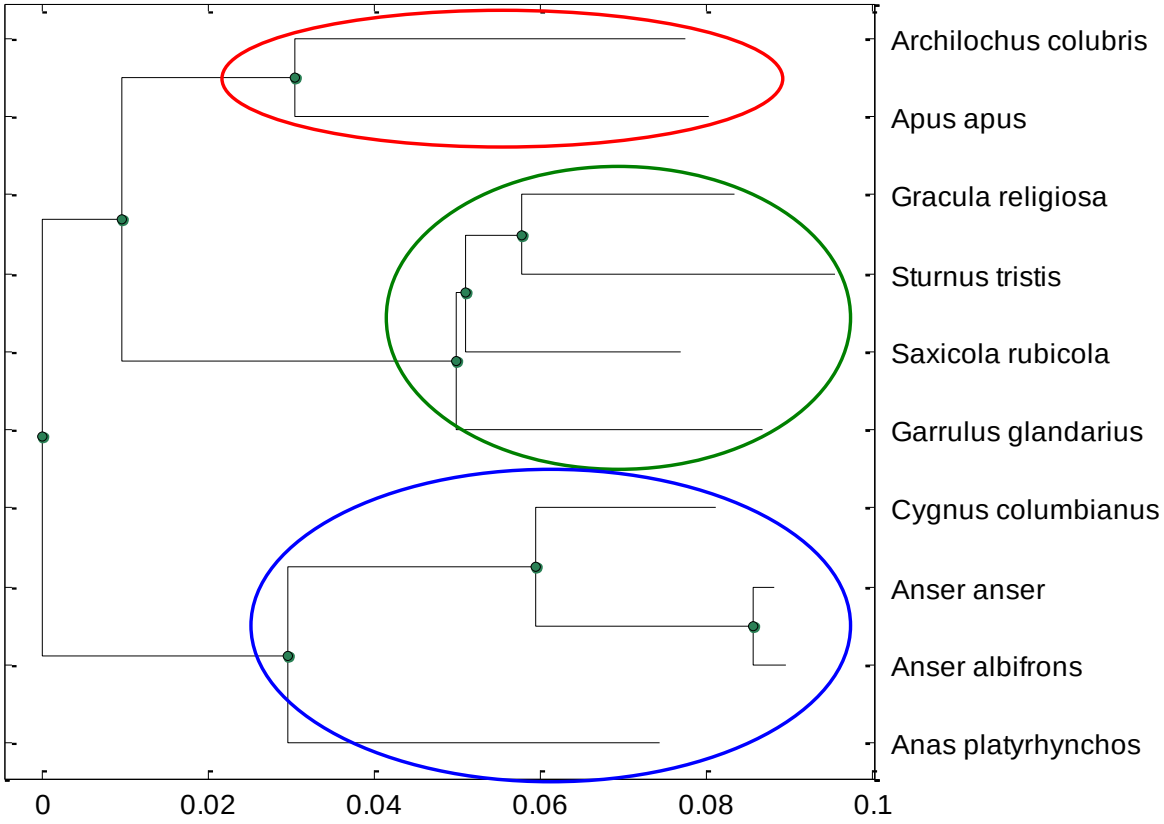
Tento fylogenetický strom by se dal rozdělit do třech shluků, které odpovídají molekulární klasifikaci. První – červený, největší shluk – reprezentuje placentálové živočichy patřící do větve *Boreoeutheria*, skupiny č. IV, do které jsou zařazeny řády sudokopytníků (*Artiodactyla*), lichokopytníci (*Perissodactyla*) i šelmy (*Carnivora*). Ve druhém shluku (zelené označení) jsou zástupci spadající do větve Boreoeutheria, skupiny č. III, do které patří řád s názvem primáti (*Primates*). Třetí, modrý shluk představuje vačnatce (*Marsupialia*), kteří tvoří samostatný nadřád.

3.2 Ptáci – Aves

Zvolení zástupci třídy ptáků obsahují 325 aminokyselin. Tato skupina dosahuje až 75 % podobnosti s 243 identickými pozicemi. Také distanční vzdálenost v této skupině je velmi malá – maximální distanční vzdálenost odpovídá hodnotě 0,1. Následuje tabulka se zástupci třídy ptáků (Tabulka 5) a fylogenetický strom sestrojený za použití matice PAM60 z důvodu vysoké podobnosti jednotlivých zástupců (Obrázek 9).

Tabulka 5 Zástupci třídy ptáků

Identifikátor	Český název	Latinský název	Řád	Čeleď
Q8HN25	Rorýs obecný	Apus apus	Svišťouni	Rorýsovití
A8RM84	Kolibřík rubínohrdlý	Archilochus colubris	Svišťouni	Kolibříkovití
C0L5Z4	Brambořníček černohlavý	Saxicola rubicola	Pěvci	Drozdovití
F8V312	Sojka obecná	Garrulus glandarius	Pěvci	Krkavcovití
G1D7D6	Loskuták posvátný	Gracula religiosa	Pěvci	Špačkovití
F1DTC7	Majna obecná	Sturnus tristis	Pěvci	Špačkovití
A6ZIZ9	Kachna divoká	Anas platyrhynchos	Vrubozubí	Kachnovití
B5M425	Husa velká	Anser anser	Vrubozubí	Kachnovití
Q85U60	Husa běločelá	Anser albifrons	Vrubozubí	Kachnovití
Q2QFS4	Labuť malá	Cygnus columbianus	Vrubozubí	Kachnovití



Obrázek 9 Fylogenetický strom ptáků

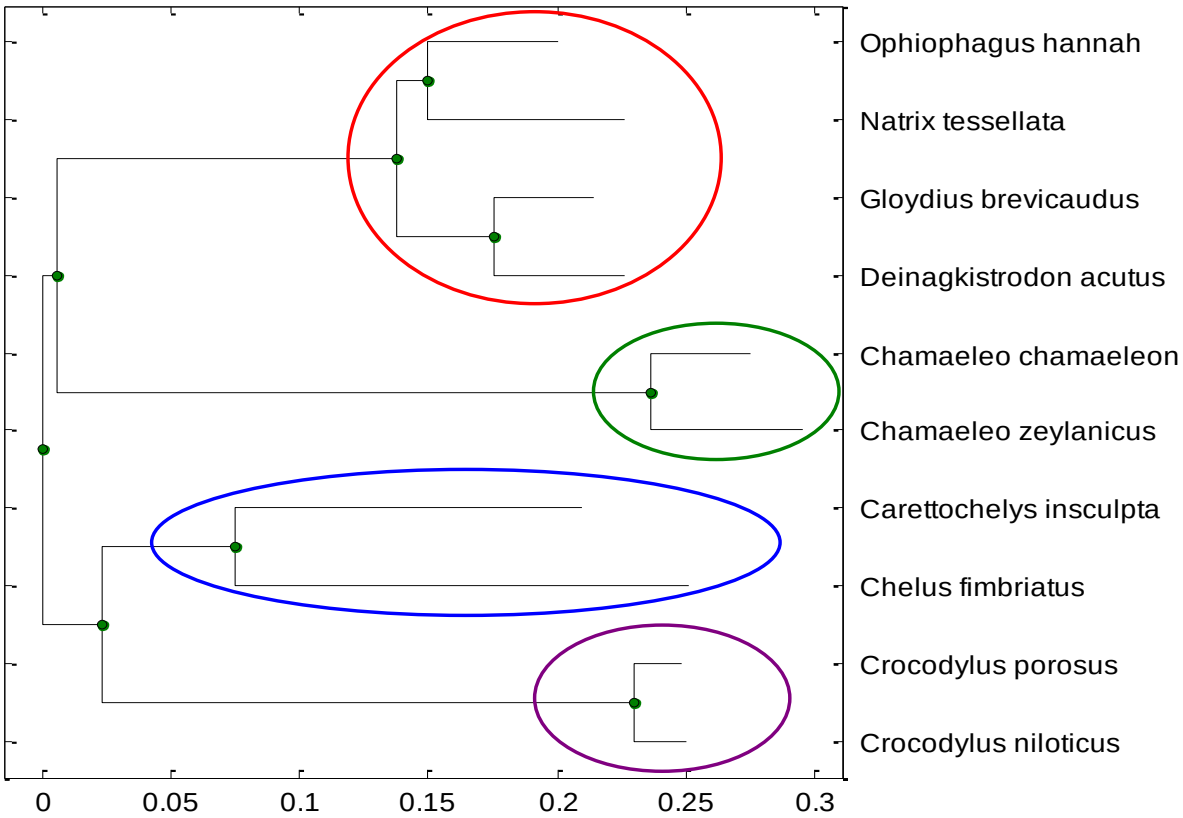
Ze sestrojeného fylogenetického stromu jsou jasně patrné tři shluky, které představují jednotlivé řády. Červeně zvýrazněný shluk představuje řád svišťovi (*Apodiformes*), zeleně zvýrazněný je řád pěvců (*Passeriformes*), následuje modře zvýrazněný řád vrubozubých ptáků (*Anseriformes*).

3.3 Plazi – Reptilia

Sekvence zástupců třídy plazů obsahují 320 aminokyselin, které jsou identické na 136 pozicích, což odpovídá podobnosti okolo 42 %. Distanční hodnota je maximálně ve výši 0,3. Tabulka 6 představuje vybrané jedince a fylogenetický strom sestrojený za použití matice PAM100.

Tabulka 6 Zástupci třídy plazů

Identifikátor	Český název	Latinský název	Řád	Čeleď
A0EQ81	Krokodýl mořský	Crocodylus porosus	Krokodýli	Krokodýlovití
F5CAJ1	Krokodýl nilský	Crocodylus niloticus	Krokodýli	Krokodýlovití
B7S6I0	Chameleon obecný	Chamaeleo chamaeleon	Šupinatí	Chameleonovití
B7S6A2	Chameleon zeylanicus	Chamaeleo zeylanicus	Šupinatí	Chameleonovití
B6DA95	Gloydius brevicaudus	Gloydius brevicaudus	Šupinatí	Zmijovití
Q5FZ03	Užovka podplamatá	Natrix tessellata	Šupinatí	Užovkovití
A9X4C6	Zmije sharp	Deinagkistrodon acutus	Šupinatí	Zmijovití
Q8WA20	Kobra královská	Ophiophagus hanna	Šupinatí	Korálovcovití
D5FW22	Karetka novoguinejská	Carettochelys insculpta	Želvy	Karetkovití
G0XMN1	Matamata třásnitá	Chelus fimbriatus	Želvy	Matamatovití



Obrázek 10 Fylogenetický strom třídy plazů

Tento fylogenetický strom (Obrázek 10) byl rozdělen do dvou velkých shluků, které se dají dále dělit. První velký shluk obsahuje řád šupinatých (*Squamata*) s podřády hady (*Serpentes*), které jsou červeně označeni, a ještěry (*Sauria*), v této práci zastoupeni chameleony, které jsou zeleně označeni.

Druhý velký shluk tvoří dva nezávislé řády – modře označený je řád želv (*Testudines*), fialově označený je řád krokodýlů (*Crocodilia*).

3.4 Ryby – Osteichthyes

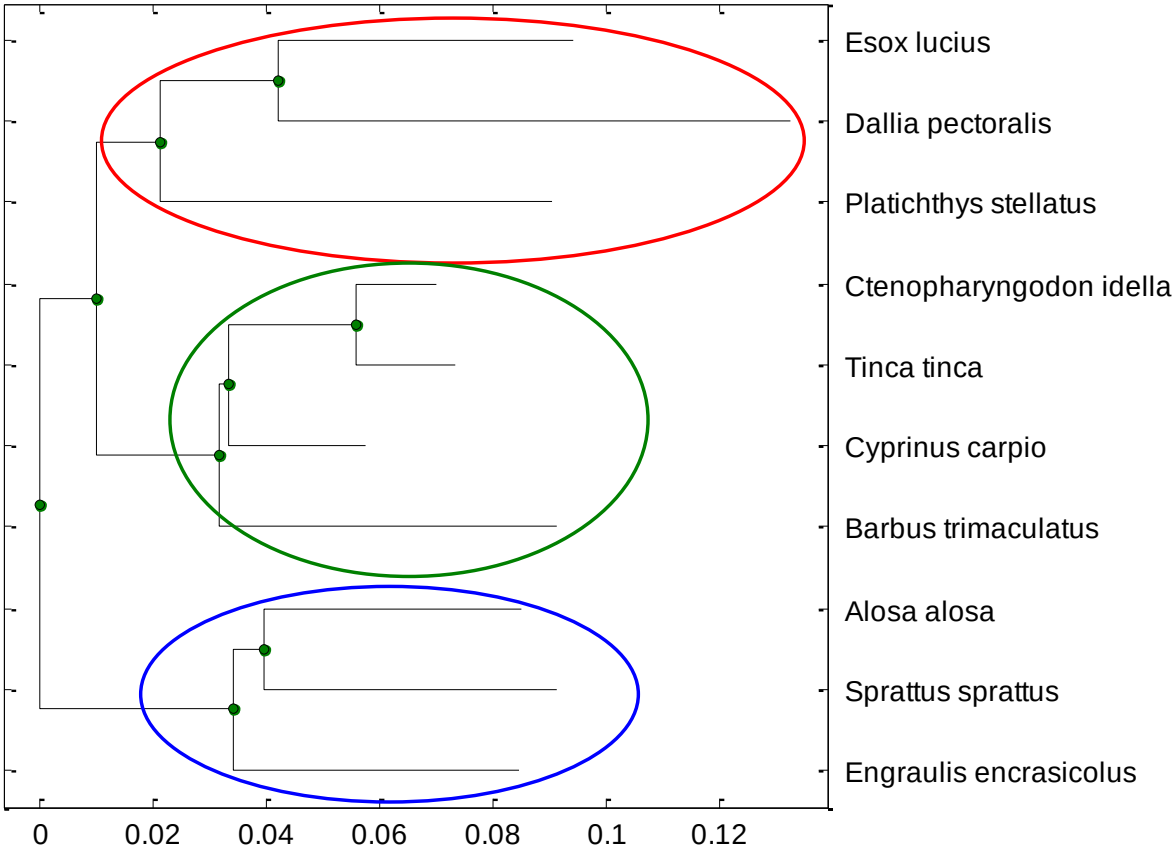
Označení ryb (*Osteichthyes*) je ve vědecké klasifikaci považováno za nadtřídu, která obsahuje dvě třídy: paprskoploutví (*Actinopterygii*) a nozdratí (*Sarcopterygii*). Druhá jmenovaná třída je označována za živé fosílie, ze kterých se pravděpodobně vyvinuli obojživelníci, a obsahuje pouze osm zástupců, které do této práce nebyly zahrnuty.

Tito zástupci mají sekvence dlouhé 324 aminokyselin. U této skupiny je stejně jako u savců vysoká shoda – přibližně 70 %, což odpovídá 227 identickým pozicím. Maximální distanční hodnota má hodnotu 0,13.

Následuje tabulka 7 s analyzovanými zástupci ryb a fylogenetický strom, který byl z důvodu vysoké podobnosti sestrojen pomocí matice PAM70.

Tabulka 7 Zástupci třídy ryb

Identifikátor	Český název	Latinský název	Řád	Čeleď
B3F0J1	Platýs bradavičnatý	Platichthys stellatus	Platýsi	Platýskovití
Q85D72	Dálie aljašská	Dallia pectoralis	Štikotvární	Dálie
Q85D59	Štika obecná	Esox lucius	Štikotvární	Štikovití
B0LXC7	Amur bílý	Ctenopharyngodon idella	Máloostní	Kaprovití
Q14F93	Kapr obecný	Cyprinus carpio	Máloostní	Kaprovití
A0ZRE3	Parma obecná	Barbus trimaculatus	Máloostní	Kaprovití
A0ZPS5	Lín obecný	Tinca tinca	Máloostní	Kaprovití
Q15FR6	Placka pomořanská	Alosa Alosa	Bezostní	Sleďovití
A5PIU0	Sleď obecný	Sprattus sprattus	Bezostní	Sleďovití
A5PID4	Sardel obecná	Engraulis encrasicolus	Bezostní	Sardelovití



Obrázek 11 Fylogenetický strom ryb

Do analýzy byly zahrnuty čtyři řády ryb, které vytváří tři shluky (Obrázek 11). Červeně je znázorněn první, který obsahuje řád platýsů (*Pleuronectiformes*) a řád štikotvárných ryb (*Esociformes*). Následuje zeleně znázorněná skupina řádu maloostných (*Cypriniformes*), které v této práci reprezentuje čeleď kaprovitých (*Cyprinidae*), a modře znázorněný řád bezostných (*Clupeiformes*), do které se řadí čeleď sledí (*Clupeidae*) a sardelí (*Engraulidae*).

3.5 Hmyz – Insecta

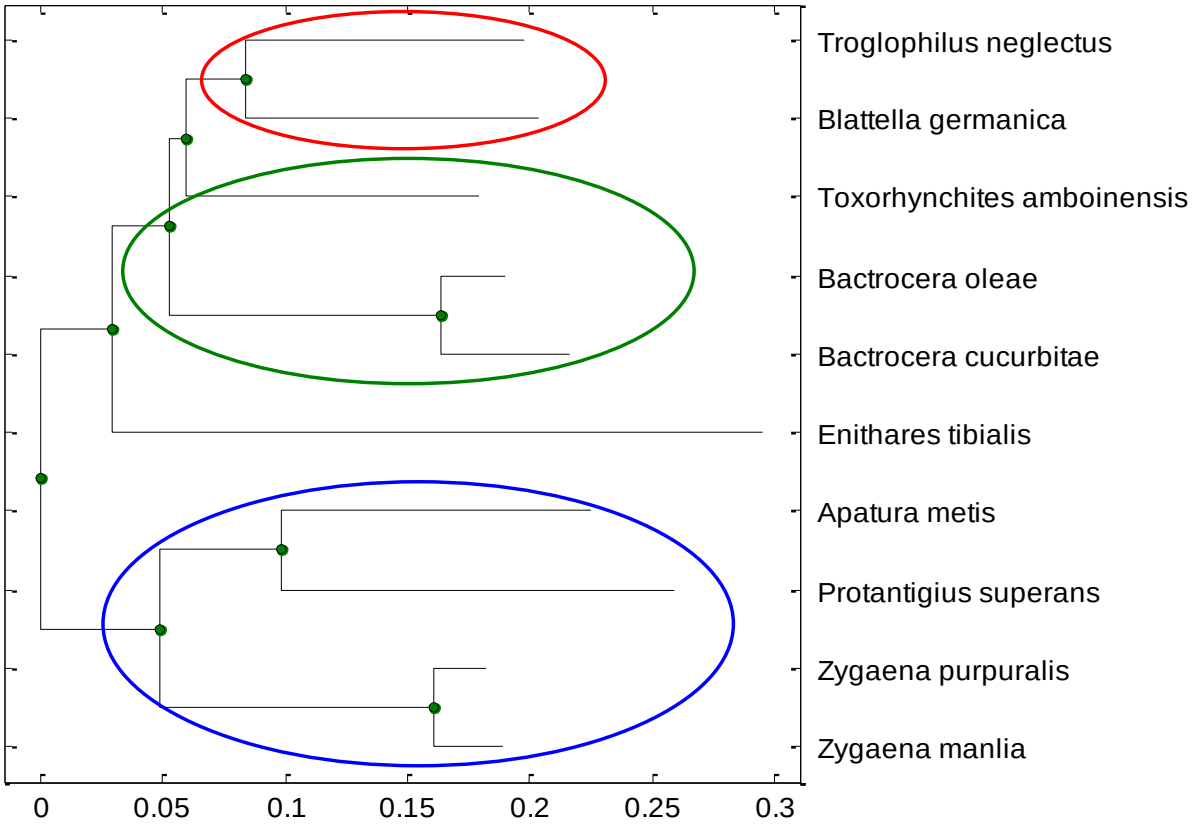
Hmyz je nejvíce různorodá a nejpočetnější skupina živočichů na světě, která zahrnuje více než milión popsaných druhů, je však odhadováno, že dalších 5 – 9 miliónů ještě nebylo objeveno.

Podle vědecké kvalifikace se hmyz řadí do říše živočichů (*Animalia*), kmene členovců (*Arthropoda*), podkmene šestinozí (*Hexapoda*), třídy hmyzu (*Insecta*) a dále se dělí na podtřídu bezkřídlí (*Apterygota*) a podtřídu křídlatí (*Pterygota*) do které spadá naprostá většina dnes žijícího hmyzu.

V této diplomové práci bylo zvoleno deset zástupců rod řadících se do pěti řádů. Vzhledem k různorodosti této skupiny není překvapivé, že shoda v proteinových sekvencích panuje pouze na 130 pozicích z 313, což představuje podobnost pouze 41 %. Distanční vzdálenost dosahuje hodnoty 0,3, která je také vyšší, jak u předchozích tříd. Z tohoto důvodu bylo nutné zvolit vyšší matici PAM než u výše sestrojených fylogenetických stromů (konkrétně matici PAM 100).

Tabulka 8 Zástupci třídy hmyzu

Identifikátor	Český název	Latinský název	Řád	Čeleď
G3M8W6	Vrtule okurková	Bactrocera cucurbitae	Dvoukřídlí	Octomilkovití
E3SV14	Vrtule olivovníková	Bactrocera oleae	Dvoukřídlí	Octomilkovití
F1BCF4	Komár Toxorhynchites	Toxorhynchites amboinensis	Dvoukřídlí	Komárovití
A8WXJ8	Vřeteluška purpurová	Zygaena purpuralis	Motýli	Vřeteluškovití
F6MF24	Motýl batolec	Apatura metis	Motýli	Babočkovití
A8WXG1	Vřeteluška mangová	Zygaena manlia	Motýli	Vřeteluškovití
G3CUR1	Protantigius superans	Protantigius superans	Motýli	Můrovití
C5HIV1	Splešťule blátivá	Enithares tibialis	Polokřídlí	Splešťulovití
B6DEF1	Cvrček ruský	Troglophilus neglectus	Rovnokřídlí	Kobylky
C6F3W0	Rus domácí	Blattella germanica	Švábi	Blarrellidae



Obrázek 12 Fylogenetický strom hmyzu

Zástupci řádu rovnokřídlých (*Orthoptera*) a švábů (*Blattodea*) tvoří společně jeden shluk, který je na obrázku 12 červeně znázorněn. Následuje zeleně znázorněný shluk dvoukřídlých (*Diptera*) a modře znázorněný shluk motýlů (*Lepidoptera*). Bokem stojí zástupce řádu polokřídlých (*Hemiptera*).

3.6 Fylogenetický strom

Následující fylogenetický strom (Obrázek 13) je sestrojen z padesáti proteinových sekvencí, které byly výše představeny. Délka zarovnaných sekvencí je 325 aminokyselin, ve kterých je 76 identických pozic, z čehož lze odvodit, že celková shoda je pouze 23 %. Toto nízké číslo shody vedlo k použití vysoké skórovací matice – použita byla matice PAM350.

Při základní analýze fylogenetického stromu jsou patrné dva velké shluky, které jsou reprezentovány dvěma kmeny živočichů – kmenem členovců (*Arthropoda*) a kmenem strunatců (*Chordata*). První zmíněný obsahuje pouze třídu hmyzu (*Insecta*) - červeně znázorněn. Je patrné, že tato třída dosahuje nejvyšších hodnot distancí – je tedy od ostatních druhů nejvíce vzdálená. Druhý shluk je rozmanitější – v této práci je zastoupen čtyřmi třídami. Modře znázorněna je třída plazů (*Reptilia*), následována je zeleně znázorněnou třídou ptáků (*Aves*), žlutě zakreslenou třídou savců (*Mammalia*) a poslední je fialová nadtřída ryb (*Osteichthyes*) reprezentovaná třídou paprskoploutvých (*Actinopterygii*).

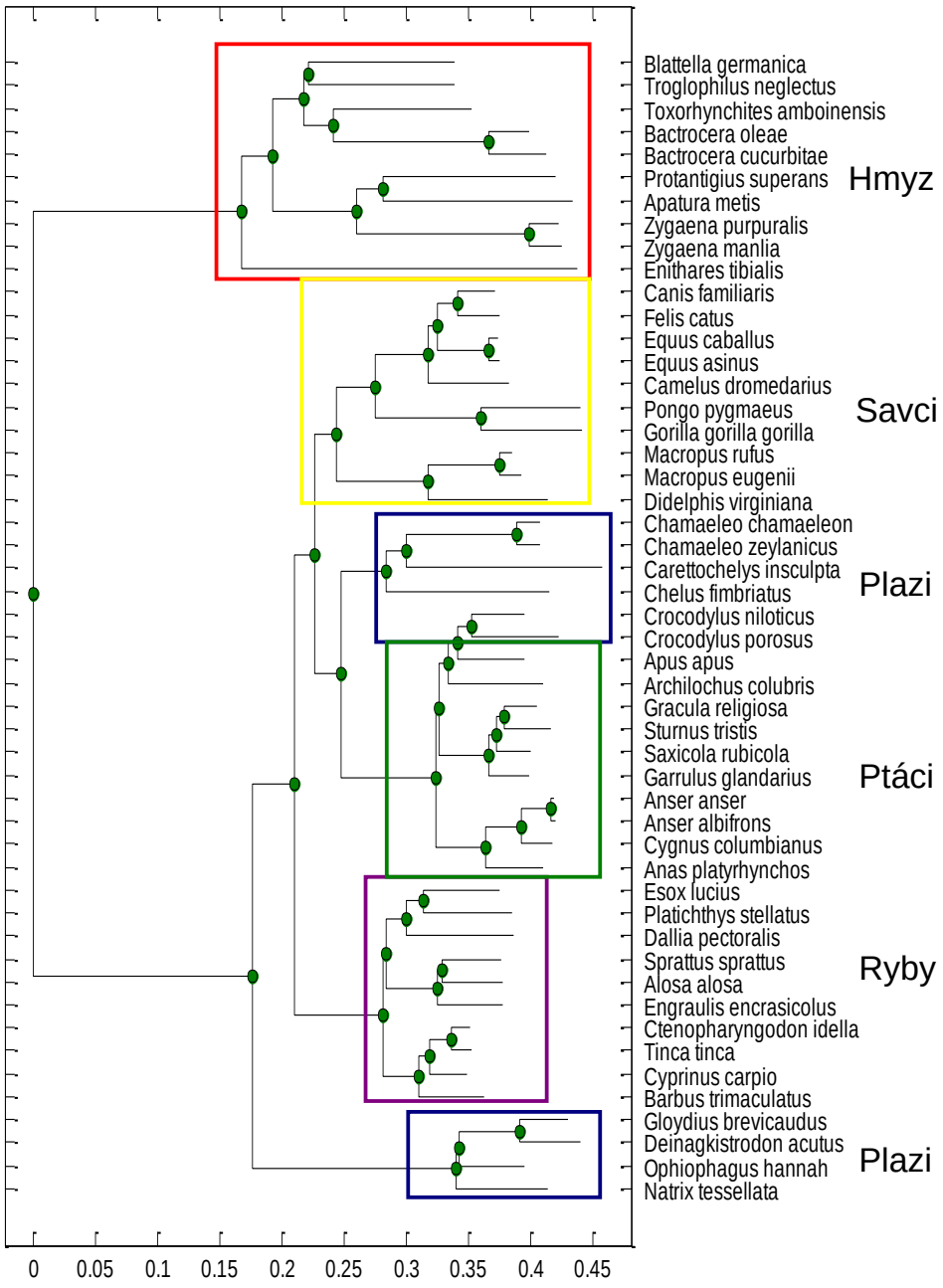
Jak již bylo řečeno v úvodu této podkapitoly, pro sestavení stromu bylo nutné použít vysokou hodnotu skórovací matice – PAM350, která je několikanásobně vyšší, jak skórovací matice použité při sestrojování fylogenetických stromů pro jednotlivé třídy. Pro přehlednost byla vytvořena přehledná tabulka – Tabulka 9.

Tabulka 9 Souhrnné informace o živočišných zástupcích

	Délka sekvence	Počet shodných pozic	Procentuální shoda	Skórovací matice
Savci	318	188	60%	PAM80
Ptáci	325	243	75%	PAM60
Plazi	320	136	42%	PAM100
Ryby	324	227	70%	PAM70
Hmyz	313	130	41%	PAM100
Komplet	325	76	23%	PAM250

Tato skórovací matice přináší do analýzy nepřesnosti. Jedné takové zásadní si lze všimnout u modře znázorněné třídy plazů a zeleně znázorněné třídy ptáků. Větve řádu krokodýlů (zástupci *Crocodylus porosus* a *Crocodylus niloticus*) byly chybně zařazeny do tří ptáků, zástupci hadů tvoří zcela oddělený shluk v blízkosti ryb. Celý shluk plazů se tedy rozpadl na tři malé shluky.

Tuto nepřesnost by měly fylogenetické superstromy eliminovat.



Obrázek 13 Klasický fylogenetický strom

4 Realizace metody průměrného konsensu

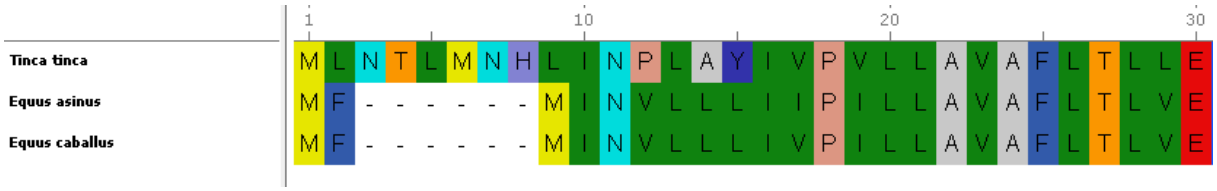
Následující kapitola představí principielní řešení problematiky superstromů, kterou je možné rozdělit do třech částí. První část je věnována obecné práci se sekvencemi – zarovnávání, výpočet evolučních vzdáleností a tvorbě fylogenetických stromů. Druhá část je zaměřena na přípravnou práci pro vytváření superstromů, která obsahuje výpočty chybějících distančních hodnot třemi různými metodami. Poslední, třetí, část shrnuje dosažené výsledky do podoby fylogenetického superstromu.

4.1 Fylogenetické stromy

Základním krokem pro tvorbu superstromů bylo vytvoření databáze sekvencí ve formátu FASTA, které jsou podrobně představeny v kapitole číslo tři. Ze získané sady bylo vybráno pět zástupců, které přiblíží funkce programu. Konkrétně se jedná o tři zástupce ze třídy savců (*Canis familiaris* – pes domácí, *Felis catus* – kočka domácí a *Gorilla gorilla* – Gorila nížinná), jeden ze třídy ptáků (*Gracula religiosa* - loskuták posvátný) a jeden plaz (*Natrix tessellata* – užovka podplamatá). Pro nejpřehlednější popis funkce metody jsou tyto sekvence rozděleny do dvou databází, které mají tři sekvence společné (představovány třídou savců) a pouze jednu odlišnou (pták, resp. plaz).

Oba soubory jsou do programu načteny a z důvodu různých délek zarovnány. K tomuto účelu byla vytvořena funkce pojmenovaná `parditmat`. Účelem této funkce je globálně párově zarovnat sekvence pomocí proporcionální distance a metody Jukes-Cantora. Na výstup je zaslána zarovnaná sekvence společně s vypočítanými hodnotami distancí.

V Matlabu slouží ke globálnímu zarovnávání příkaz `multialign`, který nebyl pro použitá data vhodný, jelikož mezi jednotlivými zástupci napříč třídami není vysoká podobnost. Vytvořená funkce je při zarovnávání šetrnější.



Obrázek 14 Zarovnávání sekvencí

Obrázek 14 vystihuje problematiku zarovnávání sekvencí různorodých organismů. Zde jsou dva zástupci lichokopytníků (*Equus asinus* a *Equus caballus*) zarovnány s línem obecným (*Tinca tinca*). Zatímco mezi lichokopytníky došlo pouze k jedné substituci na pozici sedmnáct, lín obecný se liší na třinácti pozicích z třiceti.

V následujícím kroku byly spočítány distanční hodnoty mezi sekvencemi pomocí příkazu *seqpdist*. U tohoto příkazu lze volit několik variant, proto bude probrán podrobněji.

```
dist=seqpdist(S1,'Method','jukes-cantor','Indels',
'score','Scoringmatrix', smat,'PairwiseAlignment', true,
'GapOpen',3, 'ExtendGap',3);
```

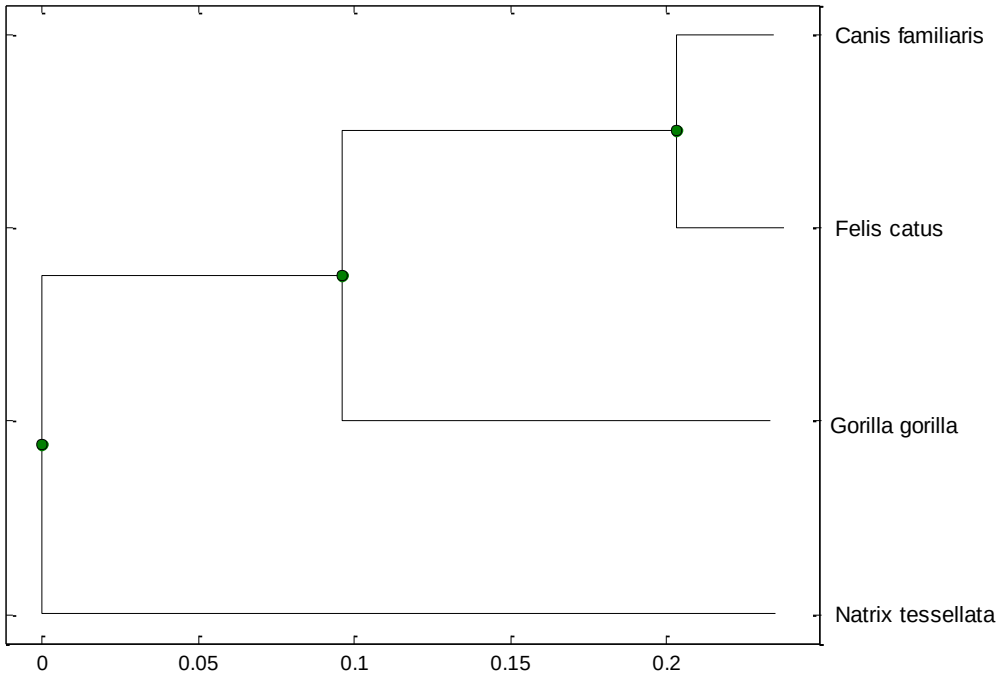
Do proměnné *S1* je uložena načtená sekvence. *Jukes-Cantor* je metoda zvolená pro výpočet distančních vzdáleností mezi dvěma sekvencemi, *indels* je řetězec, který určuje, jakým způsobem bude skript pracovat s mezerami – výchozí hodnota je nastavena na skóre – jedná se tedy o penalizaci mezer. Následuje volba skórovací matice pro globální zarovnání. Tuto hodnotu je nutné volit na základě předběžné analýzy nebo podle vlastních zkušeností uživatele. Pokud jsou si sekvence velmi podobné, volí se hodnota matice PAM nízká. Pokud jsou sekvence velmi vzdálené (obsahují malý počet shodných aminokyselin), volí se hodnota matice PAM vysoká. Pro sestrojování fylogenetických superstromů byla vytvořena funkce, která určuje hodnotu PAM podle zjištěné průměrné distance. Byly vymezeny intervaly a k nim přiřazené patřičné hodnoty PAM. Výstup této funkce je označen *smat* (zkratka skórovací matice).

Příkaz *PairwiseAlignment* kontroluje globální zarovnání uvnitř sekvence, hodnota *true* byla zvolena z důvodu nestejně dlouhých sekvencí. Poslední dva příkazy slouží k penalizaci mezer.

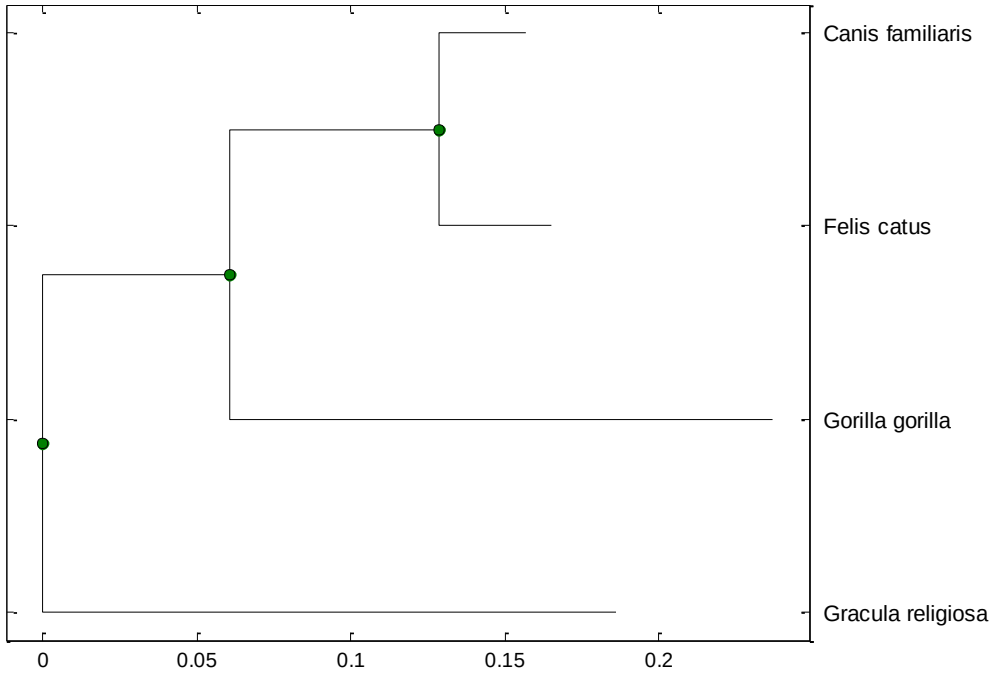
Po těchto úpravách již máme možnost sestrojit fylogenetický strom. K tomuto účelu slouží příkaz *seqneighjoin*, který spočítá fylogenetický strom z proměnné *dist* pomocí metody nejbližšího souseda.

```
tree = seqneighjoin(dist,'firstorder',S1)
plot(tree)
```

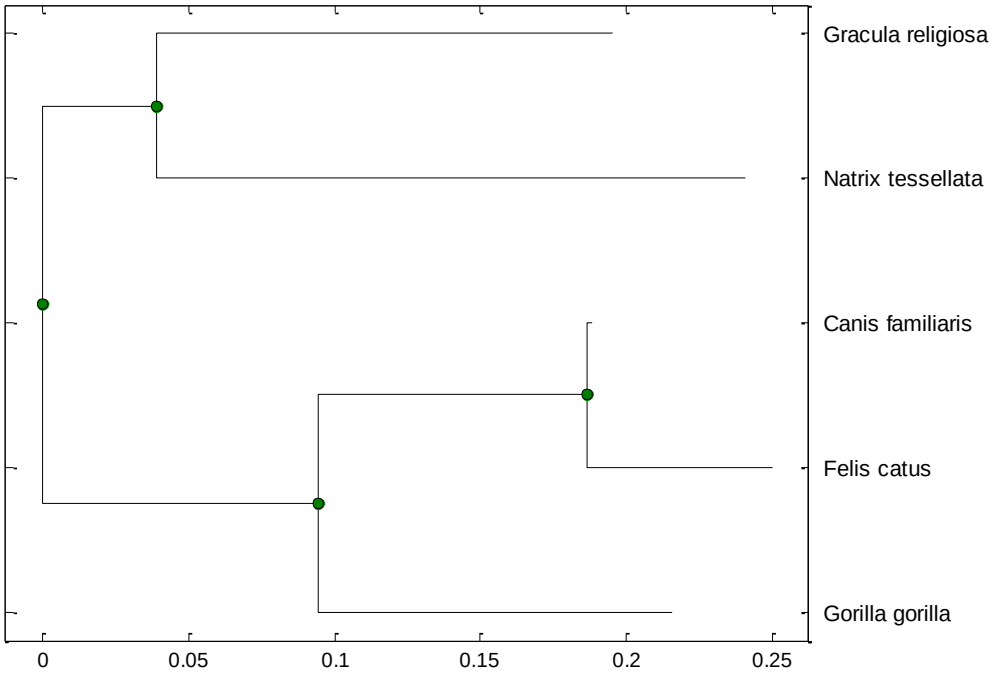
Výsledné fylogenetické stromy jsou zobrazeny na následujících obrázcích (15, 16), ze kterých je patrné, že *Canis familiaris* je nejbližší s *Felis catus*. Interpretovat by bylo možné také shluk, který v obou případech správně vytvořila třída savců.



Obrázek 15 První zdrojový strom



Obrázek 16 Druhý zdrojový strom



Obrázek 17 Klasický fylogenetický strom

Fylogenetický strom je sestrojen ze všech použitých sekvencí (Obrázek 17). Klasická fylogenetická analýza chybně vytvořila v dendrogramu dva shluky – správně spojila všechny zástupce savců do jednoho shluku. Druhý shluk tvoří zástupci plazů (*Natrix tessellata*) a ptáků (*Gracula religiosa*), což neodpovídá evolučním znalostem. Tato chyba může být způsobena nesprávným zarovnáním sekvencí nebo chybně nastavenou maticí skórovací maticí PAM.

4.2 Výpočet fylogenetického superstromu

Druhá část je již zaměřena na přípravu tvorby superstromu. Z posloupnosti distancí je nutné vytvořit čtvercovou distanční matici s nulovými prvky na diagonále. Výstupem pro naše sekvence jsou hodnoty uvedené v tabulce 10 a 11.

Z těchto dvou distančních matic chceme vytvořit výslednou distanční matici pro všechny sekvence. Je nutné rozlišit názvy sekvencí a v programovém prostředí Matlab je správně indexovat. Základními příkazy pro spojování a porovnávání řetězců jsou příkazy `strcat` a `strcmp`.

V podstatě mohou při tvoření výslední distanční matice nastat následující tři případy:

1. Daná sekvence je obsažena v obou souborech, jsou známy dvě distanční hodnoty. V tomto případě dané hodnoty distancí zprůměrujeme prostým aritmetickým

průměrem. Do výsledné matice je na danou pozici zapsána právě tato hodnota. Tento případ nastává např. u *Felis catus* a *Gorilla gorilla*.

2. Sekvence je obsažena pouze v jednom souboru, tedy i jen v jedné distanční matici. V tomto případě je do výsledné hodnoty zapsána hodnota sekvence, která je známa. Jako příklad lze uvést sekvence *Canis familiaris* a *Natrix tessellata*, případně *Canis familiaris* a *Gracula religiosa*.
3. Třetí možností je neznalost ani jedné hodnoty distance, jelikož se jedná o dvě odlišné sekvence. V tomto případě hodnotu v matici označíme např. písmenem X. S touto pozicí bude dále pracováno. Toto nastává u *Natrix tessellata* a *Gracula religion*. Výsledná distanční matice pro všechny sekvence je zapsána v tabulce 12.

Tabulka 10 Distanční matice pro první set sekvencí

	Canis familiaris	Felis catus	Gorilla gorilla	Natrix tessellata
Canis familiaris	0			
Felis catus	0,0651	0		
Gorilla gorilla	0,2795	0,4657	0	
Natrix tessellata	0,2753	0,4763	0,4688	0

Tabulka 11 Distanční matice pro druhý set sekvencí

	Canis familiaris	Felis catus	Gorilla gorilla	Gracula religiosa
Canis familiaris	0			
Felis catus	0,0651	0		
Gorilla gorilla	0,2795	0,3361	0	
Gracula religiosa	0,2753	0,3569	0,4233	0

Tabulka 12 Výsledná distanční matice

	Canis familiaris	Felis catus	Gorilla gorilla	Natrix tessellata	Gracula religiosa
Canis familiaris	0				
Felis catus	0,0651	0			
Gorilla gorilla	0,2795	0,4009	0		
Natrix tessellata	0,2753	0,4763	0,4688	0	
Gracula religiosa	0,2753	0,3569	0,4233	X	0

Nyní je nutné dopočítat neznámou distanci, která je označena fiktivní hodnotou X. V praxi se k dopočítání neznámé distance používají dvě metody – ultrametrická metoda a aditivní metoda, které byly popsány v podkapitole 2.2.2. Obě metody v prvním kroku vyhledávají počet neznámých distancí (v tomto vzorovém případě je neznámá pouze jedna hodnota distance). V této práci je navíc přidána třetí metoda – dopočítávání neznámých pozic aritmetickým průměrem.

4.2.1 Ultrametrická metoda

Neznámá distance je získána ze svého okolí známých hodnot distancí jako maximální nalezená hodnota v řádku a sloupci, ve kterém se nachází hledaná distance. V cyklu jsou takto postupně získány všechny neznámé distance, které jsou doplněny do výsledné matice, ze které lze konstruovat fylogenetický superstrom.

V prvním kroku je nutné ověřit nerovnost distancí, ze kterých je získána hledaná hodnota. Pokud je podmínka splněna, ve for cyklu ohraničeným velikostí matice (*x_konec1*, *y_konec1*) jsou postupně vyhledávány a doplňovány výsledné hodnoty. Pomocná proměnná pom představuje počet sdílených taxonů. V tomto ukázkovém případě jsou sdílené tři sekvence, nastavena je *pom = 3*.

```
% 1) Ultrametrická metoda – doplnění hodnot distancí

if matice(i,:) == matice(:,j)
     disp('Hledanou distanci nelze dohledat – nesplňuje
               podmínku ultraparametrické metody')

elseif

     for i = 1 : x_konec1
          for j = 1 : y_konec1
               if matice(i,j) == X
                  matice(i,j) = max(max(matice(i,1:pom))…
                          (max(matice(1:pom,j))));
               end
          end
     end
end
```

V tomto konkrétním případě byla nalezena jedna neznámá distance na pozici [4,4]. Výpočtem pomocí ultrametrické metody dostáváme hodnotu 0,4233, jak je vidět z tabulky 13. Hodnoty, ze kterých je neznámá distance dopočítávána, jsou znázorněny zeleně.

Tabulka 13 Dopočítaná distanční matice ultrametrickou metodou

	Canis familiaris	Felis catus	Gorilla gorilla	Natrix tessellata	Gracula religiosa
Canis familiaris	0				
Felis catus	0,0651	0			
Gorilla gorilla	0,2795	0,4009	0		
Natrix tessellata	0,2753	0,4763	0,4688	0	
Gracula religiosa	0,2753	0,3569	0,4233	0,4233	0

4.2.2 Aditivní metoda

Aditivní metoda dopočítává neznámou distanci *d(i,j)* jak z taxonů, ve kterých se nachází neznámá distance, tak z taxonů v těsné blízkosti. Na výstup je zaslán maximální nalezený součet zmenšený o hodnotu distance, ve kterém se taxony protnou. V tomto případě nejsou hodnoty ve vertikální poloze známy, jelikož neznámá distance je umístěna v pravém dolním rohu matice. Více informací o této metodě je uvedeno v podkapitole 2.2.2.

V prvním kroku opět nutné zjistit počet chybějících hodnot a zkontrolovat podmínku rozdílných hodnot součtů.

```
 % 2) Aditivní metoda

if (matice(i,1:pom) +(matice(i-pom:i-1,j-1))') ==
   (matice(i-1,1:pom) + (matice(i-pom:i-1,j))')
     disp('Hledanou distanci nelze dohledat - nesplňuje
          podmínku aditivní metody')
elseif
    for i = 1 : x_konec1
         for j = 1 : y_konec1
              if matice(i,j) == X

MAX1 = (max(matice(i,1:pom) +(matice(i-pom : i-1,j-1))'));
MAX2 = max(matice(i-1,1:pom) + (matice(i-pom:i-1,j))');
matice(i,j)=max(MAX1,MAX2)-matice(i-1,j);

              end
         end
    end
end
```

Hodnota neznámé distance je 0,4084.

Tabulka 14 Dopočítaná distanční matice aditivní metodou

	Canis familiaris	Felis catus	Gorilla gorilla	Natrix tessellata	Gracula religiosa
Canis familiaris	0				
Felis catus	0,0651	0			
Gorilla gorilla	0,2795	0,4009	0		
Natrix tessellata	0,2753	0,4763	0,4688	0	
Gracula religiosa	0,2753	0,3569	0,4233	0,4084	0

4.2.3 Aritmetický průměr

Metoda aritmetického průměru je obdobou aditivní metody. Metoda vychází z maximálního součtu hodnot známých distancí, který je podělen hodnotou, která odpovídá počtu sdílených sekvencí. Průměrují se hodnoty v horizontální i vertikální poloze. V tomto případě se průměruje pouze v horizontální poloze, jelikož hodnota neznámé distance je umístěna v pravém horním rohu matice.

```
%  3) Aritmetický průměr okolí

     for i = 1 : x_konec1
          for j = 1 : y_konec1
               if matice(i,j) == 65

                    matice(i,j) = (max(matice(i,1:pom))+
                                  ((matice(i-pom))'))/pom;
               end
          end
      end
```

Dopočítaná distance má hodnotu 0,3518.

Tabulka 15 Dopočítaná distanční matice aritmetickým průměrem okolí

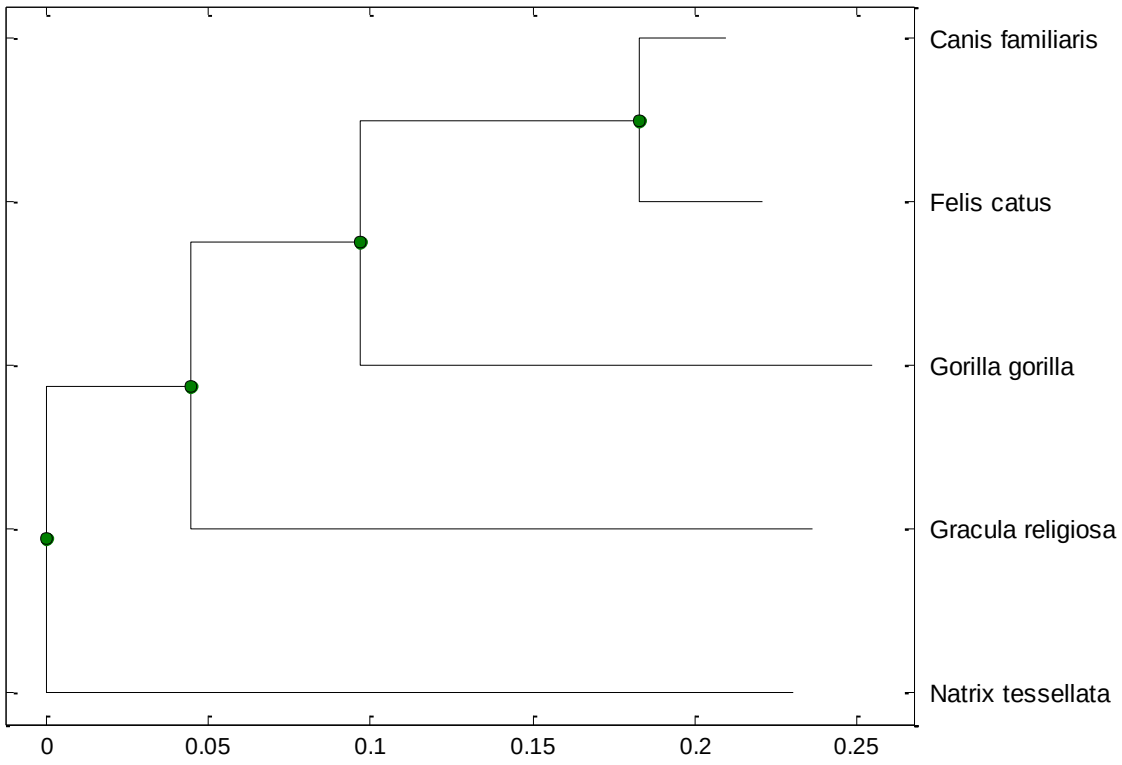
	Canis familiaris	Felis catus	Gorilla gorilla	Natrix tessellata	Gracula religiosa
Canis familiaris	0				
Felis catus	0,0651	0			
Gorilla gorilla	0,2795	0,4009	0		
Natrix tessellata	0,2753	0,4763	0,4688	0	
Gracula religiosa	0,2753	0,3569	0,4233	0,3518	0

Dopočítávání chybějících hodnot má nespornou výhodou v odpadnutí podmínky nestejné hodnoty distancí. Tato metoda je však velmi obecná a zjednodušující. U malých souborů sekvencí by to ve výpočtu nemělo představovat velkou chybu. Problém by mohl nastat u objemově velkých souborů sekvencí se vzdálenými druhy, kdy je počet neznámých hodnot distancí veliký. Při práci s velkým objemem dat by do výpočtu zanášela velké zobecnění.

4.3 Tvorba fylogenetického superstromu

Následuje zobrazení fylogenetického superstromu. Jelikož ve zkušební matici byla pouze jedna neznámá, vycházejí všechny tyto stromy velmi podobné. Obrázek 18 reprezentuje fylogenetický superstrom ve kterém byly hodnoty dopočítávány pomocí ultrametrické metody.

Z fylogenetických superstromu je patrné, že byl odstraněn druhý shluk, který vypočítala klasická fylogenetická analýza. První shluk obsahující savce zůstal zachován. Tento strom není v rozporu s tradičními hypotézami fylogenetických vztahů.

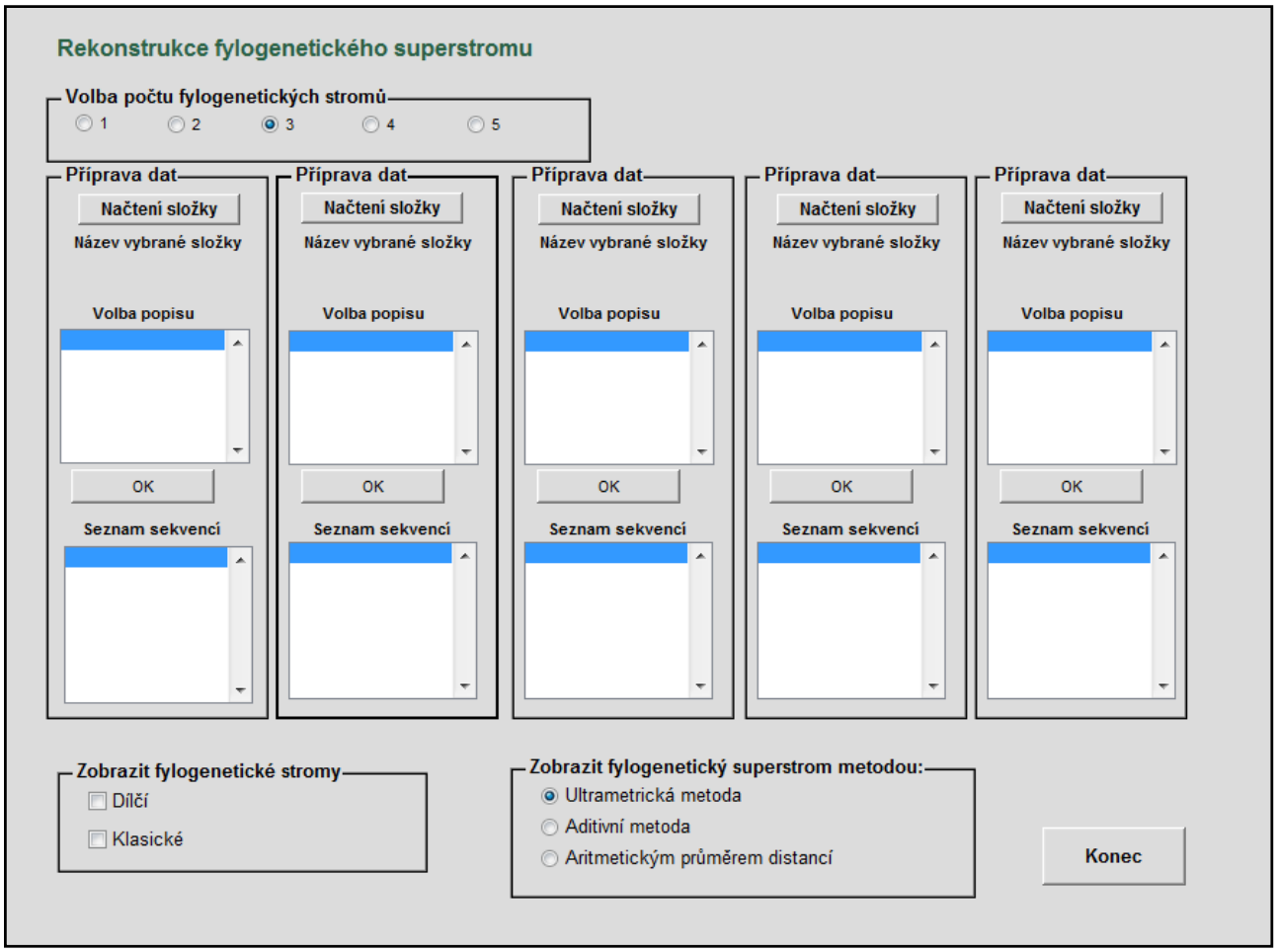


Obrázek 18 Ukázkový fylogenetický superstrom - metoda ultrametrická

5 Softwarová aplikace pro konstrukci superstromu

Následující kapitola je věnována popisu vytvořeného uživatelského interface v Matlab GUI, které uživateli usnadňuje a zpřehledňuje sestrojování fylogenetických stromů. Rozdělit by se dalo na tři části – první obecně slouží k volbě základních parametrů sekvencí, druhá k sestrojování klasických fylogenetických stromů a třetí k sestrojování fylogenetických superstromů. Pro ukončení programu bylo vytvořeno tlačítko „Konec“. Při zmáčknutí se výpočty ukončí a hlavní ovládací panel se uzavře.

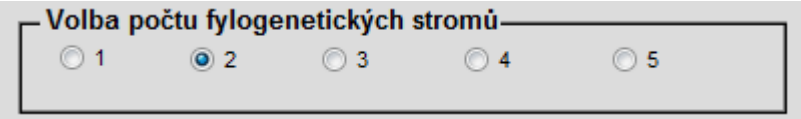
Základní schéma uživatelského prostředí je znázorněno na obrázku 19.



Obrázek 19 Grafické rozhraní programu – hlavní ovládací panel

5.1 Příprava dat

Přípravná fáze je nejrozsáhlejší částí. Postupně je nutné zvolit základní parametry pro analýzu. V prvním kroku přípravy dat má uživatel volbu počtu setů sekvencí v rozsahu 1-5, jak je znázorněno na obrázku 20. Podle zvolené hodnoty se základní okno upraví – nepotřebná okna v dalších krocích nejsou viditelná. Pro konstrukci fylogenetických superstromů je nutné zvolit minimálně dva sety. V opačném případě se uživateli objeví chybová hláška, ve které je uvedeno, že nezvolil dostatečný počet setů sekvencí.

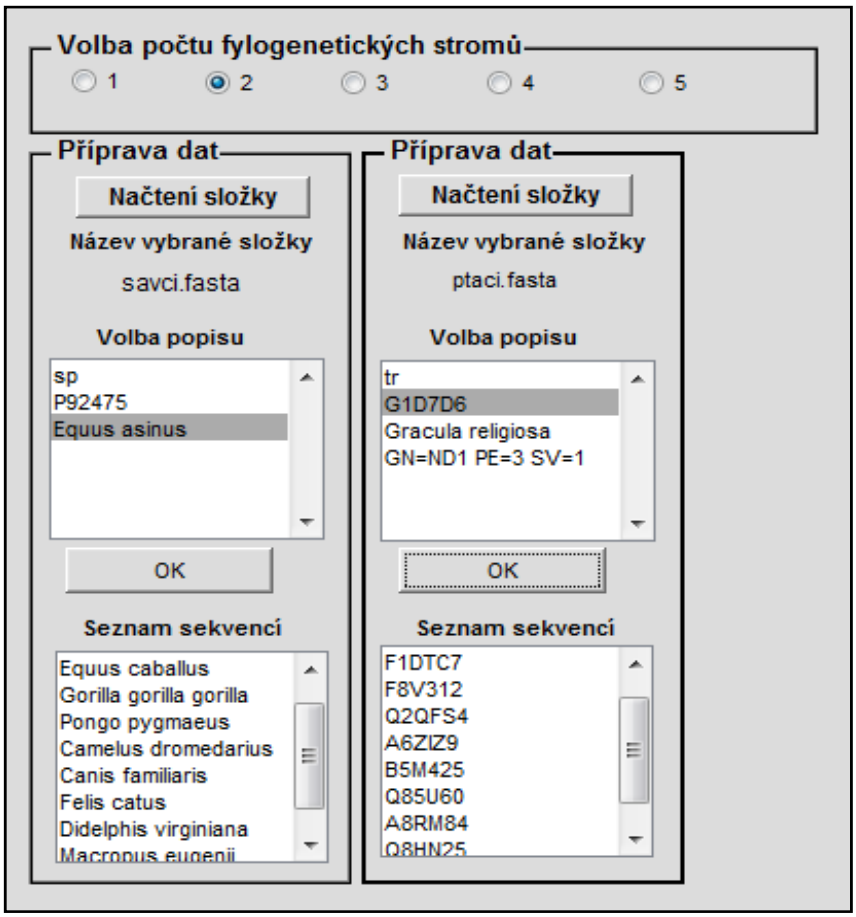


Obrázek 20 Grafické rozhraní – volba počtu zpracovávaných souborů sekvencí

Následuje načtení souboru. Po stisknutí tlačítka *Načtení složky* dojde k vyvolání dialogového okna pro výběr datových souborů se sekvencemi. Tyto soubory musí být ve formátu FASTA, který je vhodný pro zápis sekvenčních dat používaných v bioinformatice.

Název vybrané složky se v okně zobrazí pod již zmíněným tlačítkem *Načtení složky*. Název listů ve fylogramu není složen z celého názvu FASTA. Pro přehlednost zápisu je tento název rozdělen na jednotlivé popisy, které jsou vzájemně odděleny znakem „|“, případně jiným identifikátorem - v tomto případě písmeny „GN“, „PE“ za latinským názvem organismu.

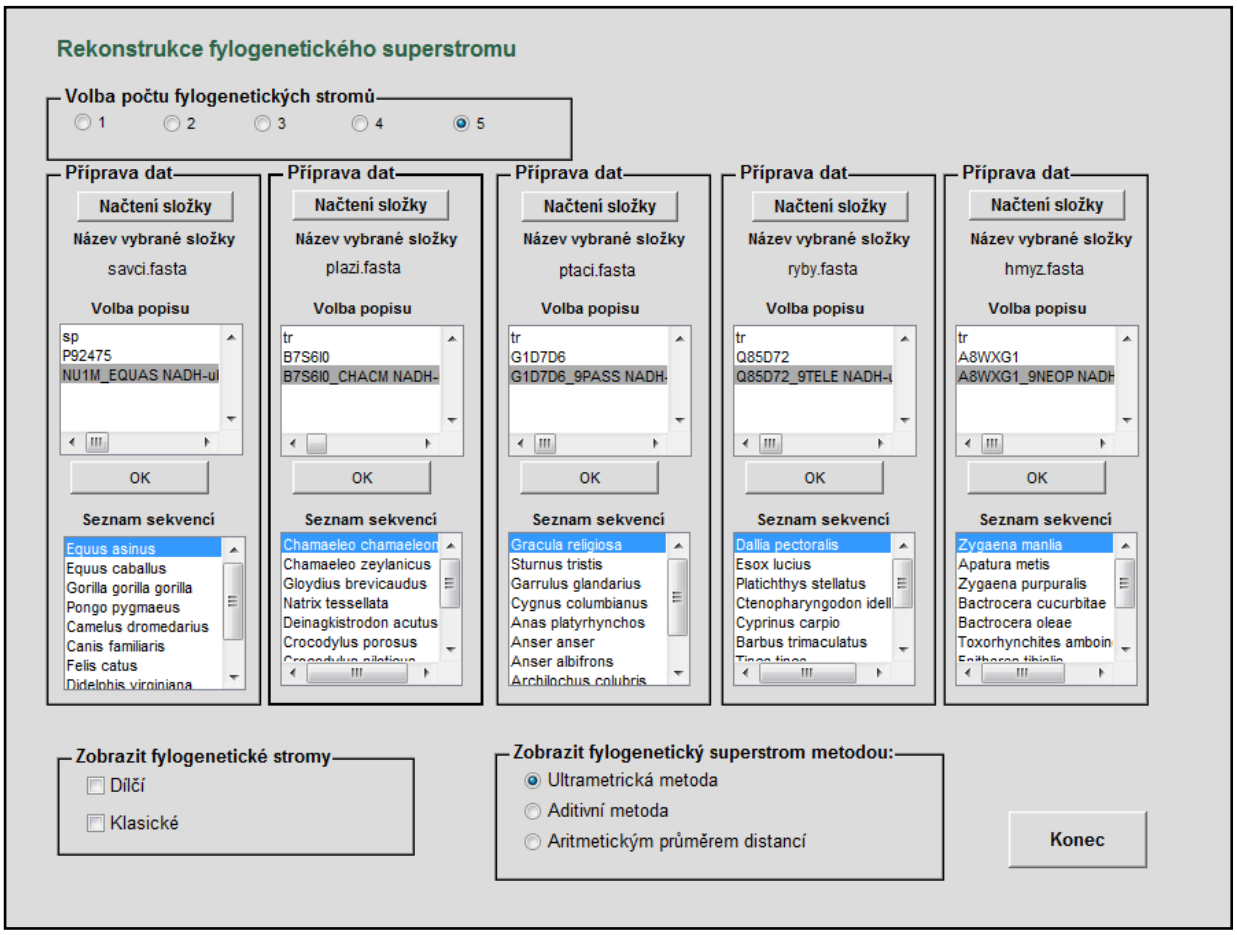
Jako nejpřehlednější se jeví pojmenování organismu jeho rodovým i druhovým jménem, případně jedinečným identifikátorem sekvence, který však neposkytuje žádné informace o organismu. Tato volba zápisu závisí na uživateli – výběr se provádí myší. Z důvodu přehledné orientace v sestrojených fylogenetických stromech je vhodné jednou zvolenou možnost zachovat pro všechny datové sety. Na obrázku 21 jsou znázorněny oba zápisy.



Obrázek 21 Grafické rozhraní – příprava dat

Po stisknutí tlačítka „OK“ se zobrazí seznam sekvencí obsažených v datovém setu. Tato diplomová práce pracuje s pěti soubory sekvencí, jak bylo uvedeno v předchozí kapitole, jedná se o sekvence proteinu ND1, který se nachází v mitochondriích. Bylo vybráno pět tříd živočichů – třída hmyzu, ryb, ptáků, plazů a savců. Základní okno po nahrání všech setů sekvencí je znázorněno na obrázku 22.

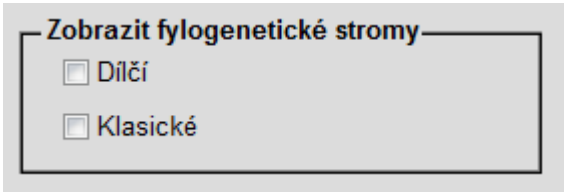
V prvním kroku byl nahrán soubor se sekvencemi savců, ve druhém plazů, třetí jsou ptáci, čtvrté ryby a poslední nahraný soubor je složka se zástupci hmyzu. Pro popis větví byla vybrána možnost popisu latinským názvem organismu.



Obrázek 22 Grafické rozhraní programu – načtená data

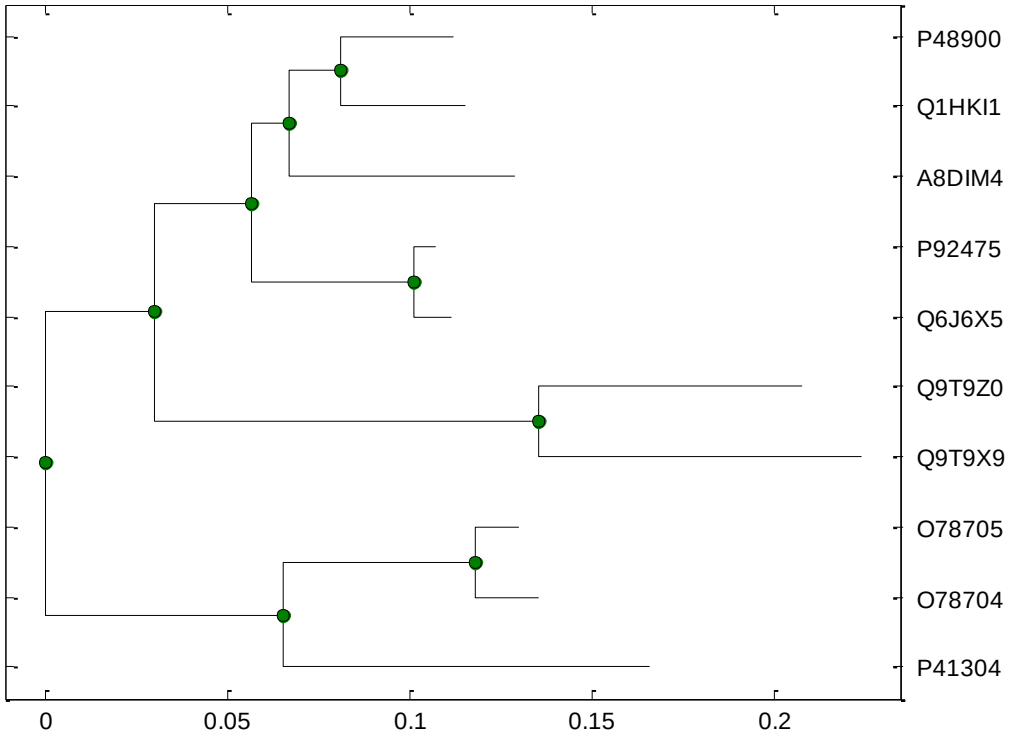
5.2 Vykreslení fylogenetických stromů

V dalším kroku se uživateli nabízí možnost sestrojení klasických fylogenetických stromů. Na obrázku 23 jsou patrné dvě možnosti – zobrazení dílčích fylogenetických stromů a zobrazení klasického fylogenetického stromu, který je sestrojen ze všech načtených sekvencí.

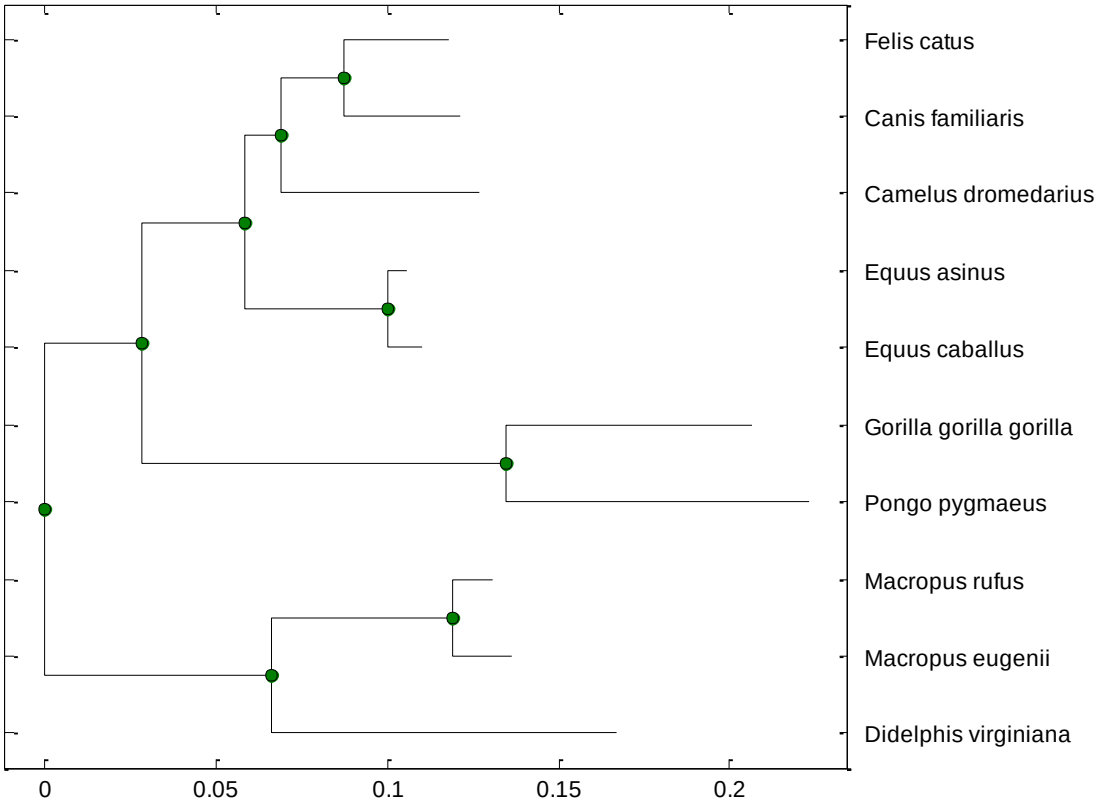


Obrázek 23 Grafické rozhraní programu – vykreslení fylogenetických stromů

Při zaškrtnutí možnosti dílčí fylogenetické stromy se vykreslí právě takový počet fylogenetických stromů, jako bylo nahráno sekvencí. Samozřejmostí je popis jednotlivých listů. Jako příklad vykreslení dílčích fylogenetických stromů je uveden obrázek 24, kde jsou listy pojmenované identifikačním číslem, a obrázek 25, kde jsou listy pojmenované latinskými názvy.

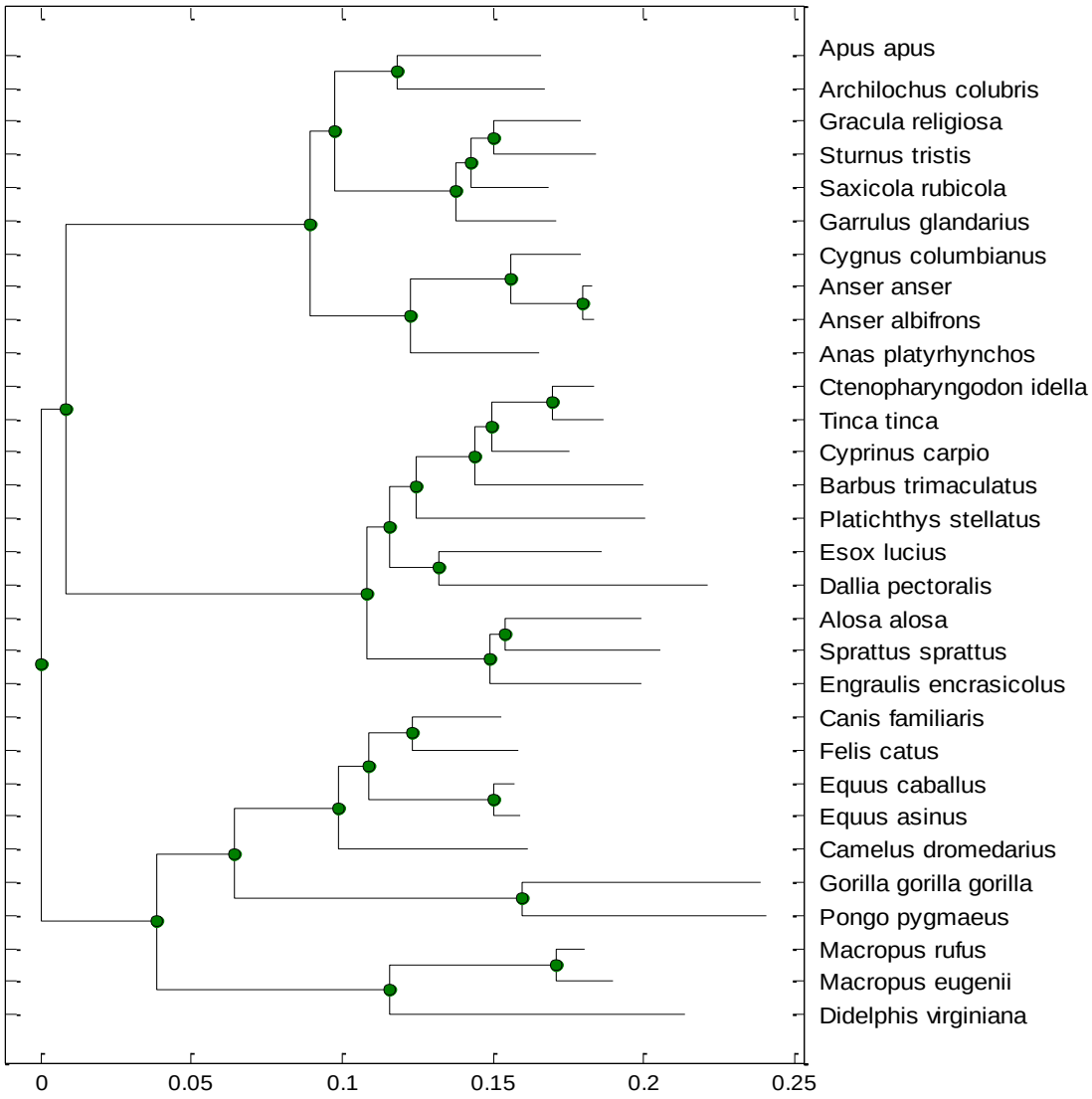


Obrázek 24 Ukázka klasického fylogenetického stromu savců – popis identifikátorem



Obrázek 25 Ukázka klasického fylogenetického stromu savců s popisem pomocí názvu

Pokud chce uživatel získat kompletní fylogenetický strom sestrojený ze všech načtených sekvencí, zobrazí se po zaškrtnutí možnosti „Klasické fylogenetické stromy“. V tomto případě jsou nahrané sekvence spojeny do jednoho souboru, ze kterého je fylogenetický strom rekonstruován. Matice PAM je nastavena na vyšší hodnotu než v předchozím případě (PAM350), jelikož není předpokládána vysoká podobnost mezi třídami živočichů. Na následujícím obrázku 26 je vykreslen klasický fylogenetický strom pro zástupce třídy savců, ryb a ptáků.



Obrázek 26 Ukázka fylogenetického superstromu tříd savců, ryb a ptáků

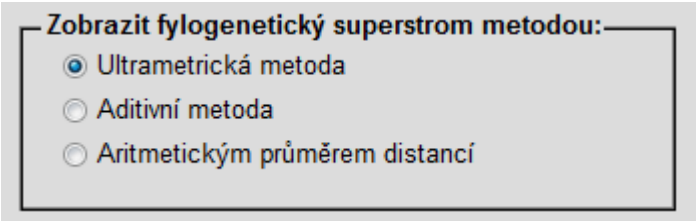
V obrázku jsou jasně rozpoznatelné tři shluky. První (vrchní) tvoří ryby, prostřední je tvořen zástupci ryb a spodní zástupci třídy savců.

5.3 Vykreslení fylogenetického superstromu

Třetí podkapitola je věnována samotné rekonstrukci fylogenetických superstromů. Tato metoda vychází z předpokladu sdílených sekvencí mezi jednotlivými datovými sety. V této diplomové práci jsou sdílené vždy první sekvence organismů obsažené v datových setech. Počet sdílených taxonů je roven počtu zpracovávaných setů. Pokud uživatel využije všech pět souborů, budou mít společných pět sekvencí. Aby bylo možné program univerzálně využít, byl zvolen náhodně vždy první zástupce ze souboru.

Konkrétně se jedná o kombinace zástupců *Zygaena manlia* (vřeteluška purpurová) ze třídy hmyzu, *Dallia pectoralis* (dálie aljašská) ze třídy ryb, *Chamaeleo chameleon* (Chameleon obecný) ze třídy plazů, *Gracula religion* (loskuták posvátný) ze třídy ptáků a jako poslední je zvolen *Equus asinus* (osel africký) ze třídy savců.

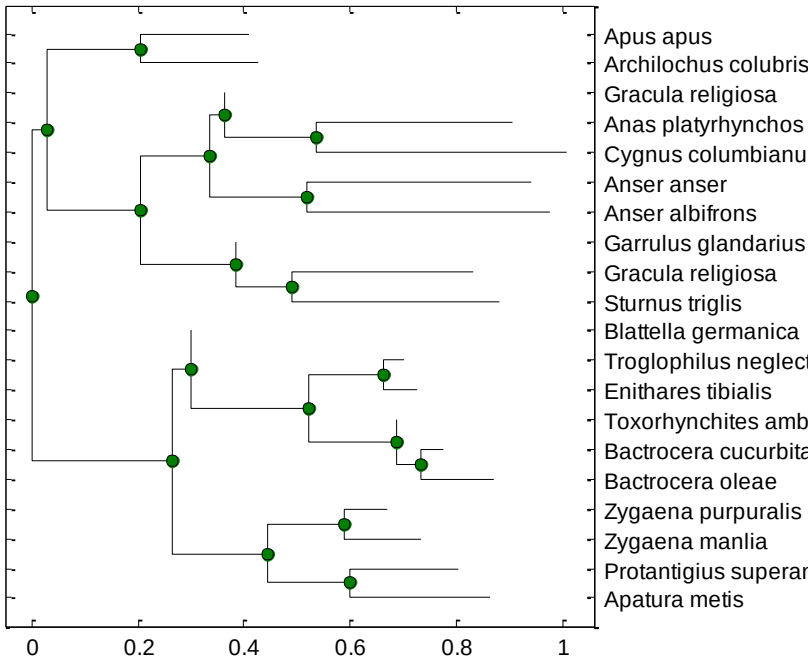
Po vytvoření nových databázových sekvencí bylo nutné dopočítat chybějící pozice distancí v matici, ze které má být fylogenetický superstrom konstruován. K tomu slouží tři analytické metody – ultrametrická, aditivní metoda a dopočítávání aritmetickým průměrem distancí, jak je uvedeno na obrázku 27.



Obrázek 27 Grafické rozhraní – volba metody dopočítání chybějících distancí

Vykreslení fylogenetického superstromu je opět závislé na počtu zvolených datových setů zvoleného v prvním kroku. Pokud jsou zvoleny dvě datové sady sekvencí po deseti sekvencích, je nutné ke každému setu přidat jednoho sdíleného zástupce – analýza tak probíhá na jedenácti sekvencích v každé datové sadě.

Pro názornou ukázku výstupu byla zvolena ultrametrická metoda počítaná na řádu ptáků a hmyzu (obrázek 28).



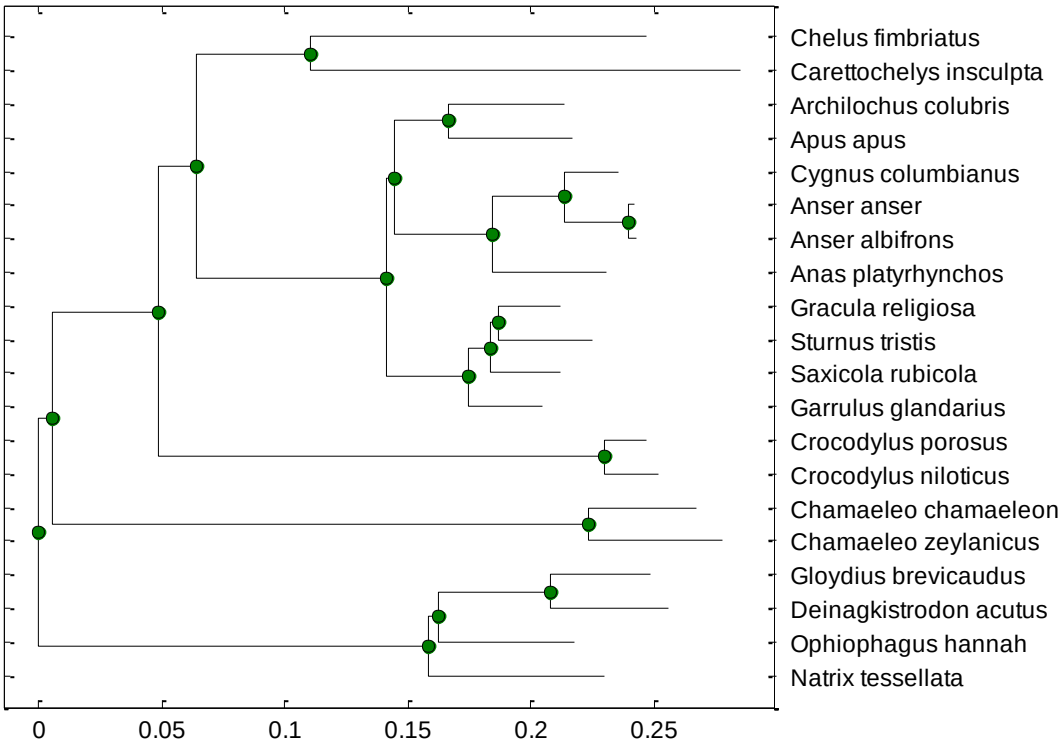
Obrázek 28 Fylogenetický superstrom ptáků a hmyzu

Rekonstrukce zachovala oddělené třídy. Vrchní shluk představuje třídu ptáků, spodní třídu hmyzu. Ve třídě plazů je jasně oddělitelný shluk svišťounů, kteří tvoří první dvě větve. Následují řády vrubozubích a pěvců. Třída hmyzu je reprezentována dvěma shluky, ze kterých vyčnívá *Blatella germanica* (rus domácí), jediný zástupce švábů v této diplomové práci. Tvoří shluk společně s dvoukřídlýma, polokřídlýma a rovnokřídlýma. Poslední shluk v tomto obrázku je tvořen třídou motýlů.

6 Diskuze

Následující kapitola je věnována reprezentaci dosažených výsledků rekonstrukce fylogenetických superstromů a jejich interpretaci. V podkapitole 3.6 byl sestrojen fylogenetický strom ze všech použitých sekvencí, který zaváděl chybu v podobě celkového rozpadnutí shluku plazů při rekonstrukci. Z tohoto důvodu bude diskuse prováděna právě na třídě plazů společně se třídou ptáků.

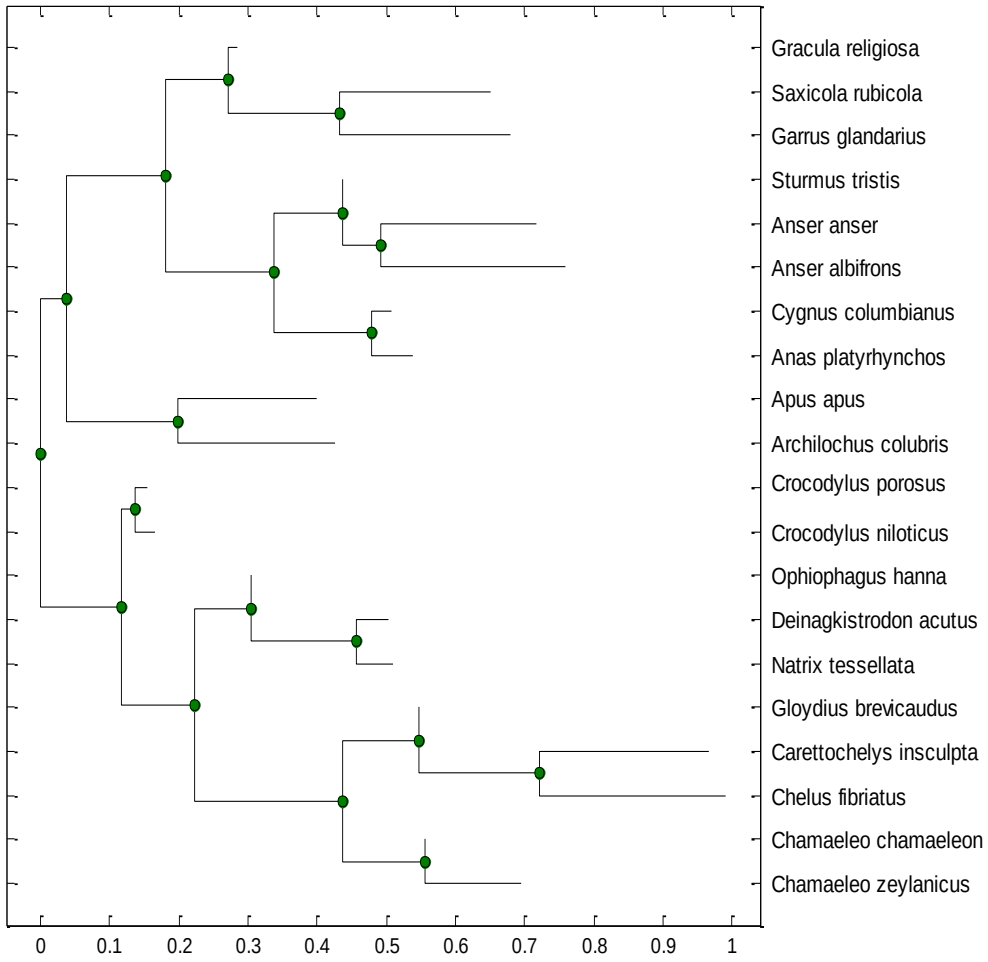
Výstupem obou analýz by měly být dvě jasně oddělitelné třídy a několik patrných řádů. Již sestrojený klasický fylogenetický strom naznačuje, jak nestabilní třída plazů je. Z obrázku 29 je patrné, že shluk plazů zůstal zachován, zatímco třída plazů se rozpadla do čtyř nezávislých shluků. Spodní tvoří řád šupinatých, ke kterému mají však patřit také zástupci chameleonů. Řád želv je chybně přiřazeny ke třídě ptáků. Na pomezí stojí řád krokodýlů.



Obrázek 29 Klasický fylogenetický strom ptáků a plazů

Při konstrukci fylogenetického superstromu jsou na výběr tři možnosti dopočítávání neznámých distancí. Jedná se o podrobně probrané metody v podkapitole 4.2 – ultrametrickou, aditivní a metodu průměrování. Výsledné matice hodnot distancí jsou uvedeny v příloze.

Na obrázku 30 je zobrazen fylogenetický strom, který neznámé dopočítával pomocí ultrametrické metody.



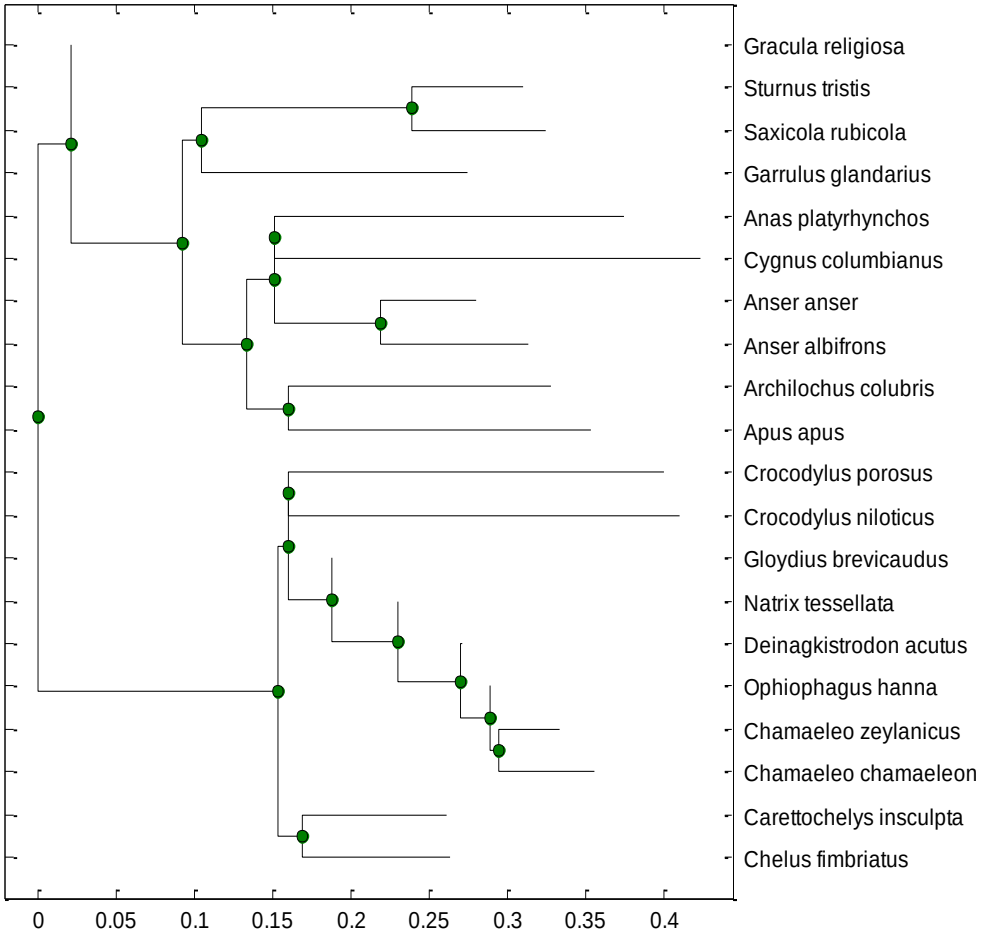
Obrázek 30 Fylogenetický superstrom třídy plazů a ptáků – ultrametrická metoda

Ultrametrická metoda obě třídy jasně oddělila, zachovány zůstaly také jednotlivé řády. Vrchní shluk v dendrogramu reprezentuje třída ptáků. Nejblíže k nim má řád krokodýlů, který je však správně zahrnut do shluku, který obsahuje třídu plazů.

Druhou možností sestrojení fylogenetického superstromu je použití aditivní metody (obrázek 31). Ta zdaleka nedosahuje takových výsledků, jako metoda ultrametrická. Z fylogenetického stromu lze vyčíst dvě třídy, bohužel však zařazení do jednotlivých řádů není nikterak přesné.

Prvním velkým nedostatkem aditivní metody je zařazení zástupce *Gracula religiosa* (loskuták posvátný), který tvoří oddělenou první větev. Stejně jako *Sturnus tristis*, *Saxicola rubeola* a *Garrulis glandarius* by měl tvořit jeden shluk (řád pěvců). Dalším nedostatkem je rozpadnutí shluku řádu šupinatých u plazů.

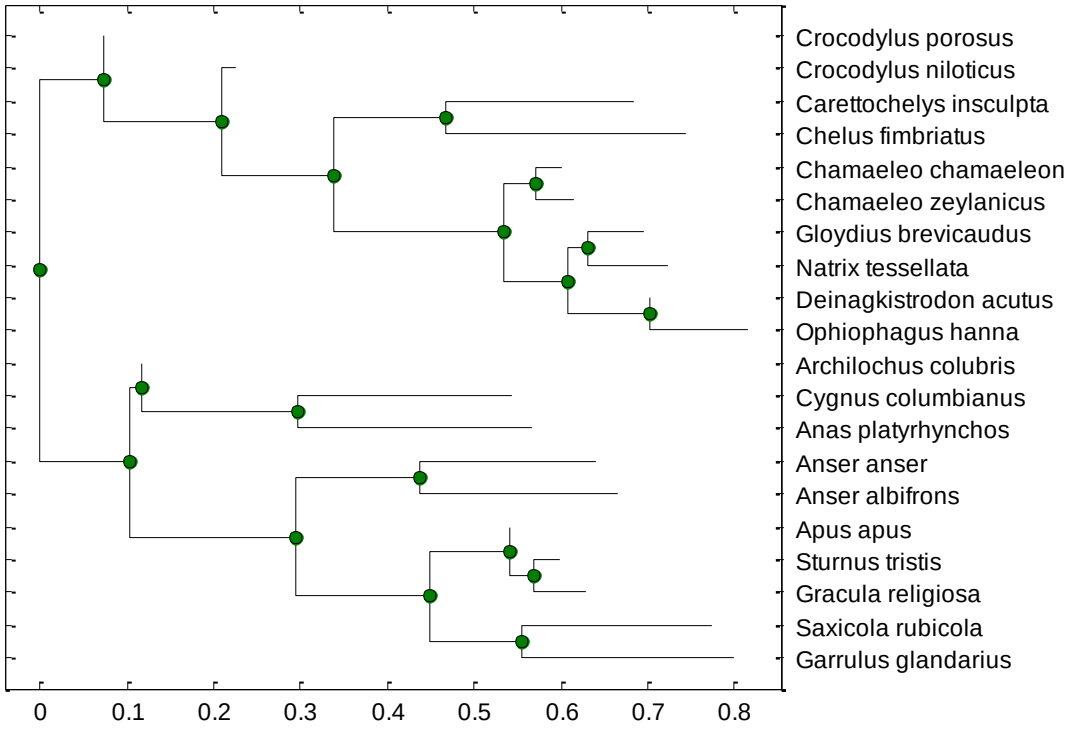
Chybu této metody lze spatřovat v menšiteli, který odečítáme od hodnoty maximálního součtu. Pokud je tato hodnota nízké číslo nebo příliš vysoké číslo, v analýze se objeví vysoká (resp. nízká) hodnota distance, která vnáší do analýzy odchylku.



Obrázek 31 Fylogenetický superstrom plazů a ptáků - aditivní metoda

Poslední možností zobrazení fylogenetických superstromů je metodou aritmetického průměrování (obrázek 32). Tato metoda oddělila oba shluky, nedošlo však k oddělení řádů. Zcela se rozpadl shluk krokodýlů u plazů. U ptáků se rozpadly všechny shluky reprezentující řády. Jedinou výhodou této metody je, že správně oddělila řád šupinatých, kam patří chameleoni i hadi.

Tato metoda dopočítávání chybějících distancí je příliš obecná. Pokud je neznámých distancí vysoký počet, mohla by vést k přílišnému zjednodušení a častému opakování určité hodnoty distance, která analýzu fylogenetických stromů zkreslí.



Obrázek 32 Fylogenetický superstrom plazů a ptáků - aritmetický průměr

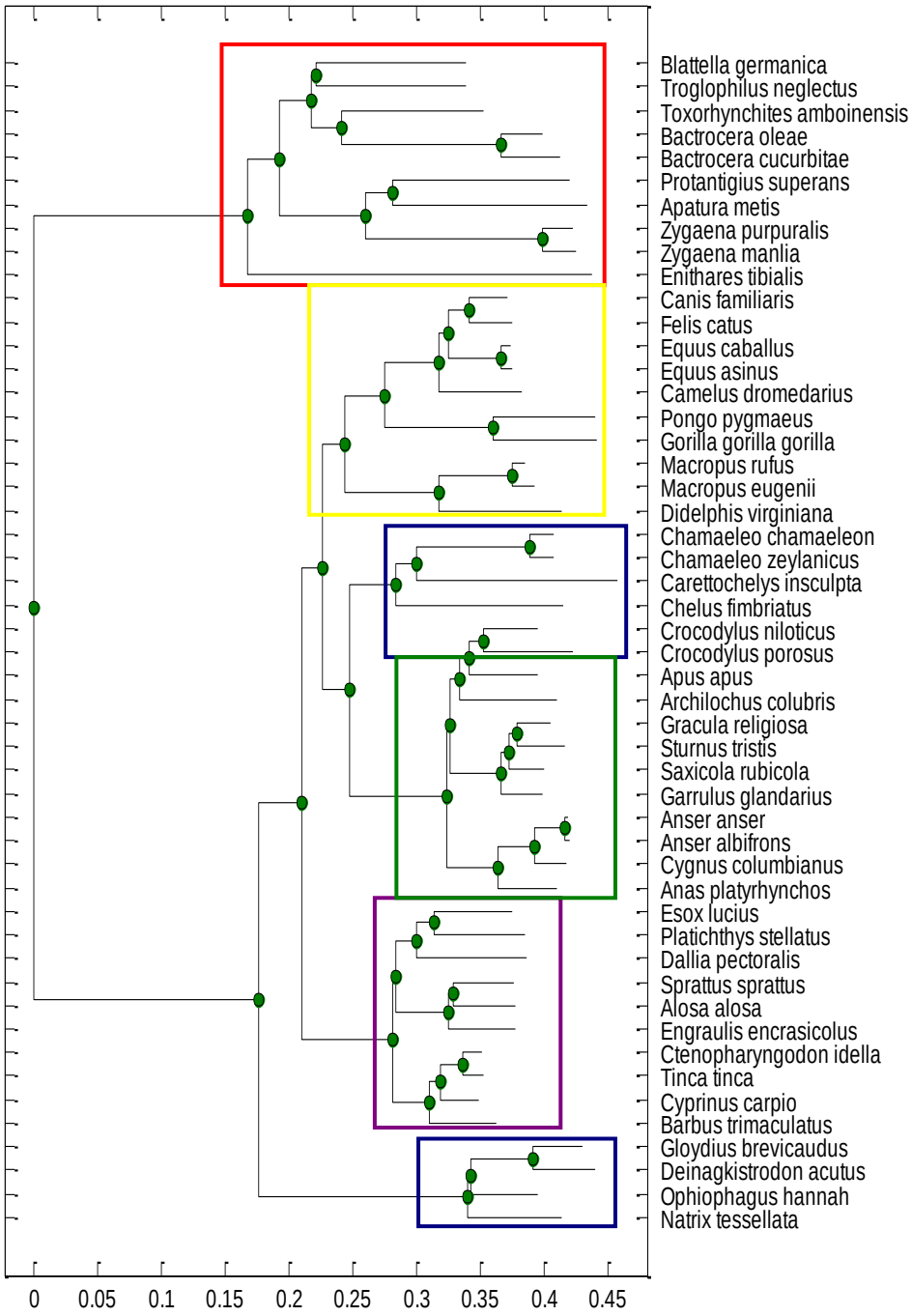
Tato ukázka naznačuje, že všechny tři metody perfektně analyzují na úrovni tříd, jsou tedy o stupeň kvalitnější, než rekonstrukce klasických fylogenetických stromů, u které došlo k zpochybnění evolučních teorémů. Jako přesnější by se dala označit metoda ultrametrická, která zachovává uspořádání jak tříd, tak řádů živočichů.

Hlavním nedostatkem klasické fylogenetické rekonstrukce je volba skórovací matice, která je snadno volitelná pro podobné sekvence, hůře pak pro databáze sekvencí, o kterých uživatel nemá před analýzou žádné informace. V tomto případě je nutno zvolit jednu hodnotu konstantní, která výsledný fylogenetický strom zkresluje. Při rekonstrukci fylogenetických superstromů tento problém odpadá, jelikož volíme skórovací matici vždy pro jeden set sekvencí, nikoli pro více spojených datových sad.

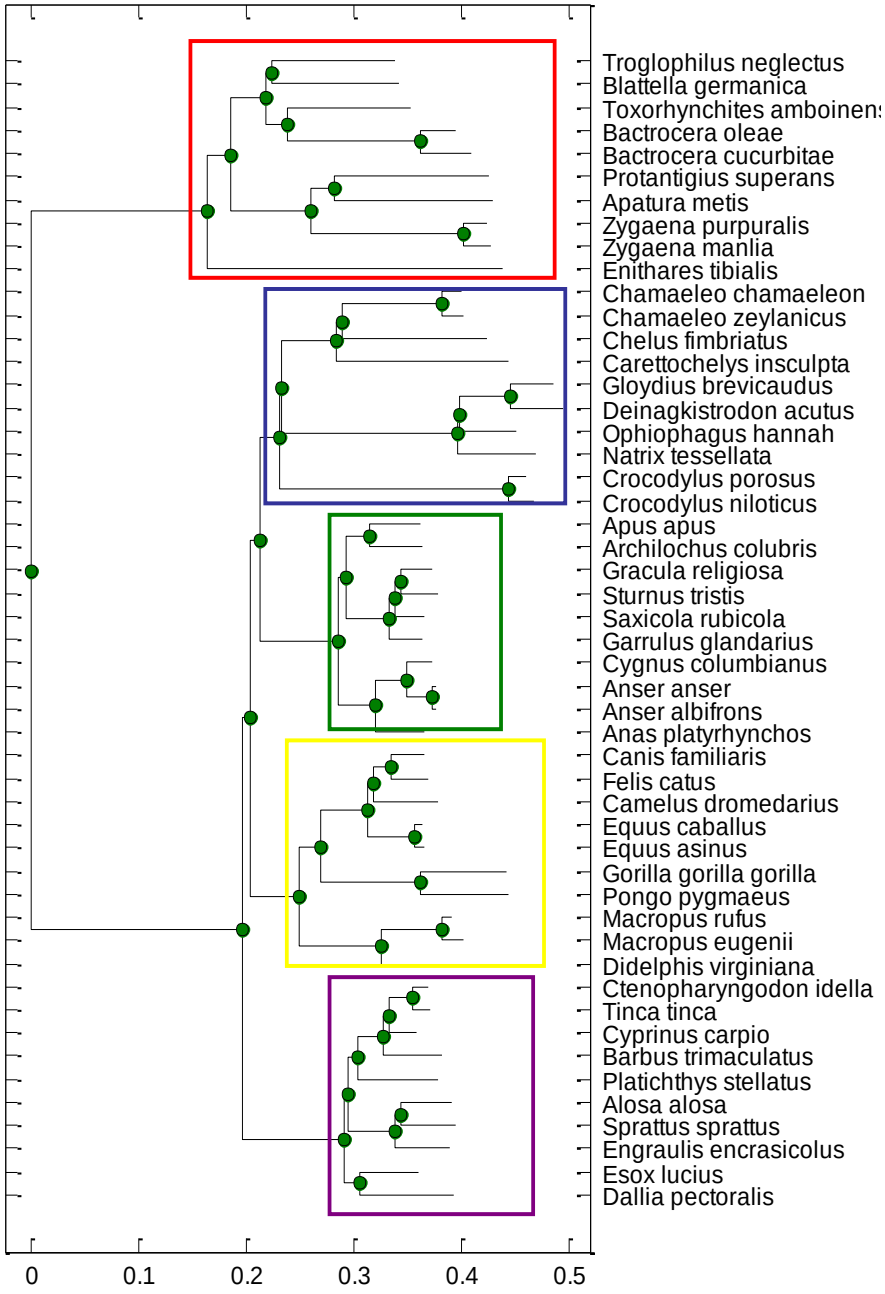
Jedním z důvodů nepřesného zařazování by mohla být volba dat. V této diplomové práci byla demonstrace prováděna na mitochondriálních genech, které jsou těžko klasifikovatelné, jelikož u těchto probíhá evoluce výrazně rychleji. Tyto geny podléhají selekčním tlakům více než klasická mitochondriální DNA.

Chyby by bylo možné eliminovat zvýšením počtu zástupců v jednotlivých třídách případně vyšším počtem druhů zařazených do analýzy. Jako řešení bych zde volila přidání třídy obojživelníků (*Amphibia*).

Jedním z cílů této diplomové práce bylo odstranění nedostatků fylogenetického stromu, který byl sestrojen ze všech padesáti sekvencí v podkapitole 3.4. Jak je vidět z obrázku 33, dva zástupci této třídy byli chybně zařazeno ke třídě ptáků (*Crocodylus niloticus, Crocodylus porosus*). Druhá část třídy plazů je přiřazena do blízkosti třídy ryb. Celý shluk se rozpadl na tři nezávislé části. Ultrametrická metoda fylogenetických superstromů tuhle závažnou chybu odstranila (obrázek 31).



Obrázek 33 Klasický fylogenetický strom



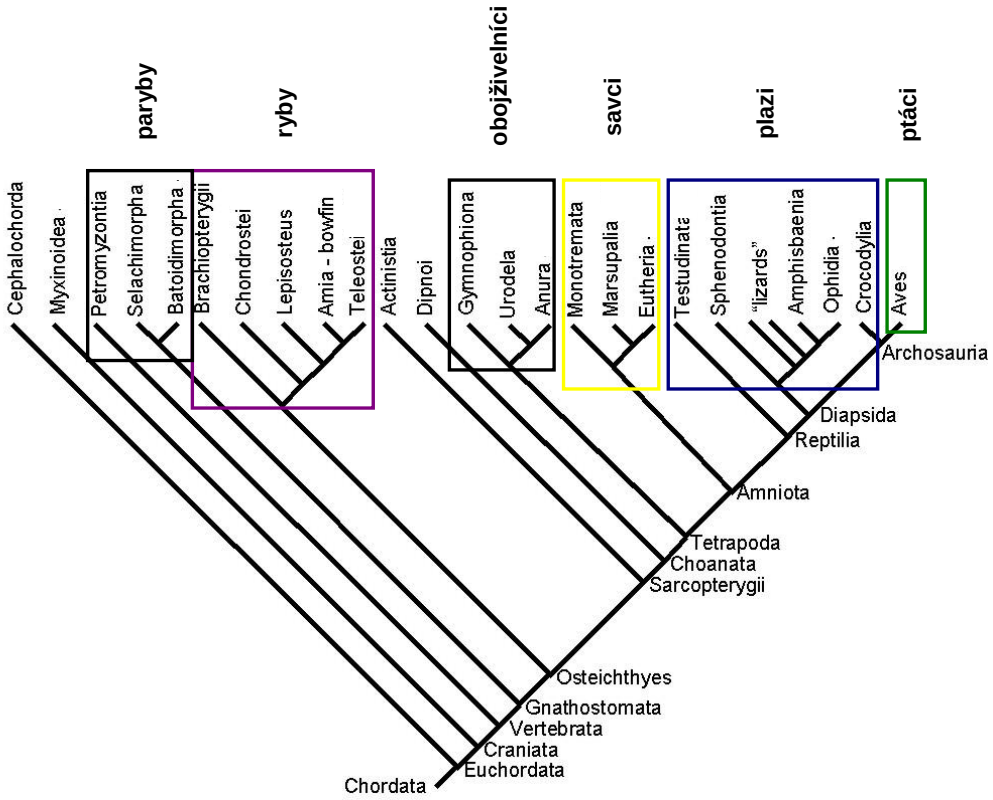
Obrázek 34 Fylogenetický superstrom - ultrametrická metoda

Superstrom na obrázku 34 obsahuje oddělené shluky všech pěti tříd. Zachovány zůstaly také jednotlivé řády živočichů. Tento fylogenetický superstrom není ve výrazném rozporu s tradičními hypotézami fylogenetických vztahů.

Jsou zde jasně patrné dva velké shluky reprezentující kmeny členovců (*Arthropoda*) – horní shluk, a kmen strunatců (*Chordata*). Druhý shluk je tvořen plazi (*Reptilia*), ptáky (*Aves*), savci (*Mammalia*) a ve spodní části jsou zástupci paprskoploutvých ryb (*Actinopterygii*). Nejvzdálenějším shlukem v superstromu je fialově označená třída ryb.

Zásadním důvodem by mělo být vodní prostředí, ve kterém se nachází jako jediní zástupci živočichů v této práci, a způsob jejich rozmnožování. Zbylé tři třídy z kmene strunatců jsou suchozemští živočichové, jsou proto spojeni do jednoho uzlu.

Pro ověření správného sestrojení fylogenetického superstromu je přiložen obrázek 35 vystihující základní schéma evoluce obratlovců (*Vertebrata*).



Obrázek 35 Fylogenetický strom obratlovců [28]

Nejvzdálenější třídou je fialově znázorněná třída paprskoploutvých ryb. V bodě popsaném „Amniota“, který je označením pro blanaté živočichy, si lze povšimnout postupného dělení na třídu savců (žlutě znázorněna), rozsáhlou třídu plazů a poslední list značící třídu ptáků.

Tento strom odpovídá výslednému superstromu, při analýze nedošlo k rozporu s tradičními hypotézami fylogenetických vztahů. Chyby, které nastaly při rekonstrukci fylogenetického stromu klasickými metodami, byly odstraněny.

Výhody fylogenetických superstromů byly v této práci zmíněny již několikrát. Mezi kladné stránky těchto metod patří možnost využití dat z více databází a více zdrojů, což umožňuje získat kvalitnější výsledky prováděných fylogenetických analýz a řešit případy, u kterých nebylo řešení doposud možné. Další nespornou výhodou je možnost rozvržení problému mezi několik malých případů. V této práci to bylo jasně prokázáno v případě tří a více analyzovaných setů. Řešení se rozložilo do několika kroků, které vedly ke vzniku

výsledného superstromu. Metoda průměrného konsensu, která byla realizována, přinesla výsledky, které nejsou v rozporu s evolučními představami. Bohužel se v této práci ukázala jako funkční pouze metoda ultrametrická, která dosahovala nejlepších výsledků. Důvody lze hledat v nedostatečném počtu analyzovaných sekvencí, případně v nesprávně zvoleném typu mitochondriálních genů.

Mezi nevýhody metod fylogenetických superstromů bezesporu patří časová náročnost prováděných algoritmů, která roste s objemem vstupních dat (analyzovaných sekvencí), a je spojena s nutností kvalitní počítačové techniky.

Při volně mezi klasickými fylogenetickými stromy a fylogenetickými superstromy by měl uživatel zvážit, pro jaké účely danou analýzu provádí. Pokud se jedná o orientační schéma znázorňující fylogenezi, volba by měla vést ke klasickým fylogenetickým stromům, které přináší jednoduché znázornění a je nutno počítat se zkreslením, které může nastat. V případě nutnosti volby spolehlivého nástinu situace a analýze velkého objemu sekvencí je vhodné využití rekonstrukce fylogenetických superstromů.

.

Závěr

Tato diplomová práce se zabývala metodami rekonstrukce fylogenetických superstromů. První část je z důvodu pochopení dalšího textu věnována molekulární fylogenetice, zpracování molekulárních dat, fylogenetickým stromům a možnostmi jejich rekonstrukce.

Druhá část práce je zaměřena na fylogenetickými superstromy. Probrány jsou důvody používání těchto stromů a metody pro jejich sestrojení. Zaměřila jsem se především na rekonstrukční metodu s názvem průměrný konsensus, která je v dalších kapitolách realizována. Jedná se o postup, kdy jsou postupně vypočítány vzdálenosti (distance) mezi všemi zdrojovými stromy. Průměrná distance každého taxonu je použitá v konečné distanční matici, ze které je sestrojen superstrom. Při použití této metody mohou nastat situace, kdy se dva taxony nevyskytují společně. V tomto případě se využívá odhad vzdáleností pomocí metody ultrametrické, aditivní a metoda dopočítávání pomocí průměrování.

Třetí část diplomové práce je zaměřena na reprezentování sekvencí, které byly pro tuto práci použity. Data zpracovávaná v této práci byla získána z databáze UniProt. Jako výchozí byl zvolen protein ND1, který se nachází v mitochondriích živočichů. Pro svou práci jsem vytvořila databázi padesáti sekvencí, které zastupují pět tříd živočichů – savců (*Mammalia*), ptáků (*Aves*), plazů (*Reptilia*), ryb (*Osteichthyes*) a hmyzu (*Insecta*).

Čtvrtá kapitola této diplomové práce představuje principielní popis výpočtu distanční matice pro vznik fylogenetických superstromů. Je realizován v programovém prostředí Matlab s Bioinformatickým toolboxem. Na vytvořených souborech sekvencí ve formátu FASTA byly provedeny úpravy vedoucí k vizualizaci zdrojového fylogenetického stromu. Z distančních matic byla vytvořena konečná matice vzdáleností. Odhad vzdáleností byl proveden metodou ultrametrickou, aditivní a aritmetickým průměrem. Po dopočítání chybějící hodnoty byl zobrazen fylogenetický superstrom.

Pátá kapitola je věnována popisu vytvořeného uživatelského interface v Matlab GUI, které by se dalo rozdělit tři části – první obecně slouží k volbě základních parametrů sekvencí, druhá k sestrojování klasických fylogenetických stromů a třetí k sestrojování fylogenetických superstromů. Jsou zde ukázány také výstupy jednotlivých funkcí.

Šestá kapitola se zabývá sestrojenými superstromy a porovnáváním jednotlivých metod sloužících k výpočtu chybějících distančních hodnot. Jsou zde ukázány nedostatky klasických fylogenetických stromů, které nerespektují tradičná hypotézy fylogenetických vztahů, a řešení v podobě superstromů. Na příkladech je zde porovnána ultrametrická, aditivní a metoda průměrování. Nejlepších výsledků dosahovala ultrametrická metoda.

Další stránky této diplomové práce zobrazují klasický fylogenetický strom tvořený všemi sekvencemi, které jsou v této práci použity, a jako jeho oprava je zobrazen fylogenetický superstrom ultrametrickou metodou. Pro ověření správnosti je sestrojený

superstrom konfrontován s veřejně dostupným fylogenetickým stromem, který vystihuje schéma evoluce všech obratlovců (*Craniata*).

Poslední část je věnována shrnutí výhod a nevýhod sestrojování fylogenetických superstromů.

Vzhledem ke stále se vyvíjejícím trendům v evoluční biologii a bioinformatice vidím v metodě rekonstrukce fylogenetických superstromů velkou budoucnost. Jejich výhody jsou nesporné, ať se jedná o možnost použití odlišných typů dat z různých zdrojů nebo možnost sestrojení fylogenetického superstromu několika malých pseudostromů. Za dobu své existence zaznamenaly velký vývoj, který bude zajisté nadále pokračovat i v budoucnosti.

Seznam zkratek

BLOSUM	Blocks Amino-acid Substitution Matrices (typ substituční matice)
GUI	Graphical User Interface
DDBJ	DNA Data Bank of Japan (japonská databáze sekvencí)
DNA	deoxyribonukleová kyselina
EMBL	European Molecular Biology Laboratory (evropská databáze sekvencí)
IUPAC	International Union of Pure and Applied Chemismy
J-C model	Jukes – Cantorův model
ME	Minimum evolution (metoda konstrukce fylogenetických stromů)
ML	Maximum Likelihood
MPR	Matrix representation with parsimony (metoda konstrukce matice pro tvorbu superstromu)
MSSA	The most simile supertree algorithm (metoda konstrukce fylogenetických superstromů
NNI	Nearest neigbor interchange
NJ	Neighbor-Joining (metoda konstrukce fylogenetických stromů)
OTU	Operation Taxonomic unit (operační taxonomická jednotka)
PAM	Point Accepted Mutation (typ substituční matice)
UPGMA	Unweighted Pair Group Method with Arithmetic Mean

Seznam příloh

Použitá literatura

[1] Bioinformatics. IUPAC Codes [online]. 8.2.20011 [cit. 2012-04-15]. Dostupné z: www.bioinformatics.org/sms/iupac.html

[2] BINIDA-EMONDS, O.R.P.: *The evolution of supertrees*. Trends in ecology & evolution (Personal evolution), vol. 19, no. 6, pp. 315-22, Jun.2004.

[3] BININDA-EMONDS, Olaf R.P, GITTLEMAN, Mike A. STEEL. The (Super)tree od Life: Procedures, Problems,and Prospects. *Annual Reviews*. 2002, č. 33, s. 265-289. DOI: 10.1146/annurev.ecolsys.33.01.0802.150511. Dostupné z: <http://sysbio.oxfordjournals.org/content/60/1/32.full.pdf+html>

[4] BUERKI, Félix FOREST, SALAMIN a Nadir ALVAREZ. Comparative Performance of Supertree Algorithms in Large Data Sets Using the Soapberry Family (Sapindaceae) as a Case Study. Systematic Biology. 2011, č. 60, s. 32-44. DOI: 10.1093/sysbio/syq057. Dostupné z: http://sysbio.oxfordjournals.org/content/60/1/32.full.pdf+html

[5] CRANDALL, Keith A. a Jens LAGERGREN. *Algorithms in Bioinformatics*. Karlsruhe. Karlsruhe: WABI, 2008, s. 113-122. ISBN 978-3-540-87360-0.

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

[7] CVRČKOVÁ, Fatima. Úvod do praktické bioinformatiky. 1. Praha : ACADEMIA, 2006. 148 s. ISBN 80-200-1360-1.

[8] DELSUC, Frédéric, Henner BRINKMANN a Hervé PHILIPPE. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*. 2006, č. 6, s. 361-375. DOI: 10.1038/nrg1603. Dostupné z: http://www.nature.com/nrg/journal/v6/n5/abs/nrg1603.html

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

[10] FLÉGR, Jaroslav. *Evoluce sekvence DNA*. In: Praktická metodologie vědy [online]. 1. vyd. 11.10.2007 [cit. 2012-04-15]. Dostupné z: http://darwin.natur.cuni.cz/~flegr/prezentace/mima_evsekvenc.ppt

[11] GATESY, John, Conrad MATHEE, Rob DESALLE a Cheryl HAYASHI. Resolution of a Supertree/Supermatrix Paradox. *Systematic Biology*. 2002, č. 51, 652–664. DOI: : 10.1080/10635150290102311. Dostupné z: <http://sysbio.oxfordjournals.org/content/51/4/652.full.pdf>

[12] HAMPL, Vladimír; NOVOTNÝ, Marián. Přírodovědecká fakulta [online]. Praha : 2010 [cit. 2011-12-04]. *Molekulární taxonomie*. Dostupné z WWW: <http://web.natur.cuni.cz/~vlada/moltax/>.

[13] International Darwin Day Foundation. *Darwinday* [online]. 19.4.2011 [cit. 2012-04-22]. Dostupné z: http://darwinday.org/resources/videos-resources/charles-darwin-and-the-tree-of-life/

[14] LITTNEROVÁ, Simona; JARKOVSKÝ, Jiří. Statistika [online]. Brno : 2010 [cit. 2011-11-06]. *Vícerozměrné statistické metody*. Dostupné z WWW: <http://www.iba.muni.cz/esf/res/file/bimat-prednasky/vicerozmerne-statisticke-metody/VSM-05.pdf>.

[15] MAKARENKOV, Vladimir; LAPOINTE, Francois-Joseph. *A weighted least-squares approach for inferring phylogenies from imcomplete distance matrices*. Bioinformatic. 2004, 132004, s. 2113-2121.

[16] National Center for Biotechnology Information (NCBI) [online]. 2011 [cit. 2011-11-22]. GenBank. Dostupné z WWW: <http://www.ncbi.nlm.nih.gov/sites/gquery>.

[17] NEČÁSEK, JAn. *Genetika*. 2. Praha : Scientia, 1993. 112 s. ISBN 80-7183-085-2.

[18] ŘEHULKA, Pavel. *Bioinformatika* [online]. Hradec Králové : 2009 [cit. 2011-12-01]. Základy bioinformatického zpracování dat v proteomice. Dostupné z WWW: <http://www.pmfhk.cz/WWW/UMP/bioinformatika_pr_cv.pdf>.

[19] SAITOU, Naruya; NEI, Masatoshi. *The Neighbor-joining Method:* A New Method for Reconstructing Phylogenetic Trees. Molekular Biology. 1987, s. 406-425.

[20] SEDLÁŘ, K. *Bootstrappingové metody ve fylogenetice: bakalářská práce.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2011. 55 s. Vedoucí práce Ing. Helena Škutková

[21] VUT, FEKT studijní materiály k předmětu Analýza biologických sekvencí (2010), garant předmětu: Provazník,I.

[22] VUT, FEKT studijní materiály k předmětu *Praktika z bioinformatiky* (2009), garant předmětu: Kolářová, J.

[23] WIENS. Missing Data, Incomplete Taxa, and Phylogenetic Accuracy. *Systematic Biology*. 2003, č. 52, s. 528-538. DOI: 10.1080/10635150390218330. Dostupné z: http://sysbio.oxfordjournals.org/citmgr?gca=sysbio;52/4/528

[24] WOLF, Yuri I., Igor B. ROGOZIN a Nick V. GRISHIN. Genome trees and the tree of life. *Trends in Genetics*. 2004, č. 18, 472–479. DOI: 10.1016/S0168-9525(02)02744-0. Dostupné z: http://www.sciencedirect.com/science/article/pii/S0168952502027440

[25] YING, Cao. Conflict Among Individual Mitochondrial Proteins in Resolving the Phylogeny of Eutherian Orders. Molecular Evolution. 2002, č. 47, s. 307-322. Dostupné z: http://nsmserver2.fullerton.edu/departments/chemistry/evolution_creation/web/YingCao.pdf

[26] ZAPLATÍLEK, Karel; DOŇAR, Bohuslav. *Matlab : tvorba uživatelských aplikací*. 1. Praha : BEN, 2004. 215 s. ISBN 80-7300-133-0.

[27] ZHANG, Wei, Zhirong SUN, Random local neighbor joining: A new method for reconstructing phylogenetic trees, *Molecular Phylogenetics and Evolution*, 2008, č.47, s. 117-128, DOI: 10.1016/j.ympev.2008.01.019. Dostupné z: http://www.sciencedirect.com/science/article/pii/S1055790308000419

[28] Zoology: Comparative Vertebrate Anatomy [online]. 2008 [cit. 2012-04-15]. Dostupné z: http://www.ou.edu/class/zoo2204/CVAhome.html