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

KOMPARAČNÍ ANALÝZA GENOMICKÝCH DAT POMOCÍ GRAFICKÉ REPREZENTACE

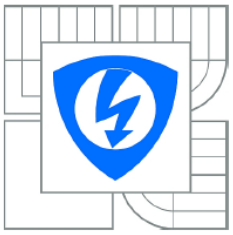
SIMILARITY/DISSIMILARITY ANALYSIS OF GENOMIC DATA ON THE BASIS OF GRAPHICAL REPRESENTATION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — JIŘÍ TĚTHAL
AUTHOR

VEDOUCÍ PRÁCE — Ing. DENISA MADĚRÁNKOVÁ
SUPERVISOR

BRNO 2011



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor
Biomedicínská technika a bioinformatika

Student:	Jiří Těthal	***ID:***	119750
Ročník:	3	***Akademický rok:***	2010/2011

NÁZEV TÉMATU:

Komparační analýza genomických dat pomocí grafické reprezentace

POKYNY PRO VYPRACOVÁNÍ:

Seznamte se s metodou identifikace živočišných druhů pomocí DNA barcodingu a grafické reprezentace DNA sekvencí pomocí výpočtu denzity jednotlivých nukleotidů a dvojic komplementárních nukleotidů. V programovém prostředí Matlab vytvořte funkci pro výpočet denzity DNA sekvence a funkci pro porovnání denzity dvou sekvencí distanční metodou.
V programovém prostředí Matlab vytvořte uživatelské grafické rozhraní, které bude sloužit pro výpočet denzity DNA sekvence a porovnávání denzity různých sekvencí. S pomocí tohoto programu vytvořte databázi DNA sekvencí mitochondriálního genu COI, které lze získat z veřejně přístupné databáze. Vytvořená databáze bude obsahovat původní sekvenci DNA a programem vypočtené denzity nukleotidů. Pro vybraný soubor dat proveďte komparační analýzu pomocí porovnávání denzity nukleotidů.

DOPORUČENÁ LITERATURA:

[1] JAN, J. Číslicová filtrace, analýza a restaurace signálů. VUTIUM, Brno, 2002.
[2] MORITZ, C., CICERO, C. DNA Barcoding: Promise and Pitfalls. PLoS Biology. 2004, vol. 2, no. 10, p. 1529-1531.

Termín zadání:	7.2.2011	***Termín odevzdání:***	27.5.2011

Vedoucí práce: Ing. Denisa Maděránková

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

ABSTRAKT

Práce se zabývá identifikací druhů živočichů pomocí denzity nukleotidů mitochondriálního genu CO1.

V první části práce jsou teoreticky shrnuty informace o DNA barcodingu a buněčné organele mitochondrii, která s touto metodou úzce souvisí.

Druhá část se již prakticky zabývá porovnáváním různých sekvencí s využitím denzity jejich nukleotidů. K tomuto účelu byl vytvořen program, který využívá dvě funkce, přičemž první ze zadaných sekvencí nukleotidů vypočítá jejich denzitu a druhá dokáže tyto denzity porovnat distanční metodou.

ABSTRACT

The work deals with the identification of species of animal through the density of nucleotids of mitochondrial gene CO1.

In the first part the theory summarized information about DNA barcoding and the mitochondria, cellular organel that with this method is closely related.

The second part deals with virtually comparing different sequences using the density of nucleotides. For this it was created program, which uses two functions, the first of the specified nucleotide sequence calculates the density and one can compare the density by distance methods.

Klíčová slova

BOLD, CBOL, denzita DNA, dendrogram, distance, DNA barcoding, druh, fasta, gen CO1, genom, iBOL, mitochondrie, mitochondriální DNA, PCR

Keywords

BOLD, CBOL, dendrogram, density of DNA, distance, DNA barcoding, species, fasta, gene CO1, genome, iBOL, mitochondria, mitochondrial DNA, PCR

Bibliografická citace

TĚTHAL, J. *Komparační analýza genomických dat pomocí grafické reprezentace*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2011. 54s. Vedoucí bakalářské práce Ing. Denisa Maděránková.

Prohlášení

Prohlašuji, že svoji bakalářskou práci na téma Komparační analýza genomických dat pomocí grafické reprezentace jsem vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.
Jako autor uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

V Brně dne 27. května 2011 .....
podpis autora

Poděkování

Děkuji vedoucí bakalářské práce Ing. Denise Maděránkové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé bakalářské práce.

V Brně dne 27. května 2011 .....
podpis autora

Obsah

Seznam ilustrací

Seznam tabulek

1. Úvod

Potřeba identifikace živočišných druhů se projevuje v mnoha odvětvích současné vědy. Od základní taxonomie, evoluční biologie přes ochranu živočichů, sledování kvality životního prostředí až po identifikaci potravin.

Tato práce se zaměřuje na jednu z metod umožňujících právě tuto identifikaci, a to metodu DNA barcodingu, která je poměrně „mladá“, avšak má velký potenciál. Je to taxonomická metoda, která k identifikaci jednotlivých druhů zvířat a rostlin využívá krátký specifický úsek nejčastěji mitochondriální DNA. Na rozdíl od klasického porovnávání morfologických znaků je tato metoda rychlá a levná a tím zaujala celou řadu vědců.

V úvodní kapitole je popsán krátký přehled problematiky určování druhů a srovnávání DNA. Protože k identifikaci využívá barcoding krátkou oblast mitochondriálního genu, je třetí kapitola věnována právě této důležité buněčné organele.

Čtvrtá kapitola pojednává o samotném teoretickém rozboru DNA barcodingu. Pátá kapitola pak poskytuje informace o databázích barcode.

Šestá a sedmá kapitola se již prakticky zabývá popisem programového řešení v prostředí Matlab. sloužícího pro porovnávání různých sekvencí s využitím denzity jejich nukleotidů. V osmé kapitole jsou shrnuty výsledky porovnání dendrogramů souboru 30 ryb.

2. Přehled problematiky určování druhů a srovnávání DNA

Ještě nedávno byly jednotlivé druhy identifikovány jen pomocí morfologických rysů, například podle tvaru, velikosti či barvy různých částí těla. Tuto práci musí provádět zkušení taxonomisté. Takto byly za 250 let popsány „pouze“ 2 miliony druhů. „Pouze“ proto, že počet druhů eukaryotních organismů na naší planetě se odhaduje v rozsahu 10 až 100 milionů.

Drtivá většina organismů na Zemi má ale rozměry menší než 1 mm a u těch je taxonomické rozlišování velmi složité, někdy bez elektronového mikroskopu takřka nemožné. Přitom tyto mikrofauny a mikroflóry představují většinu života v půdě a v oceánech a poskytují saprofytický i produktivní základ pro většinu větších organismů, včetně člověka.

Proto je zde potřeba jiného způsobu třídění. Rychlejšího, levnějšího a použitelného pro co nejvíce organismů. Hledanou cestou může být určování druhů pomocí nové metody, která využívá pro identifikaci krátké sekvence většinou mitochondriálního genu. Mluvíme o tzv. DNA Barcodingu. Ten nabízí rychlou identifikaci druhu i z malého množství vzorku (tkáně) zkoumaného jedince.

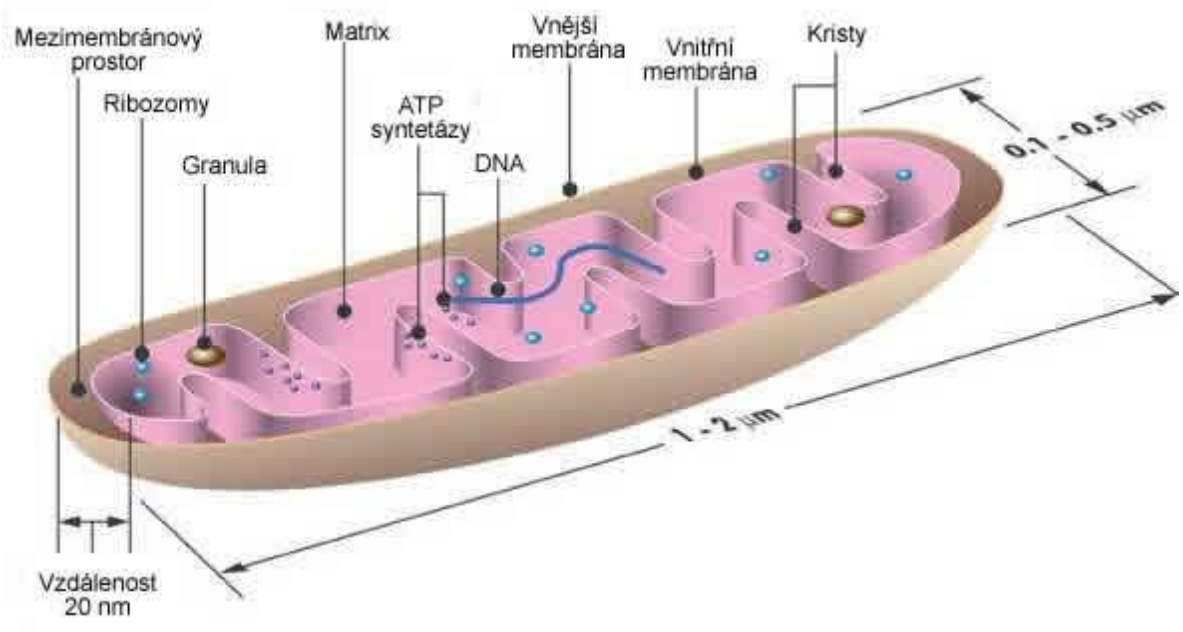
Hlavním cílem této metody není klasifikace jednotlivých druhů, ale určení neznámého vzorku. Ačkoli někteří vědci se snaží využít této cesty i k určování a zařazování nových druhů, jiní tvrdí že to sahá nad rámec možností této metody.

Obrázek 1: DNA barcoding

3. Mitochondrie

Jsou to jedny z nejdůležitějších organel eukaryotické buňky. Slouží k buněčnému dýchání a s tím spojenému získávání energie. Slovo mitochondrie pochází z řeckých slov *mitos* (stuha, vlákno) a *chondros* (zrníčko), tj. vláknitá zrníčka, protože se tak jeví pod světelným mikroskopem. Dosahují obvykle rozměrů v řádu několika mikrometrů a v buňce se jich vyskytuje několik stovek, ale i sto tisíc.

Mitochondrie je tvořena dvěma membránami – vnější a vnitřní. Vnitřní membrána vybíhá v četné výběžky – kristy. Mezi kristami se vyskytuje základní hmota mitochondrie – matrix. Ta pak obsahuje jednotlivé ribozomy, mitochondriální genom, enzymy atd.



Obrázek 2: Stavba mitochondrie

Označujeme je jako semiautonomní organely, protože jsou na buňce částečně nezávislé. Soudí se, že to jsou bakteriální buňky, které v průběhu evoluce pronikly do jiné buňky a zde se přizpůsobily (tzv. endosymbióza). Bakteriálnímu původu nasvědčuje např. dvojitá membrána, cirkulární molekula DNA nebo specifický typ dělení.

Uvnitř mitochondrií probíhá cyklus kyseliny citrónové (Krebsův cyklus, citrátový cyklus), kde je oxidován acetylkoenzym A (acetyl-CoA) a produkován je oxid uhličitý a energie ve formě vysokoenergetických elektronů. Ty z cyklu putují formou NADH a FADH2, které v dýchacím řetězci odevzdávají tyto elektrony za vzniku vody a ATP.

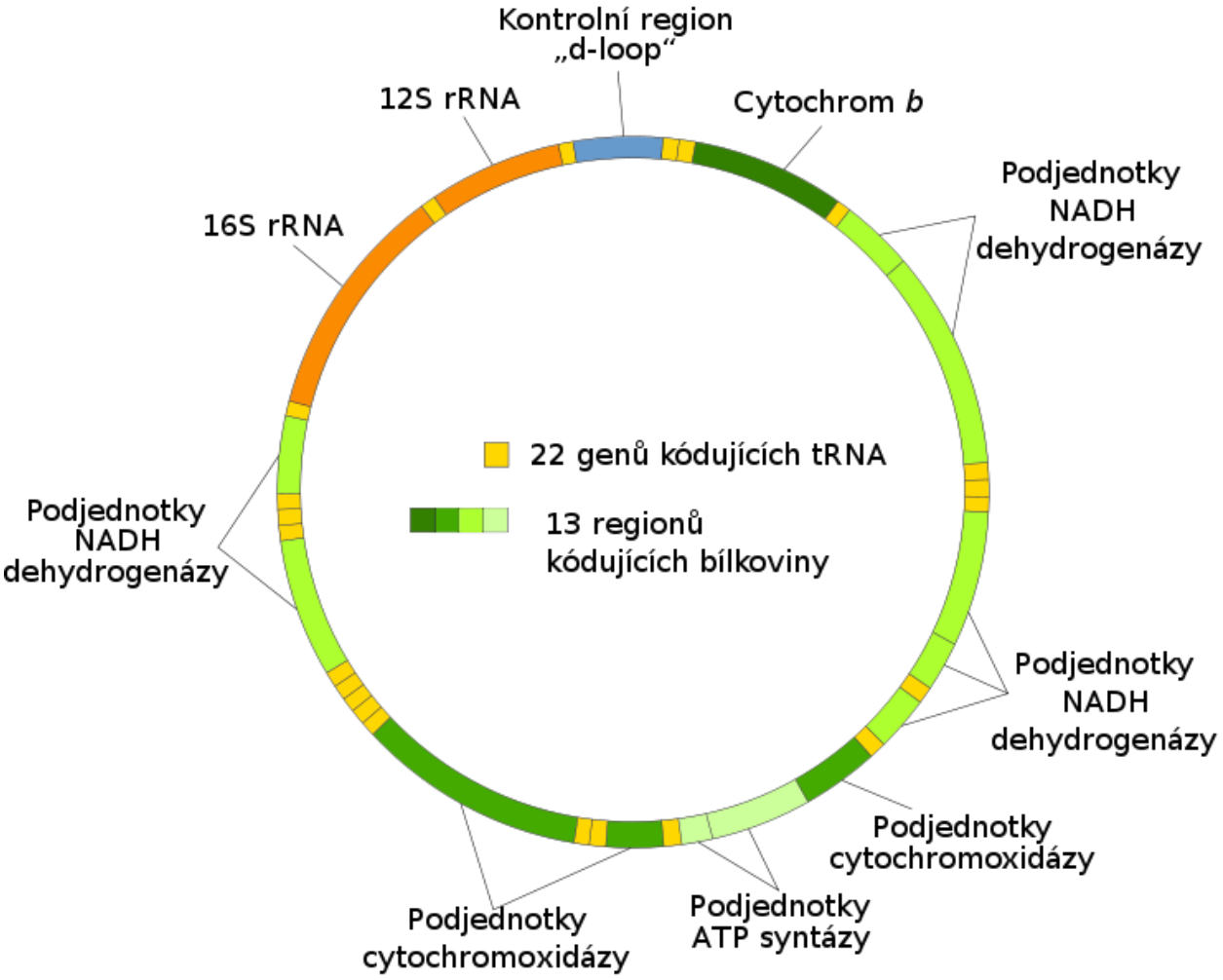
Mezi další funkce mitochondrií patří buněčná diferenciace, buněčná smrt, ale i kontrola buněčného cyklu a růstu. [1] [2]

3.1. Mitochondriální dědičnost

Mitochondrie se dělí v závislosti na energetické potřebě buňky. Když buňka potřebuje energie hodně, mitochondrie se dělí a rostou, naopak když je potřeba snížená, mitochondrie umírají nebo se stávají neaktivními. Mitochondrie se dělí podélným štěpením a rozložení mitochondrií v dceřiných buňkách je víceméně náhodné.

Mitochondriální DNA (mtDNA) je tvořena kruhovou molekulou (cirkulární tvar) a obvykle se nachází v mitochondriální matrix. Někdy je připojena k vnitřní membráně.

Například u člověka obsahuje 16569 párů bází, které tvoří 37 genů – z toho 24 genů kóduje různé části proteosyntetického aparátu mitochondrie (2 typy rRNA a 22 tRNA) a zbylých 13 kóduje mitochondriální enzymy. Většina mitochondriálních proteinů je ale kódována v jádře buňky a ty jsou sem přeneseny z cytosolu. Odlišnost proteosyntetického aparátu mitochondrie od normálního aparátu eukaryotické buňky je značná (velikost, počet molekul).



Obrázek 3.: Schéma mitochondriální DNA člověka

Charakterem se mtDNA podobá prokaryotnímu nukleoidu a ne eukaryotickým chromozomům. Na rozdíl od jaderného genomu není moc obsáhlá, ale může zaujímat většinu buněčného genomu, v závislosti na velkém počtu mitochondrií, a více identických kopií kruhové mtDNA v jedné mitochondrii. Stejně jako u prokaryot se v mtDNA nevyskytují nekódující sekvence, tzv. introny. Dále u mitochondrií oproti DNA eukaryot nalézáme jiné stopkodony (AGA, AGG). Probíhá zde rychlejší a četnější replikace DNA bez činnosti opravných mechanismů a nejsou zde přítomny bílkoviny typu histonů. [3] [4]

Tabulka 1: Kódování DNA eukaryot

Kodon	Aminokyselina	Kodon	Aminokyselina	Kodon	Aminokyselina	Kodon	Aminokyselina
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	STOP	UGA	STOP
UUG	Leu	UCG	Ser	UAG	STOP	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met (START)	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Tabulka 2: Změny kódování DNA u mitochondrií

Organismus	Kodon	Běžně	Změna
Savci	AGA, AGG	Arginine	Stop codon
	AUA	Isoleucine	Methionine
	UGA	Stop codon	Tryptophan
Bezobratlí	AGA, AGG	Arginine	Serine
	AUA	Isoleucine	Methionine
	UGA	Stop codon	Tryptophan
Kvasinky	AUA	Isoleucine	Methionine
	UGA	Stop codon	Tryptophan
	CUA	Leucine	Threonine

Mitochondriální DNA se dědí materálně (po matce), protože mitochondrie v zárodku pocházejí většinou pouze z vajíčka (při oplození vajíčka jsou otcovské mitochodrie ze spermie zničeny velkým výkonem, který spermie potřebuje při pohybu). V ojedinělých případech se v cytoplasmě vyskytují mitochondrie mateřské i otcovské (nebyly zničeny v zygotě). Takový jedinec se nazývá cybrid (cytoplasmatický hybrid).

Mutace v mtDNA mohou být zdrojem různých dědičných metabolických onemocnění způsobených mutací buď v jádře v genech pro mitochondriální enzymy, nebo v mitochondriální DNA. Mutace mohou mít vliv na oxidativní fosforylaci, cyklus kyseliny citrónové, β-oxidaci a jiné funkce mitochondrie

Vzhledem k tomu, že se zde jedná o specifický typ dědičnosti, pak je-li matka přenašečkou mutace v mtDNA, předá ji všem svým potomkům, pokud je přenašečem otec, mutaci předat nemůže.

Při dělení buňky nedochází ke kontrolovanému rozdělení mitochondrií do dceřiných buněk, proto mohou dceřiné buňky získat různý počet normálních a mutovaných mitochondrií.

Buňka, jež obsahuje pouze normální mitochondrie nebo pouze mutované mitochondrie se nachází ve stavu homoplazmie. Pokud obsahuje směs normálních a mutovaných mitochondrií, je ve stavu heteroplazmie.

Příklady chorob, které se projeví při zastoupení 60-90 % mutantních mitochondrií v daném místě tkáně: Leberova hereditární optická neuropatie, Pearsonův syndrom, MELAS a jiné. [55]

Geny otce a matky se tedy v mtDNA nerekombinují. Jelikož nemá opravné systémy, mutuje mnohem rychleji než jaderná DNA. Právě těchto vlastností využívá řada genetiků a biologů k rozlišení druhů a populací. [5] [6]

3.2. Cytochrom C oxidáza

1. **Popis:**

Je to velký transmembránový komplex proteinů, který se jako poslední účastní elektronového transportu v mitochondriálním dýchacím řetězci. Tento enzym katalyzuje redukci kyslíku na vodu a dopravuje protony (protonová pumpa) přes biologickou membránu.

Primární struktura se skládá z řetězce asi 100 aminokyselin a molekulová hmotnost je asi 12 000 daltonů. Mnoho organismů vyššího řádu má řetězec 104 aminokyselin.

Součástí komplexu proteinů u savců je integrální membránový protein složený z několika „kovových" míst a 13 bílkovinných podjednotek. Podjednotky I - II jsou syntetizované mitochondrií a zbylé podjednotky IV - XIII jsou kódované jádrem.

2. **Funkce:**

Elektrony, které pocházejí ze živin (citrátový cyklus, oxidace mastných kyselin a anaerobní glykolýza), předá kyslíku přes koenzym Q a cytochrom c. Redukuje ho 4 elektrony, přičemž vzniknou 2 molekuly vody. Přitom uvolní energii pro syntézu adenozin-5'-trifosfátu (ATP). Jako akceptor elektronů v citrátovém cyklu slouží NADH, který se oxiduje na NAD+. Meziprodukty při redukci kyslíku jsou volné radikály - peroxid vodíku a superoxid, ty však zůstávají navázané na enzym. Zjednodušeně:

4 cytochrom C (red.) + O_2 + 4 H+ → 4 cytochrom C (oxid.) + 2 H_2O

Cytochrom c se také podílí na zahájení apoptózy (řízené formy buněčné smrti, používané k usmrcení buňky v procesu vývoje, v reakci na infekci nebo poškození DNA). Uvolnění velkého množství cytochromu c do cytoplasmy vede k aktivaci proteáz, které jsou zodpovědné za zničení buňky.

3. **Využití:**

Cytochrom c je využíván při LLLT (Low-laserová terapie). Laserový paprsek s vlnovou délkou blížící se infračervené oblasti proniká tkání, kde v buňkách zvyšuje aktivitu cytochromu c, čímž se zvyšuje metabolická aktivita a uvolní se více energie potřebné pro buňku k regeneraci. [7] [8]

4. DNA barcoding

Je to taxonomická metoda, která využívá krátký specifický úsek DNA k identifikaci (přiřazení) jednotlivých druhů rostlin nebo živočichů.

4.1. Historie

Jako první o DNA barcodingu publikoval Dr. Paul Hebert (University of Guelph in Ontario, Canada) v roce 2003 (především v článku Biological identifications through DNA barcodes, Hebert et al. 2003). Možnost identifikace druhů pomocí velmi krátké sekvence genu okamžitě zaujala řadu biologů a genetiků.

4.2. „Nejvhodnější“ gen

Vědci požadovali od nejvhodnější části genu, aby byl úsek dostatečně krátký pro rychlou analýzu, a zároveň dost dlouhý na to, aby mohl spolehlivě identifikovat jednotlivé druhy.
U živočichů a mnoha eukaryot se jako nejvhodnější ukázala být pro barcoding mitochondriální DNA a její gen CO1. Tento gen vyrábí klíčový enzym - „podjednotka 1 cytochrom oxidáza“, který je součástí dýchacího řetězce, kde katalyzuje redukci kyslíku na vodu.
Tento gen obsahuje standardně 648 párů bází. Mitochondriální DNA byla vybrána také proto, že jediná buňka obsahuje až 1 000 mitochondrií, ve kterých je více kopií DNA. Proto i malý vzorek může posloužit k získání dostatečného množství DNA pro úspěšné sekvenování.
Je vhodný, protože v rámci druhu se tento úsek liší jen nepatrně, naproti tomu rozdíly mezi druhy jsou patrné. Například při porovnání člověka a šimpanze se tento gen liší asi na 60 místech.
Pro spolehlivou detekci je nutné popsat alespoň 10 jedinců každého druhu.
Klasický morfologický popis druhu potřebuje více exemplářů obou pohlaví, zatímco pomocí barcodingu analyzujeme i malou část těla.
V roce 2009 databáze sekvencí CO1 zahrnovaly nejméně 620 000 jedinců z více než 58 000 druhů zvířat, což je více než databáze jakéhokoliv jiného genu.

Pro rostliny gen CO1 nemůže být použit, protože se v rostlinách mění málo, a proto jsou rozdíly mezi druhy nepatrné. Byly tudíž určeny vhodnější markery, kterými jsou geny chloroplastů rbcL a matK . [9]

4.1. „Birds identification“

V roce 2004 provedl tým pracovníků okolo Dr. Paula Heberta výzkum u severoamerických ptáků. Byl analyzován čárový kód 260 druhů a bylo zjištěno, že každý z těchto druhů má jinou sekvenci CO1. Rozdíly sekvence CO1 mezi druhy byly v průměru 7,93 %, rozdíly v rámci druhu byly v průměru pouze 0,43 %. Proto navrhli, aby mezera (hranice) mezi druhy byla, vždy pro konkrétní studii, rovna desetinásobku průměrných vnitrodruhových rozdílů. [10]

4.2. Problémy

Například u členovců se často vyskytují symbionti, kteří mohou způsobit změny v mtDNA a to pak může vést k nesprávnému odvození evoluční historie daného organismu. Dalším problémem jsou tzv. pseudogeny – kopírování mitochondriální DNA do jaderné DNA, které pak mohou také zkreslovat výsledky analýzy. [11]

4.3. Standardní postup analýzy vzorků

1. Vzorky z různých úložišť biologických materiálů slouží k vytvoření podkladu pro vlastní identifikaci. Mezi hlavní zdroje vzorků patří například různé sbírky, zoologické zahrady, muzea, ale i volná příroda.

2. Ze vzorků se v laboratořích pomocí PCR získá požadovaná sekvence DNA čárového kódu. Tyto údaje pak putují do databází, v kterých probíhá další zpracování. V nejmodernějších laboratořích může tato analýza trvat jen několik hodin.

3. Databáze jsou nejdůležitější částí řetězce. Proto je snaha vytvořit centralizovanou databázi všech živočišných druhů. Tato databáze by sloužila k porovnání neznámého vzorku se záznamy v databázi a tím k určení původu vzorku.
   Největší 3 databáze DNA – GenBank, EMBL a DDBJ se pro záznam dat DNA barcodingu dohodly na datovém standardu CBOL.

4. Pomocí analýzy dat můžeme porovnávat jednotlivé záznamy v databázích. CBOL nabízí portál, který umožňuje vědcům jednoduše ukládat, spravovat, analyzovat a zobrazovat jejich barcoding záznamy. [9]

4.4. Získávání genomických dat

4.4.1. Polymerázová řetězová reakce (PCR - Polymerase Chain Reaction)

Slouží k rychlému namnožení určitých úseků sekvencí DNA. Je založena na principu replikace nukleových kyselin. Úseky se na začátku a na konci označí primery.
Tímto způsobem se vytvoří přibližně 10 tisíc kopií vzorového fragmentu DNA. Syntéza nového vlákna probíhá pomocí termostabilní DNA polymerázy (nejčastěji Taq polymerázy izolované z termofilní bakterie Thermus aquaticus).
PCR probíhá v zařízení zvaném termocykler, ve kterém se rychle mění teplota o několik desítek stupňů Celsia a v ideálním případě dochází ke zdvojení sekvence a její množství tak exponenciálně roste.

Průběh PCR:

1. **Denaturace -** po dobu 20-30 sekund při teplotě 94-98 °C, kdy dochází k narušení vodíkových můstků a rozpojení dvoušroubovice.
2. **Nasednutí primerů** – při teplotě 50-65 °C nasedají primery na specifická místa DNA.
3. **Syntéza DNA** – při různé teplotě (závislá na DNA polymeráze) dochází k syntéze nového vlákna DNA od 5' konce ke 3' konci.

Tento cyklus se obvykle 30krát opakuje, což vede ke vzniku obrovského množsví kopií původní sekvence (z jedné molekuly až 1 miliarda nových).
Metody se využívá nejvíce k vědeckým potřebám, kdy se sekvence analyzují, sekvenují či se z nich vytváří fylogenetické stromy.

4.4.2. Sekvenování DNA

Sekvenování DNA je vlastně zjišťování pořadí nukleových bází (A, C, G, T) v sekvencích DNA. To probíhá pomocí mnoha biochemických metod.
Mezi základní patří Maxam-Gilbertovo sekvenování a Sangerovo sekvenování, kdy se využívá gelové elektroforézy. Mezi nejnovější patří pyrosekvenování či Single molecule real-time, které využívají detekci uvolněného světelného záření. [12] [13]

4.5. CBOL a iBOL

Jedná se o dvě velké mezinárodní iniciativy na poli DNA barcodingu.

4.5.1. CBOL (The Consortium for the Barcode of Life)

Je to datový standard, který byl založen v roce 2004. Věnuje se začlenění DNA barcodingu jako nového globálního vědeckého standardu pro identifikaci druhů. Pro jeho podporu pořádá školení, semináře, konference i různé pracovní skupiny. Nicméně neprovádí žádné konkrétní získávání dat pomocí DNA barcodingu. [14]

4.5.2. iBOL (The International Barcode of Life project)

Projekt, který byl založen v roce 2007 na University of Guelph, Canada. S pomocí partnerů z 25 zemí světa se podílí na výzkumu technologií pro rychlejší analýzu druhů. Klade si za cíl do 5 let analyzovat 5 milionů exemplářů od 500 000 druhů. To má sloužit k sestavení knihovny, díky které by bylo možné rychle a levně identifikovat jednotlivé druhy pomocí DNA barcodingu. [15]

5. Databáze barcode

DNA databáze barcode a sekvencí pro identifikaci organismů se nacházejí v tabulce 3. [16]

Tabulka 3: Seznam databází

Jméno databáze	Adresa	Organismy
Algaterra	www.algaterra.org	Řasy
BOLD	www.boldsystems.org	Všechny organismy
ISTH	www.isth.info	Trichoderma (Houby)
Mycobank	www.mycobank.org	Houby
Nematol	http://nematol.unh.edu/	Hlístice
Silva	www.arb-silva.de	Bacteria, Archaea, Eukarya
Sponge Barcoding Project	www.spongebarcoding.org	Mycí houby
MOTU	http://nemhelix.cap.ed.ac.uk:880	Hlístice
Unite	http://unite.ut.ee	Houby ektomykorhizní

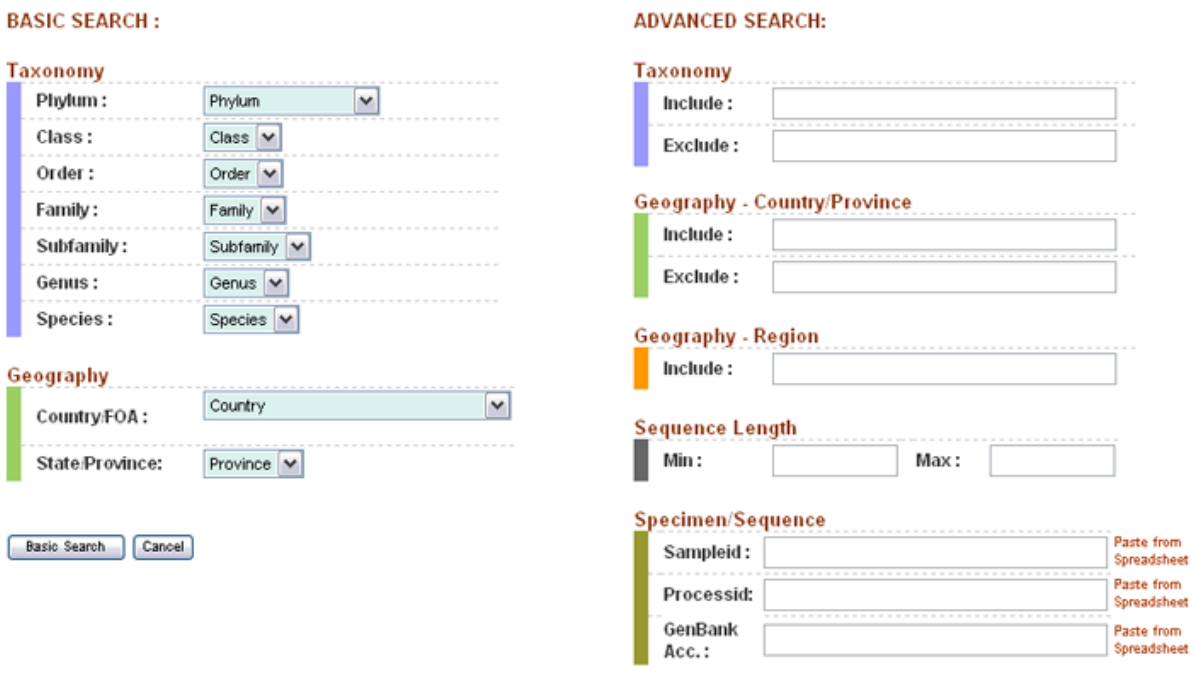
5.1.1. BOLD

Barcode of Life Data Systems (BOLD) - databáze sloužící k ukládání, analýze a zveřejňování záznamů čárového kódu DNA. BOLD je volně dostupná pro každého výzkumníka, který se zajímá o DNA barcoding. Záznamy zde splňují normy nutné pro BARCODE data společná pro všechny databáze.

BOLD má společné rozhraní pro 3 způsoby vyhledávání. A to hledání ve všech záznamech, v souboru projektů nebo v jednom konkrétním projektu.

Dále existují dva typy vyhledávání v BOLDu - základní vyhledávání a pokročilé vyhledávání. Základní vyhledávání umožňuje pomocí rolovacího menu vyhledávat podle taxonomie nebo zeměpisné polohy záznamu v BOLD. Pomocí rozevíracího výběru se obnovují volby pro další úrovně výběru.

Pokročilé vyhledávání umožňuje více specifik. Tím můžeme zúžit rozsah každého hledání. V pokročilém vyhledávání můžeme také vyloučit určitá kritéria, aby se zabránilo nechtěným výsledkům.



Obrázek 4.: Pokročilé vyhledávání na BOLDu

5.1.2. EMBL

EMBL obsahuje skoro čtrnáct miliard nukleotidů tvořících mnoho genů a genomů z různých organismů.

Databáze EMBL je organizována Evropskou molekulárně biologickou laboratoří (EMBL). Je to veřejná evropská primární nukleotidová databáze se sídlem v Anglii na adrese http://www.ebi.ac.uk/embl. Databáze je vytvářena v součinnosti s ostatními nukleotidovými databázemi GENBANK (USA) a DDBJ (Japonsko) a je velmi dobře přístupná spolu s mnoha odvozenými a dalšími databázemi přes SRS (Sequence Retrieval System) například na adrese http://srs6.ebi.ac.uk. Databáze obsahuje všechna data zaslaná vědeckou komunitou, a to bez kontroly. Z tohoto důvodu může obsahovat určité procento chyb. Manuál k databázi je k dispozici na adrese http://www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html.

Pro schválení CBOLem musí každý BARCODE záznam v EMBL obsahovat řadu povinných polí a některé důrazně doporučené pole (viz následující tabulky), aby byly BARCODE údaje ve všech databázích BARCODE projektů stejné. [17]

Tabulka 4: Povinná pole

Skupina	Pole	Popis
Druh	Centrum třídění	Centrum identifikátoru ze seznamu
	Kód sbírky	Kód pro sběr v centru
	Voucher identifikátor	Konkrétní jednoznačný identifikátor pro exemplář
Organismus	Taxonomický název	Organismu, včetně názvu a odkazu na rodokmen
Země	Země	Izolace adresy vzorku
Funkce anotace	Gen	Symbol genu sekvence
	Biologické funkce	Název, specifický rys (např. CDS)
PCR primery	Jméno dopředného primeru	Jméno dopředného primeru
	Sekvence dopředného primeru	Sekvence dopředného primeru
	Jméno reverzního primeru	Jméno reverzního primeru
	Sekvence reverzního primeru	Sekvence reverzního primeru

Tabulka 5: Doporučené pole

Skupina	Pole	Popis
Detaily vzorků	Zeměpisná šířka a délka	Souřadnice místa odběru vzorků
	Identifikovány	Jméno výzkumníka, který vzorek identifikoval
	Shromažďují	Jméno výzkumníka vzorku, který vzorek shromáždil
	Datum odběru	Datum odběru vzorku

5.2. Vyhledávání sekvencí v databázích

Způsobů, jak získat požadovaná data z databází, je několik. Každá sekvence má svůj unikátní identifikátor, podle kterého ji můžeme jednoduše vyhledat. Další možností je hledání podle klíčových slov. Mezi dalšími vyhledávacími kritérii mohou být druh organismu, autor, rok, místo původu a další.

5.3. Formáty dat

5.3.1. FASTA

V dnešní době je nejuniverzálnějším formátem pro práci s biologickými daty. Hodí se jak pro práci s nukleotidovými, tak i aminokyselinovými sekvencemi. Není příliš vhodný pro archivaci. Na rozdíl od vnitřních formátů databází neumožňuje uložit dodatečné informace pro databázové vyhledávání. To je na druhou stranu jeho výhoda při práci, jelikož se dá zapsat přímo z klávesnice. Soubor ve formátu FASTA je textovým souborem, který začíná na prvním řádku znakem > (větší než). Za tímto znakem následuje „hlavička“, ve které je obsažen název sekvence, anotace a různé další údaje, které nejsou obsažené ve vlastní sekvenci, jako například zdrojová databáze apod. Nejdůležitější částí hlavičky je identifikátor, který je reprezentován skupinou alfanumerických znaků hned za znakem >. Tento identifikátor je jedinečný a musí být v hlavičce zahrnut, další informace lze už považovat za volitelné. Za hlavičkou pak následuje vlastní sekvence ve formě surových dat, ta by neměla obsahovat mezery či prázdné řádky. Velkou výhodou FASTA formátu je možnost spojit do jednoho souboru více sekvencí, které pak mohou být zpracovány naráz. Podmínkou ale je, aby byly všechny sekvence buď nukleotidové, anebo aminokyselinové, jelikož FASTA neumožňuje přímo specifikovat typ sekvence. Nepřímo, což některé programy vyžadují, je to možné pomocí koncovky textového souboru, kde `*.nt` označuje nukleotidovou sekvenci a `*.aa` sekvenci aminokyselin. Dalšími koncovkami mohou být např: `*.fa`, `*.fas`, `*.fasta`, `*.fsa` a další, neboť neexistuje žádný standard. [17]

Identifikátory v hlavičce FASTA pro nejpoužívanější databáze:

GenBank	gi | gi-number | gb | accession | locus
EMBL Data Library	gi | gi-number | emb | accession | locus
DDBJ, DNA Database of Japan	gi | gi-number | dbj | accession | locus
NBRF PIR	pir | | entry
Protein Research Foundation	prf | | name
SWISS-PROT	sp | accession | name

6. Programové řešení

6.1. Funkce

Jedním z cílů práce bylo vytvořit funkci pro výpočet denzity DNA sekvence a funkci pro porovnání denzity dvou sekvencí distanční metodou.
Dále bylo vytvořeno intuitivní grafické rozhraní, které slouží k uživatelsky nenáročné práci s těmito funkcemi.
Jednotlivé funkce a vlastní program pro tuto bakalářskou práci jsou vytvořeny v programovém prostředí MATLAB R2010b, version 7.11.0.

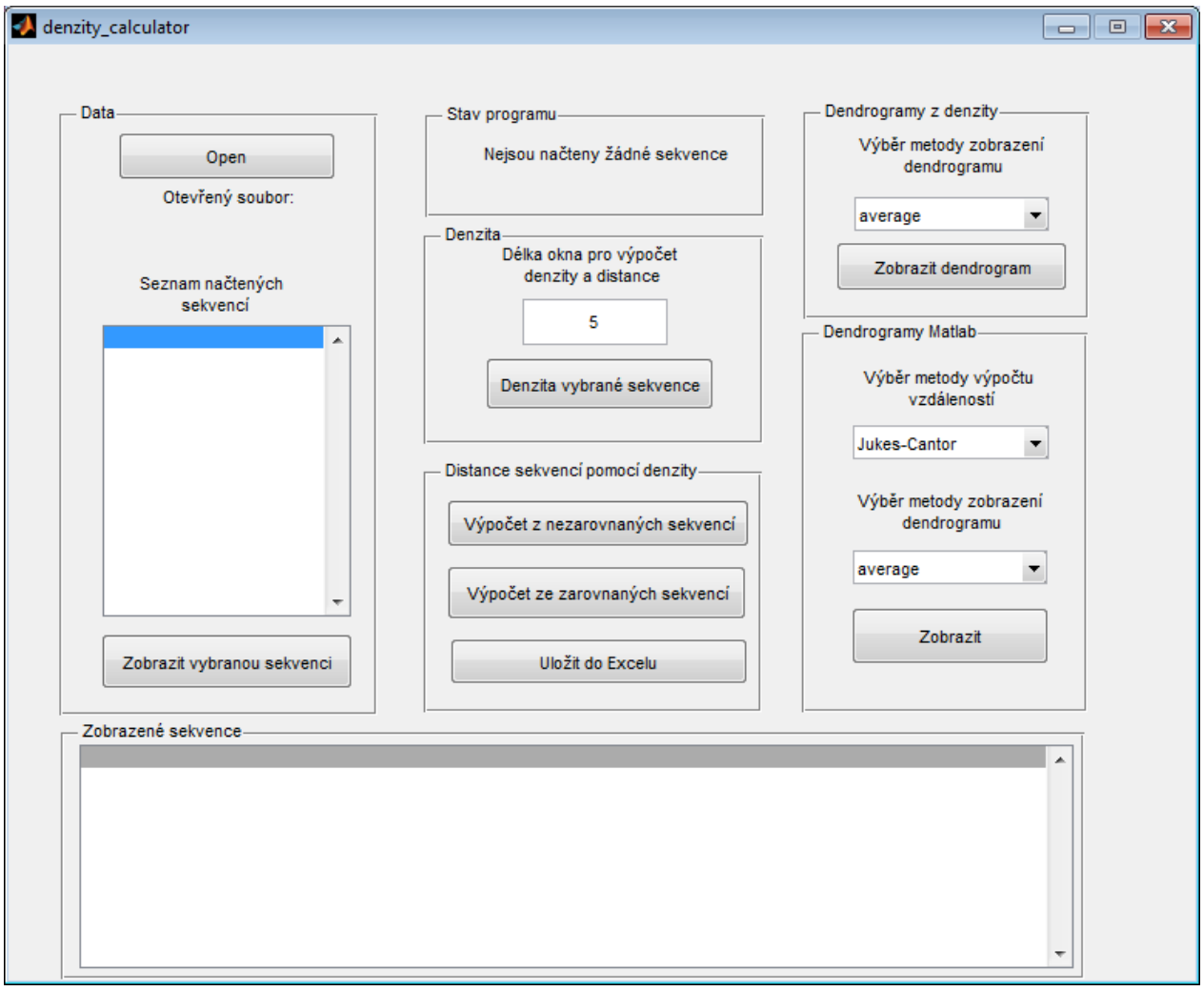
6.1.1. Funkce pro výpočet denzity DNA jednotlivých sekvencí

Funkce pro výpočet denzity zadané funkce se jmenuje *denzita_dna.m*. Tato funkce má 2 vstupy a to vstup ***sekvence,*** který načte zkoumanou sekvenci, a druhý vstup ***delkaokna***, která určuje, jak velké okno bude danou sekvenci „projíždět“. Tato funkce převede danou sekvenci do numerického formátu, a to způsobem, že za Adenin dosadí 1, za Cytosin 2, za Guanin 3 a za Thymin 4. Poté pomocí cyklu for projíždí sekvenci okno zadané délky a počítá denzitu tak, že zprůměruje hodnoty v okně (všechny sečte a vydělí délkou okna).
Výstupem je matice hodnot ***denzita***, která je pak vykreslena do grafů jak samostatně pro každou bázi, tak součtem komplementárních bází.

6.1.2. Funkce pro porovnání denzity dvou sekvencí distanční metodou

Funkce *euklid.m* porovnává denzity 2 zadaných sekvencí. Má 2 vstupy b a d, přičemž prvním vstupem je denzita jedné a druhým denzita druhé sekvence. Rozdílnost je počítána pomocí euklidovské vzdálenosti pomocí vzorce: $\rho(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, kde *n* je délka sekvencí, *x* je denzita jedné sekvence na pozici *i* a *y* denzita druhé sekvence na pozici *i*.
Hodnoty jsou zobrazeny v grafu. Výstupem je proměnná ***rozdílnost,*** která je součtem rozdílnosti jednotlivých částí sekvence.

6.2. Uživatelské prostředí programu



Obrázek 5.: Okno programu

Program je rozdělen do 7 bloků, které sdružují podobné funkce programu.

1. ***Stav programu***

- V poli Stav programu lze sledovat funkci programu a stav, ve kterém se právě nachází. Při chybě se zobrazí text obsahující zhodnocení situace a uživatelskou nápovědu pro odstranění chyby.

2. ***Data***

- Blok pro práci s daty slouží k načtení souboru se sekvencemi, zobrazení názvů sekvencí a jejich výběru pro další práci – zobrazení sekvence nebo výpočet její denzity.
- **Open** – slouží k otevření souboru dat ve formátu *.m nebo *.mat
- **Zobrazit vybranou sekvenci** – zobrazí vybranou sekvenci z názvu sekvencí

3. ***Zobrazení sekvence***

- Blok, ve kterém se zobrazují vybrané sekvence v základním formátu.

4. ***Denzita***

- Blok, ve kterém se pracuje s denzitou – její výpočet pro vybranou sekvenci a nastavení délky okna, které pak slouží i pro další bloky.
- **Délka okna** – slouží pro nastavení délky okna pro výpočet denzity a distance.
- **Denzita vybrané sekvence** – výpočet denzity vybrané sekvence a její zobrazení v grafu (v novém okně).

5. ***Distance sekvencí pomocí denzity***

- Blok, ve kterém se pracuje s distancí – její výpočet pomocí denzity pro všechny načtené sekvence a uložení tabulky vzdáleností v Excelu.
- **Výpočet z nezarovnaných sekvencí** – výpočet denzity a distancí z nezarovnaných sekvencí.
- **Výpočet ze zarovnaných sekvencí** – výpočet denzity a distancí ze zarovnaných sekvencí.
- **Uložit do Excelu** – uložení distancí do Excelovské tabulky.

6. ***Dendrogramy z denzity***

- Blok, který slouží k zobrazení dendrogramu z dříve vypočtených distancí.
- **Výběr metody zobrazení dendrogramu.**
- **Zobrazit dendrogram.**

7. ***Dendrogramy Matlab***

- Blok, který slouží k výpočtu vzdálenosti sekvencí pomocí algoritmů programu Matlab a zobrazení jejich dendrogramu.
- **Výběr metody výpočtu vzdálenosti –** metody pro výpočet vzdálenosti sekvencí.
- **Výběr metody zobrazení dendrogramu.**
- **Zobrazit** – zobrazení dendrogramu.

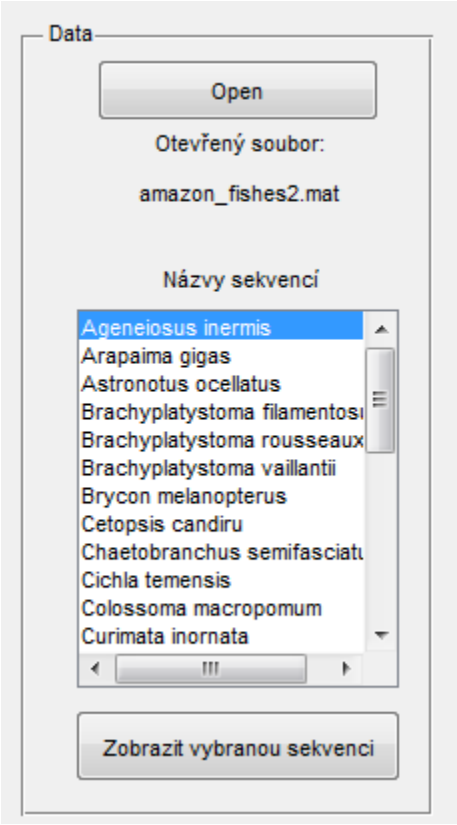
7. Stručný popis a průběh při analýze

Aplikace denzity_calculator je software určený pro práci se sekvencemi nukleotidů. Mezi hlavní funkce pro práci s genomickými daty patří výpočet denzity vybraných sekvencí, výpočet distancí mezi sekvencemi v načteném souboru dat a zobrazení fylogenetických stromů různými metodami.

Vstupním souborem je soubor s příponou .mat (MATLAB file), obsahující dvě proměnné – Names a Sequences.

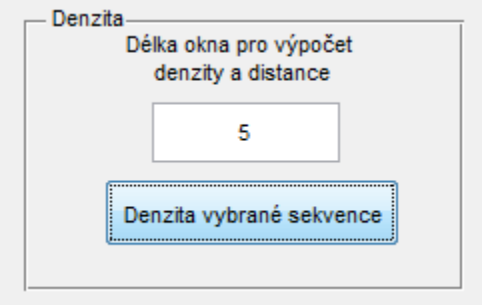
Pro přípravu takového souboru ze sekvencí ve formátu FASTA je vhodné použít speciální program. V této práci k tomu posloužil již dříve vytvořený program v rámci předmětu APBI.

Výchozí soubor se do programu nahrává pomocí tlačítka Open. Po stisku tlačítka je spuštěno dialogové okno pro výběr požadovaného *.mat nebo *.m souboru.



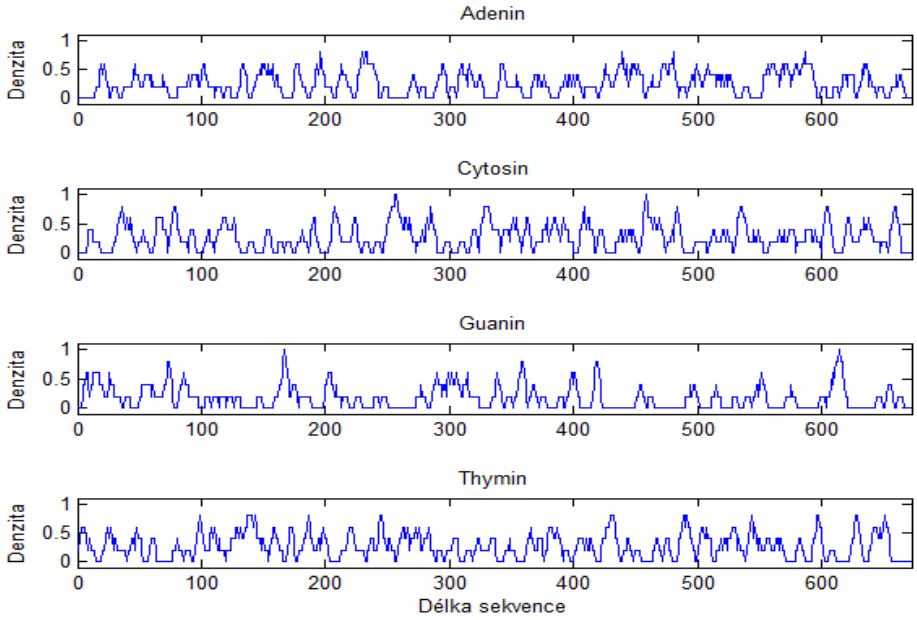
Obrázek 6.: Načtení souboru

Jeho název se pak objeví pod tlačítkem Open. Pokud soubor obsahuje povinné proměnné (Names a Sequences), je obsah proměnné Names zobrazen v listboxu Seznam načtených sekvencí. V tomto listboxu také vybíráme sekvence pro další práci. Vybranou sekvenci můžeme zobrazit v základním tvaru („ACTG“) v listboxu Zobrazené sekvence. Dále můžeme pro nastavenou délku okna vypočítat její denzitu a tu pak zobrazit ve 2 grafech – k tomu je využívána externí funkce *denzita_dna*, která je popsána výše. Délka okna musí být lichá, musí být minimálně 5 a maximálně 1/20 délky sekvence.



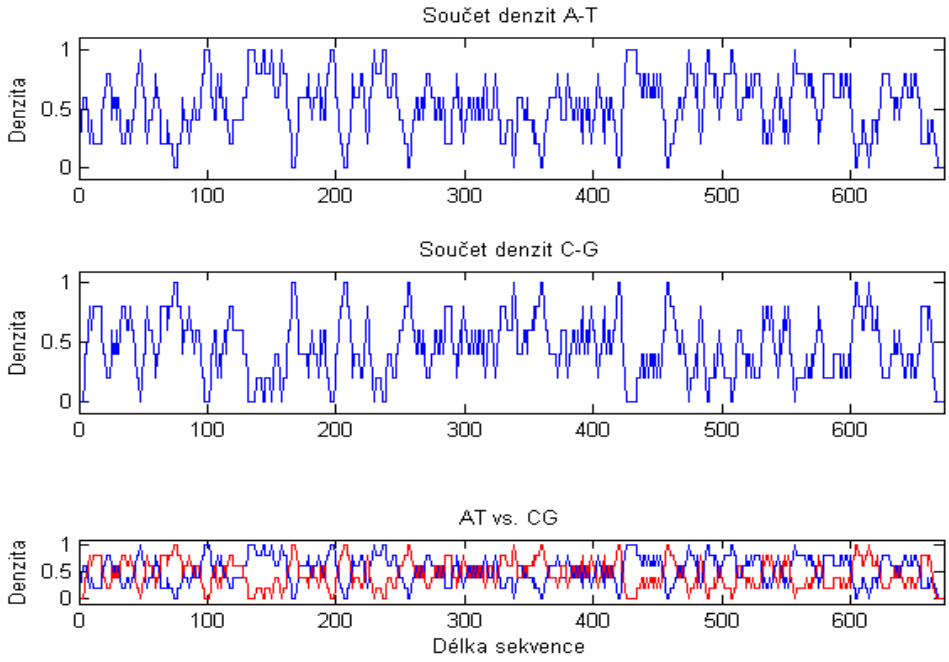
Obrázek 7.: Nastavení délky okna

První graf zobrazuje denzitu jednotlivých nukleotidů (Adenin, Cytosin, Guanin, Thymin) v závislosti na pozici v sekvenci.



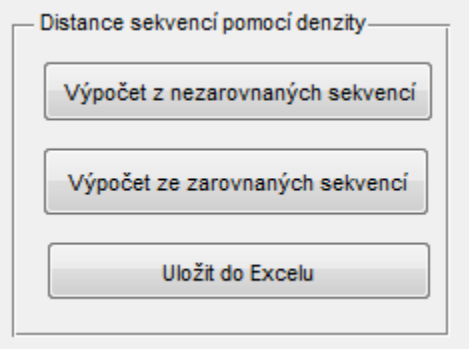
Obrázek 8.: Denzita jednotlivých nukleotidů Ageneiosus inermis

Druhý graf zobrazuje součet denzit komplementárních nukleotidů A-T a C-G a jejich zobrazení v jednom grafu.



Obrázek 9.: Součet denzit komplementárních nukleotidů Ageneiosus inermis

Pro celý soubor načtených dat pak můžeme využít funkce pro výpočet distancí denzit všech sekvencí pomocí funkce *euklid*. Zde můžeme zvolit buď výpočet z nezarovnaných sekvencí, nebo ze zarovnaných, kde jsou nejprve všechny sekvence zarovnány pomocí funkce multialign a teprve pak je počítána denzita a euklidovská distance.



Obrázek 10.: Výpočet distance

Vypočítané denzity pak můžeme uložit do Excelu.

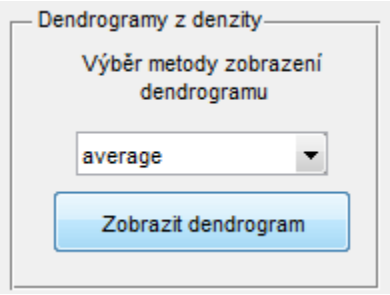
Tabulka 6: Příklad distancí

0	160,0255	159,5336	139,9718	139,3811
160,0255	0	152,887	160,9033	146,5583
159,5336	152,887	0	152,7735	143,5822
139,9718	160,9033	152,7735	0	35,53647
139,3811	146,5583	143,5822	35,53647	0

Zobrazení fylogenetického stromu z námi vypočtených distancí provedeme pomocí tlačítka Zobrazit dendrogram. V rolovacím menu můžeme zvolit jednu z metod sestavení stromu. Na výběr zde máme metody Average, Single, Complete, Centroid, Weightet a Median.

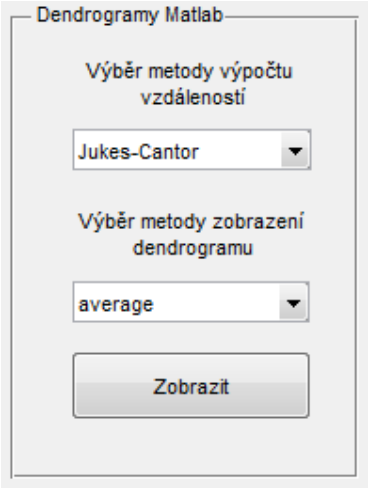
Popis metod:

- Single – spojení pomocí nejbližšího souseda, používá nejmenší vzdálenost mezi objekty dvou skupin.
- Complete – spojení pomocí nejvzdálenějšího souseda, používá největší vzdálenost mezi objekty dvou skupin.
- Average – používá průměrné vzdálenosti mezi všemi páry objektů jakýchkoli dvou skupin.
- Centroid –používá euklidovskou vzdálenost mezi těžištěm dvou skupin.
- Weightet – používá speciální vážený výpočet průměru.
- Median – používá euklidovskou vzdálenost mezi váženými těžišti dvou skupin.



Obrázek 11.: Výběr metody sestavení dendrogramu

Jako doplněk je zde sestavení dendrogramu pomocí funkcí, kterými disponuje MATLAB, kde můžeme pro výpočet vzdáleností zvolit metody Jukes-Cantor, p-distance či alignment-score a pro sestavení stromu je opět na výběr z metod Average, Single, Complete, Centroid, Weightet a Median.



Obrázek 12.: Výběr metody sestavení dendrogramu

8. Diskuze řešení

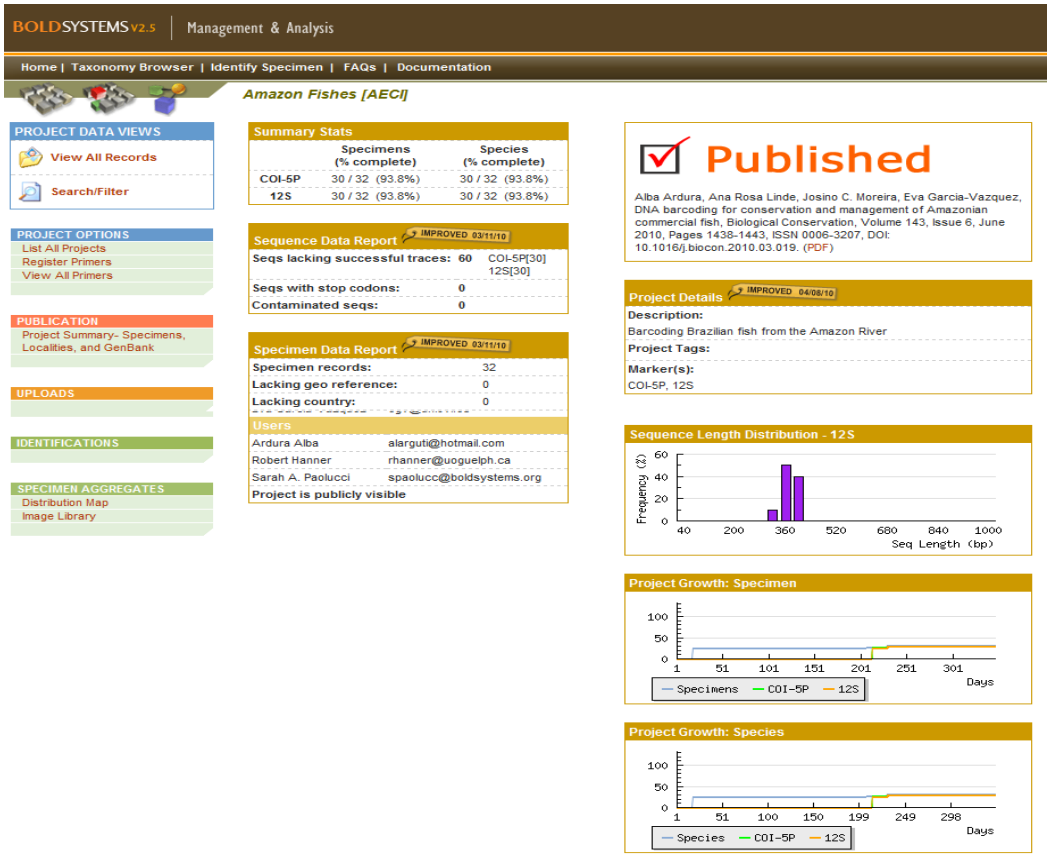
V následující kapitole je popsána analýza souboru 30 druhů ryb (Amazon fishes), získání jejich sekvencí, vytvoření dendrogramů z denzit a jejich komparační analýza.

8.1. Získání sekvencí

Sekvence byly získány z veřejné databáze BOLDsystems, která je dostupná na adrese www.boldsystems.org. Po bezplatné registraci a přihlášení je možné stahovat vybrané sekvence či celé projekty sekvencí. Ze sekce databáze FishBol byla vybrána skupina Amazon Fishes (AECI), která obsahuje 32 záznamů a z toho je u 30 druhů získána sekvence CO1.

Barcoding Fish (FishBOL)	Pub	Specimens	Species	Species with Sequences Markers [stat]	Sequences Markers [stat]	Project Tags
AMNHI African fishes from AMNH Ichthyology		463	183	COI-5P[167], D-loop[23], RAG1[22]	COI-5P[410], D-loop[103], RAG1[94]	
AECI Amazon Fishes	✓	32	32	COI-5P[30], 12S[30]	COI-5P[30], 12S[30]	

Obrázek 13.: Databáze FishBol



Obrázek 14.: Data Amazon fishes

Identification	Specimen Page	Sequence Page	Length [Ambig] COI-5P	12S
Ageneiosus inermis	Bocudo	AECI004-10	662 [0n]	356 [0n]
Arapaima gigas	Pirarucu	AECI022-10	651 [0n]	344 [0n]
Astronotus ocellatus	Acara-acu	AECI001-10	648 [0n]	324 [0n]
Brachyplatystoma filamentosum	Filhote	AECI011-10	652 [0n]	371 [0n]
Brachyplatystoma rousseauxii	Dourada	AECI010-10	655 [0n]	344 [0n]
Brachyplatystoma vaillantii	Piramutaba	AECI020-10	605 [0n]	0
Brycon melanopterus	Matrinxa	AECI014-10	652 [0n]	328 [0n]
Cetopsis candiru	Candiru	AECI007-10	679 [0n]	371 [0n]
Chaetobranchopsis orbicularis	Acara branco	AECI030-11	0	393 [0n]
Chaetobranchus semifasciatus	Acara	AECI028-11	655 [0n]	379 [0n]
Cichla temensis	Tucunare	AECI026-10	645 [0n]	337 [0n]
Colossoma macropomum	Tambaqui	AECI025-10	651 [0n]	349 [0n]
Curimata inornata	Branquinha	AECI006-10	656 [0n]	0
Geophagus proximus	Acara reco	AECI027-10	641 [0n]	366 [0n]
Heros efasciatus	Acara preto	AECI031-11	0	364 [0n]
Hoplarchus psittacus	Acara azulao	AECI029-11	655 [0n]	337 [0n]
Leporinus piau	Piau	AECI018-10	652 [0n]	365 [0n]
Mylossoma duriventre	Pacu	AECI015-10	674 [0n]	336 [0n]
Osteoglossum bicirrhosum	Aruana	AECI003-10	674 [0n]	339 [0n]
Oxydoras niger	Cuiu-cuiu	AECI008-10	649 [0n]	329 [0n]
Phractocephalus hemioliopterus	Pirarara	AECI017-10	657 [0n]	318 [0n]
Piaractus brachypomus	Pirapichinga	AECI021-10	635 [0n]	378 [0n]
Pimelodus blochii	Mandi	AECI013-10	653 [0n]	379 [0n]
Plagioscion squamosissimus	Pescada	AECI016-10	656 [0n]	350 [0n]
Prochilodus nigricans	Curimata	AECI009-10	669 [0n]	315 [0n]
Pseudoplatystoma fasciatum	Surubim	AECI024-10	639 [0n]	361 [0n]
Pseudoplatystoma tigrinum	Pintado	AECI019-10	651 [0n]	324 [0n]
Pterygoplichthys joselimaianus	Acari bodo	AECI005-10	653 [0n]	378 [0n]
Satanoperca lilith	Acara tucunare	AECI032-11	645 [0n]	382 [0n]
Schizodon fasciatus	Aracu	AECI002-10	652 [0n]	315 [0n]
Semaprochilodus insignis	Jaraqui	AECI012-10	649 [0n]	351 [0n]
Triportheus elongatus	Sardinha	AECI023-10	605 [0n]	356 [0n]

Obrázek 15.: Délky sekvencí jednotlivých druhů

8.2. Výpočet distancí

Distance neboli euklidovské vzdálenosti, které nám program vypočítá, slouží k porovnání podobnosti/rozdílnosti zadaných sekvencí. Jsou to vlastně sumy rozdílností jednotlivých částí sekvence a z toho vyplývá, že čím je číslo vyšší, tím jsou vybrané sekvence rozdílnější. Naopak nula znamená, že se jedná o naprosto identické sekvence. V tabulce 7 vidíme příklady distancí pro různé délky okna. Na diagonále je vidět nula, což znamená, že je do kříže porovnána stejná sekvence.

Tabulka 7: Příklady vypočtených distancí pro okno délky 5 a 25

Délka okna 5	1.	2.	3.	4.	5.
1. Ageneiosus inermis	0	160,0255	159,5336	139,9718	139,3811
2. Arapaima gigas	160,0255	0	152,887	160,9033	146,5583
3. Astronotus ocellatus	159,5336	152,887	0	152,7735	143,5822
4. Brachyplatystoma filamentosum	139,9718	160,9033	152,7735	0	35,53647
5. Brachyplatystoma rousseauxii	139,3811	146,5583	143,5822	35,53647	0

Délka okna 25	1.	2.	3.	4.	5.
1. Ageneiosus inermis	0	90,97646	85,52837	80,33838	78,74768
2. Arapaima gigas	90,97646	0	84,89931	90,49691	81,25972
3. Astronotus ocellatus	85,52837	84,89931	0	88,0868	80,3339
4. Brachyplatystoma filamentosum	80,33838	90,49691	88,0868	0	31,38999
5. Brachyplatystoma rousseauxii	78,74768	81,25972	80,3339	31,38999	0

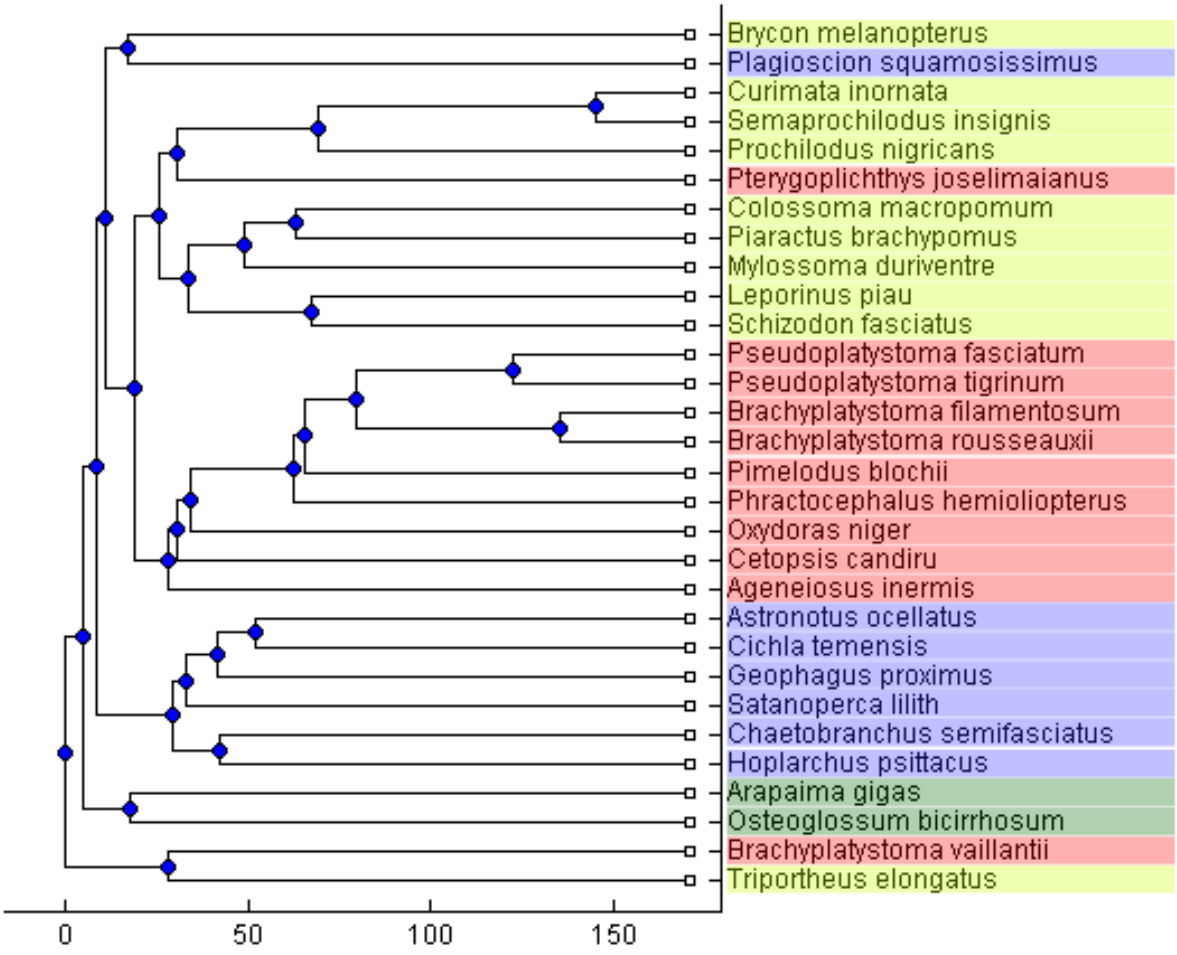
8.3. Dendrogramy

Následující tabulka 8 zobrazuje soubor 30 ryb Amazon fishes a jejich zařazení do řádu a čeledi. Řády jsou pak pro lepší přehlednost barevně odlišeny jak v tabulce, tak ve vytvořených dendrogramech. Druhy jsou rozděleny do 4 řádů, a to tak, že žlutě je označen řád Characiformes, zeleně Osteoglossiformes, modře Perciformes a červeně Siluriformes.

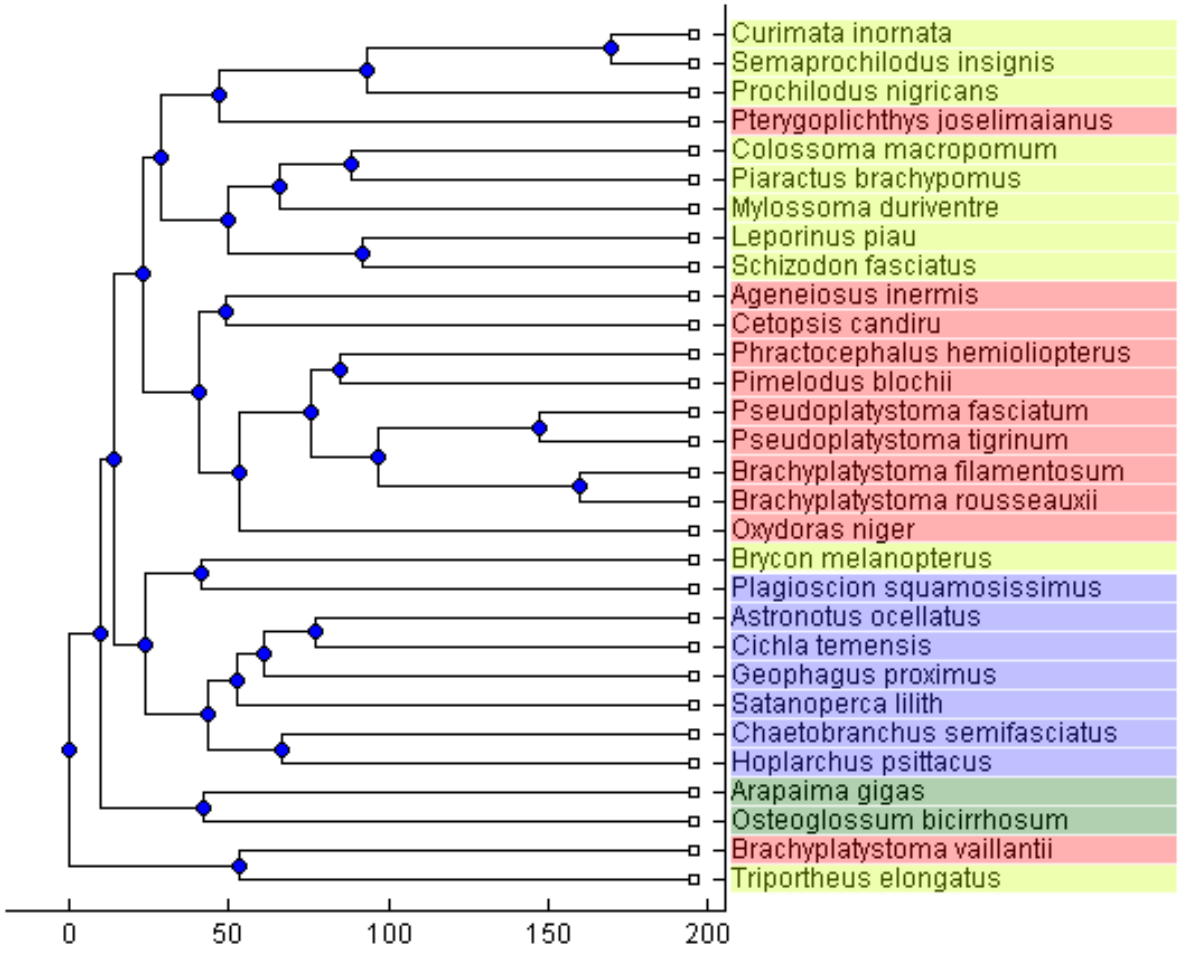
Tabulka 8: Seznam druhů Amazon fishes a jejich systematické zařazení

Druh (Species)	Řád (Order)	Čeleď (Family)
Brycon melanopterus	Characiformes	Characidae
Colossoma macropomum	Characiformes	Characidae
Curimata inornata	Characiformes	Curimatidae
Leporinus piau	Characiformes	Anostomidae
Mylossoma duriventre	Characiformes	Characidae
Piaractus brachypomus	Characiformes	Characidae
Prochilodus nigricans	Characiformes	Prochilodontidae
Semaprochilodus insignis	Characiformes	Prochilodontidae
Schizodon fasciatus	Characiformes	Anostomidae
Triportheus elongatus	Characiformes	Characidae
Arapaima gigas	Osteoglossiformes	Osteoglossidae
Osteoglossum bicirrhosum	Osteoglossiformes	Osteoglossidae
Astronotus ocellatus	Perciformes	Cichlidae
Cichla temensis	Perciformes	Cichlidae
Geophagus proximus	Perciformes	Cichlidae
Hoplarchus psittacus	Perciformes	Cichlidae
Chaetobranchus semifasciatus	Perciformes	Cichlidae
Plagioscion squamosissimus	Perciformes	Sciaenidae
Satanoperca lilith	Perciformes	Cichlidae
Ageneiosus inermis	Siluriformes	Auchenipteridae
Brachyplatystoma filamentosum	Siluriformes	Pimelodidae
Brachyplatystoma rousseauxii	Siluriformes	Pimelodidae
Brachyplatystoma vaillantii	Siluriformes	Pimelodidae
Cetopsis candiru	Siluriformes	Cetopsidae
Oxydoras niger	Siluriformes	Doradidae
Phractocephalus hemioliopterus	Siluriformes	Pimelodidae
Pimelodus blochii	Siluriformes	Pimelodidae
Pseudoplatystoma fasciatum	Siluriformes	Pimelodidae
Pseudoplatystoma tigrinum	Siluriformes	Pimelodidae
Pterygoplichthys joselimaianus	Siluriformes	Loricariidae

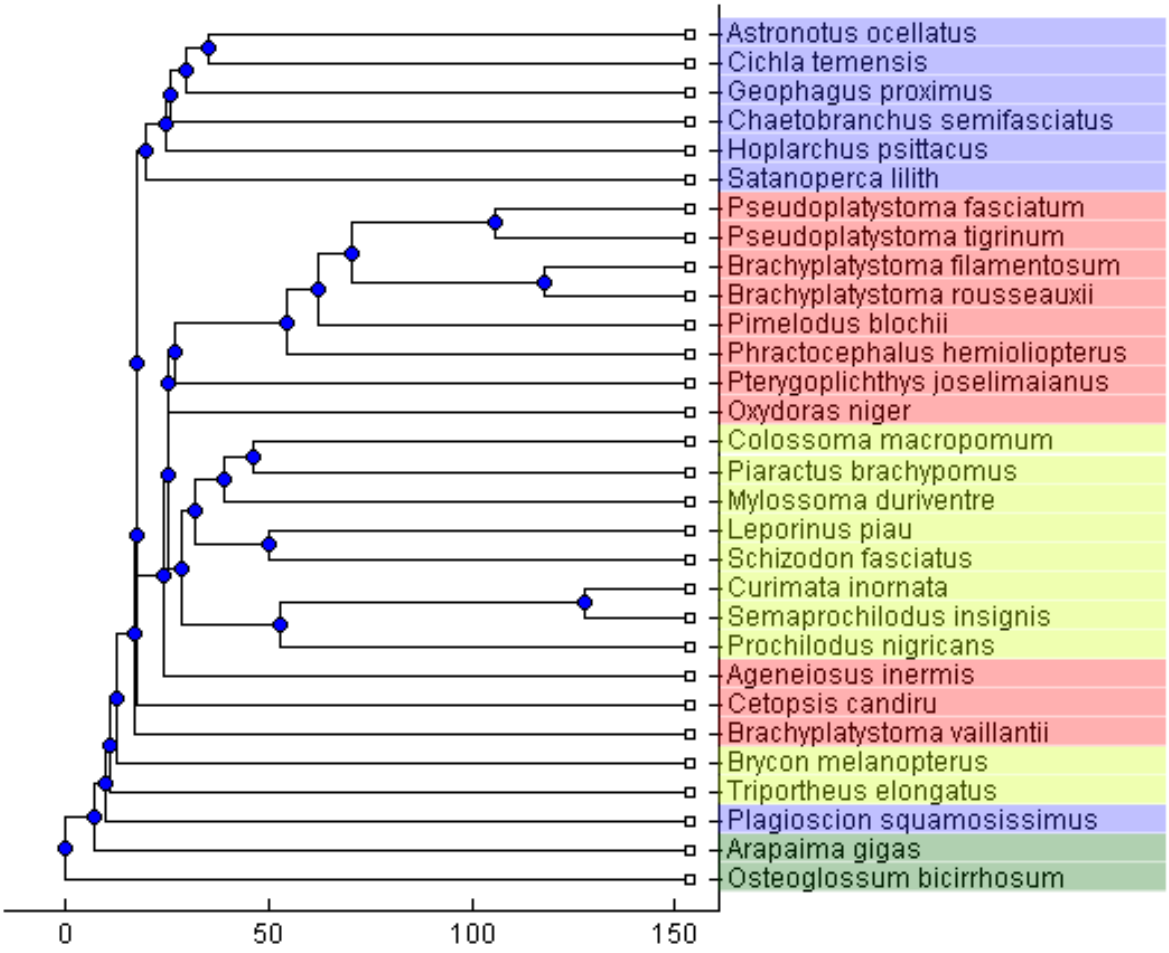
Na následujících obrázcích můžeme vidět dendrogramy vytvořené ze souboru dat Amazon fishes a sestavené pomocí metod average, single a complete. Denzita zde byla vypočítána pro různé délky okna, a to 5, 15 a 25.



Obrázek 16.: Dendrogram - délka okna 5, metoda sestavení - average



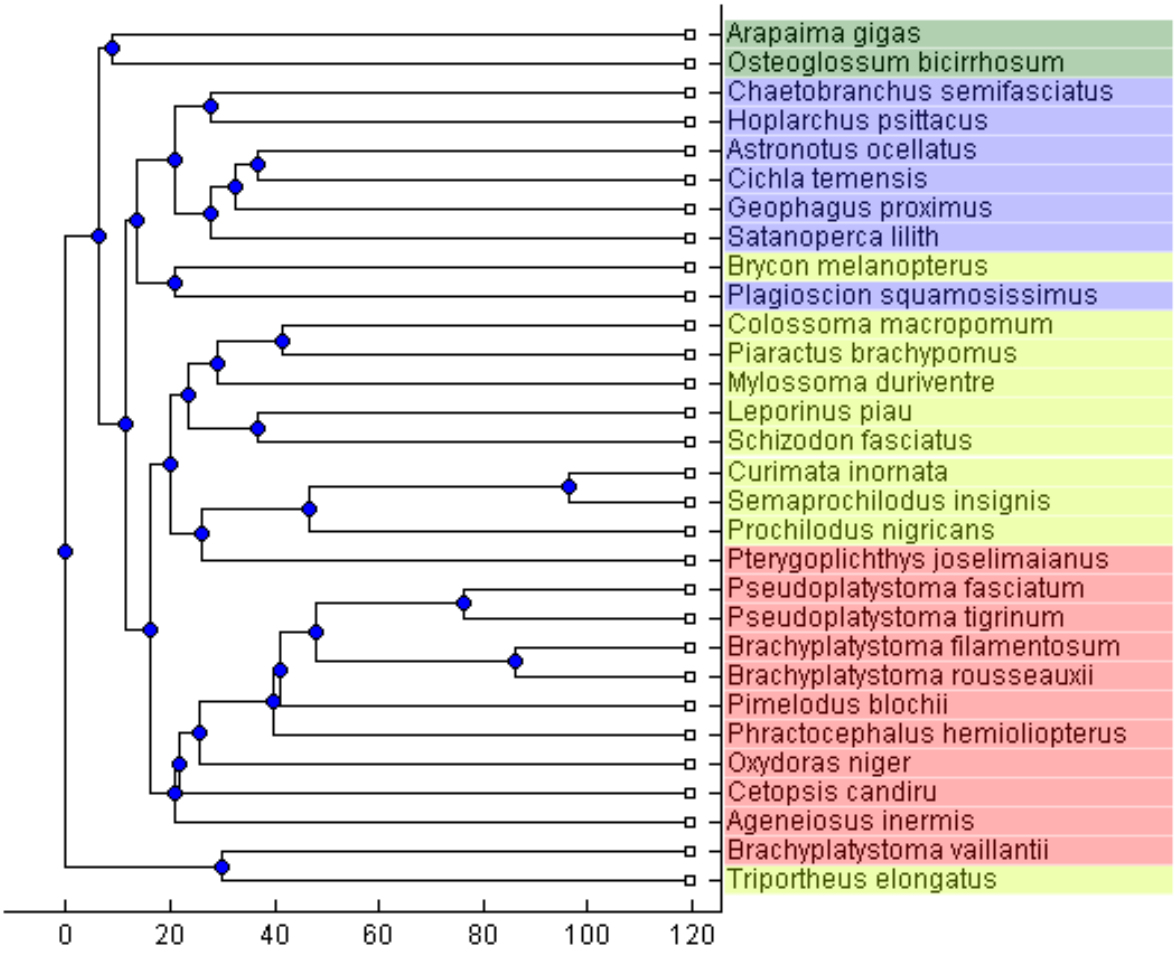
Obrázek 17.: Dendrogram - délka okna 5, metoda sestavení complete



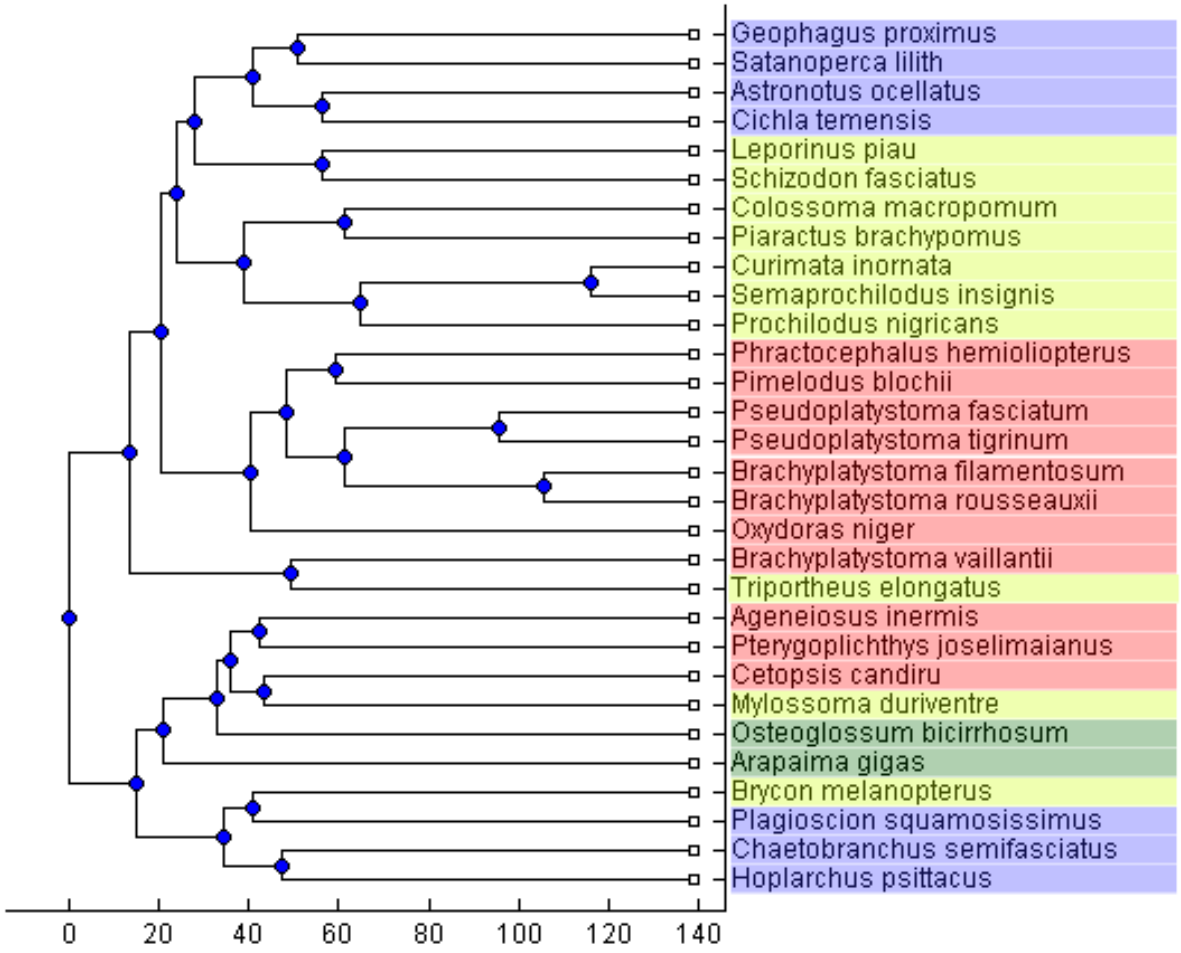
Obrázek 18.: Dendrogram - Délka okna 5, metoda sestavení single

Dendrogramy sestavené z denzit s délkou okna 5, až na některé výjimky, k sobě správně přiřadily jednotlivé řády. Mezi tyto výjimky patří Plagioscion squamosissimus, který není ve všech třech metodách přiřazen mezi řád perciformes. To může být způsobeno tím, že všechny ostatní druhy řádu Perciformes patří na rozdíl od tohoto druhu ke stejné čeledi. Dále se od ostatních stejného řádu osamostatňuje Pterygoplichthys joselimaianus a Brachyplatystoma vaillantii, přičemž poslední zmíněná má o 50 párů bází kratší sekvenci, což po zarovnání může být důvodem odlišnosti.

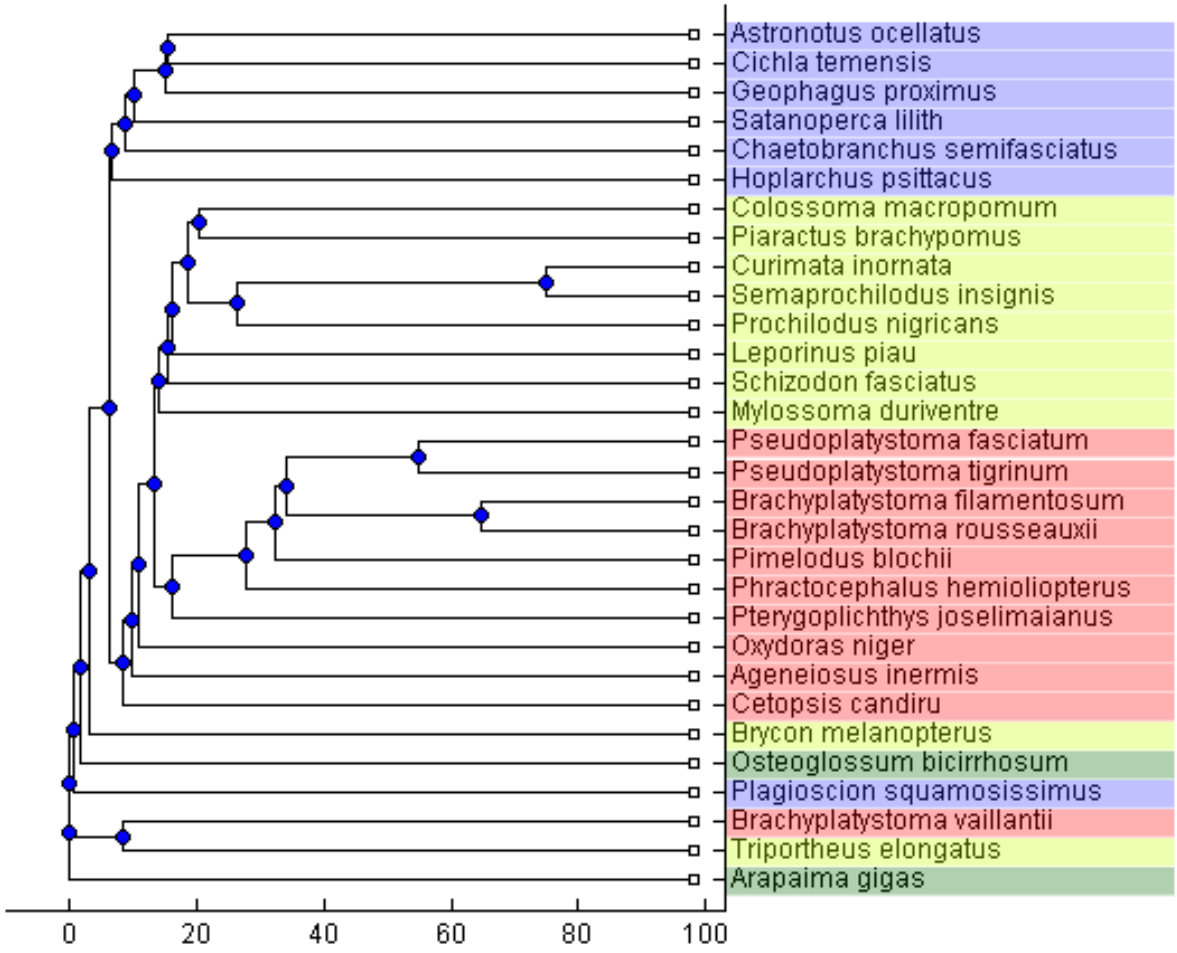
K dobrým výsledkům lze přičíst, že v dendrogramu jsou podobné druhy přiřazeny k sobě. Jsou to ty, které mají stejné rodové jméno a liší se jen jménem druhovým. Zde vidíme takto přiřazené například druhy Pseudoplatystoma či Brachyplatystoma.



Obrázek 19.: Dendrogram - délka okna 15, metoda sestavení average



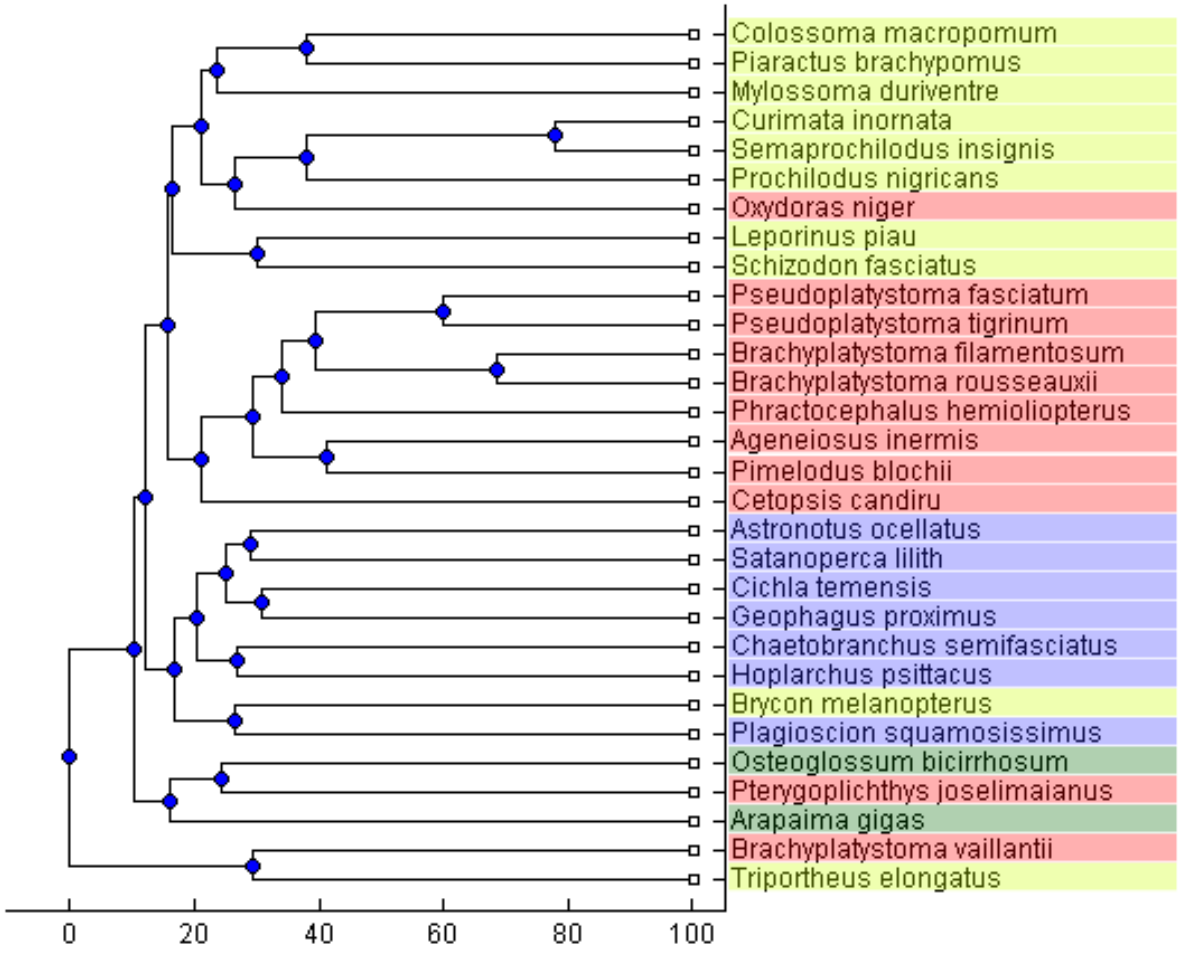
Obrázek 20.: Dendrogram - délka okna 15, metoda sestavení complete



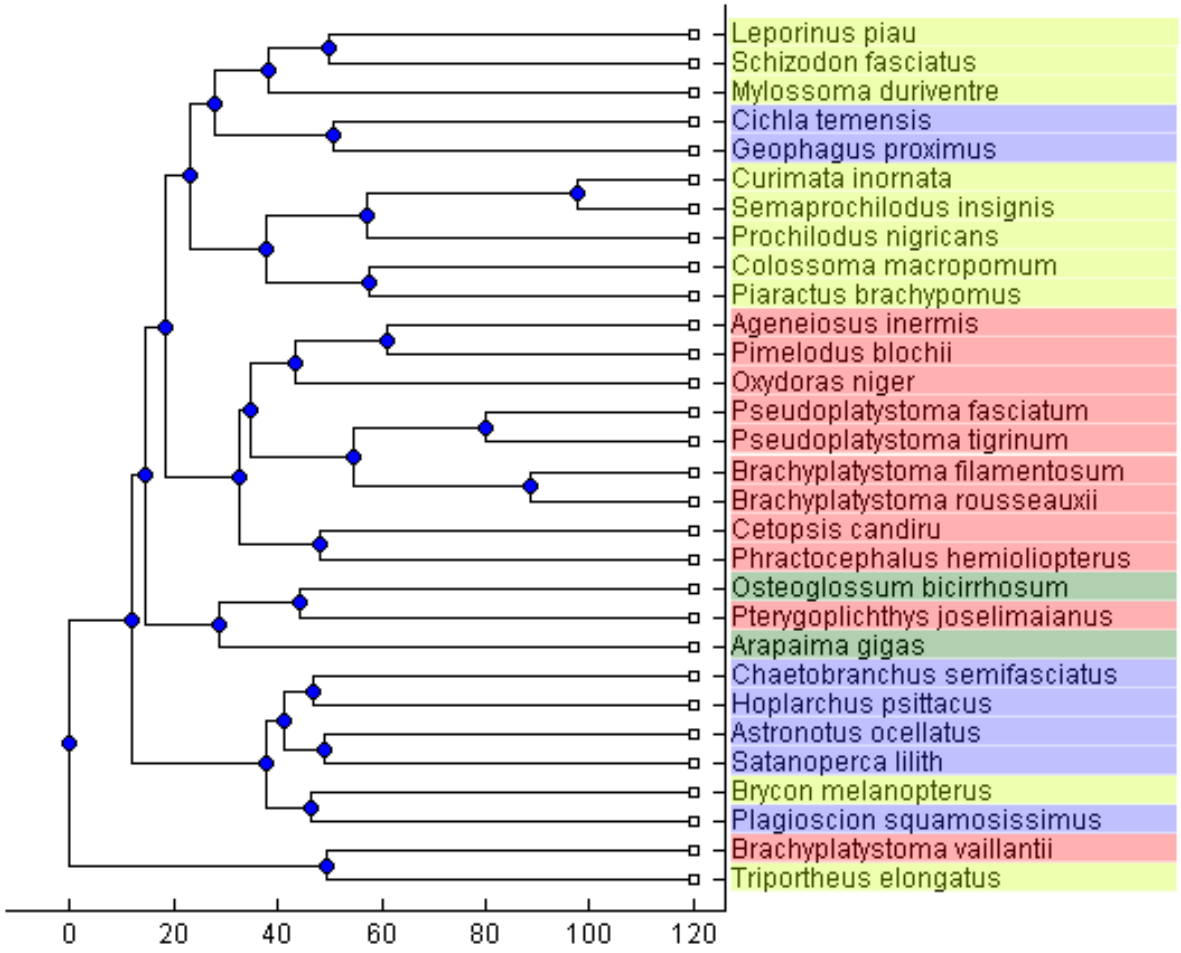
Obrázek 21.: Dendrogram - délka okna 15, metoda sestavení single

Dendrogramy sestavené z denzit s délkou okna 15 měly asi nejlepší výsledky. Od svého řádu Characiformes se nejvíce vzdaloval druh Triportheus elongatus. To může být způsobeno opět kratší sekvencí o 50 bází oproti ostatním druhům stejného řádu. Dále se vzdaloval Plagioscion squamosissimus, který se, jak už bylo vysvětleno dříve, řadí jako jediný k jiné čeledi v rámci řádu.

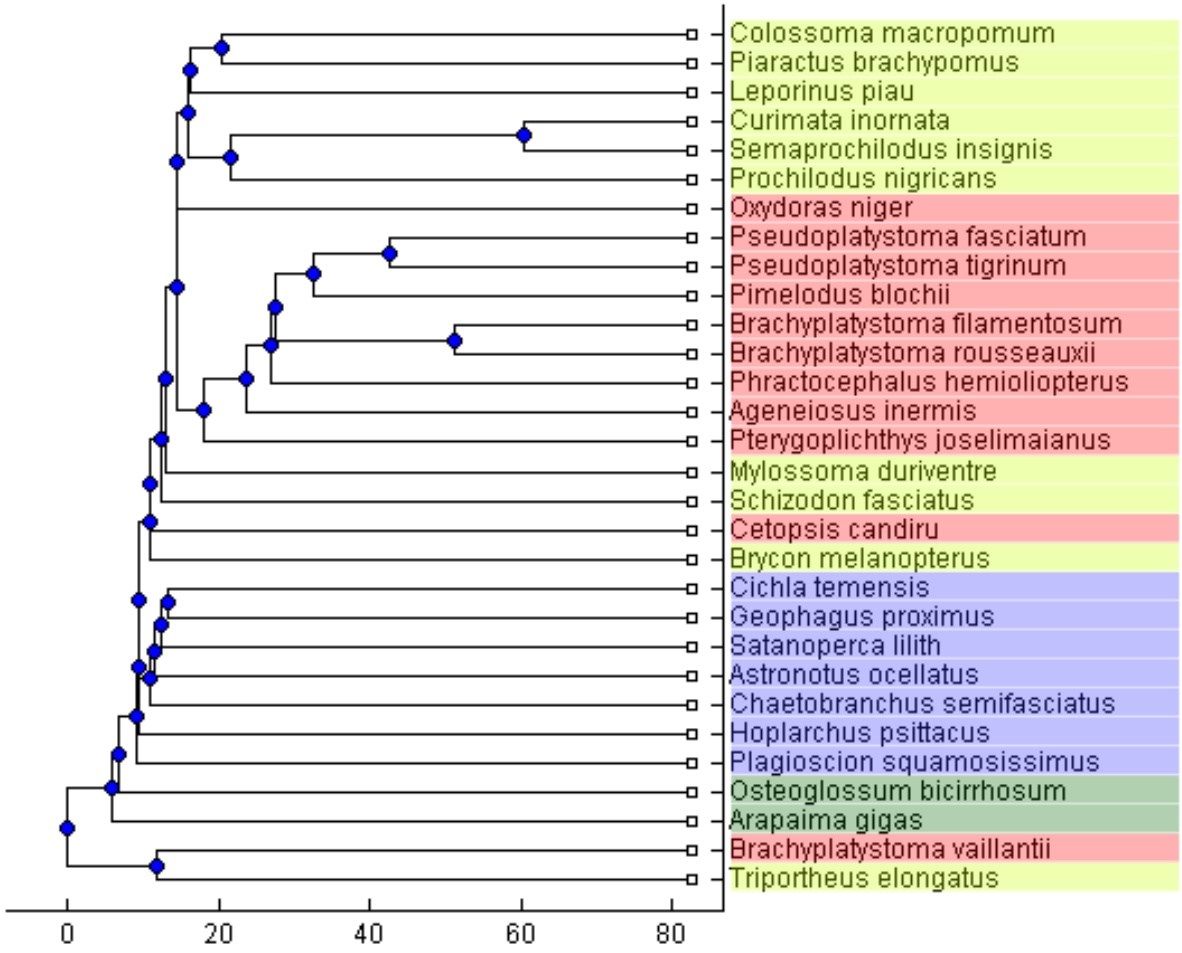
Velmi blízké druhy se stejným rodovým jménem Pseudoplatystoma jsou zde opět přiřazeny ke stejné větvi.



Obrázek 22.: Dendrogram - délka okna 25, metoda sestavení average

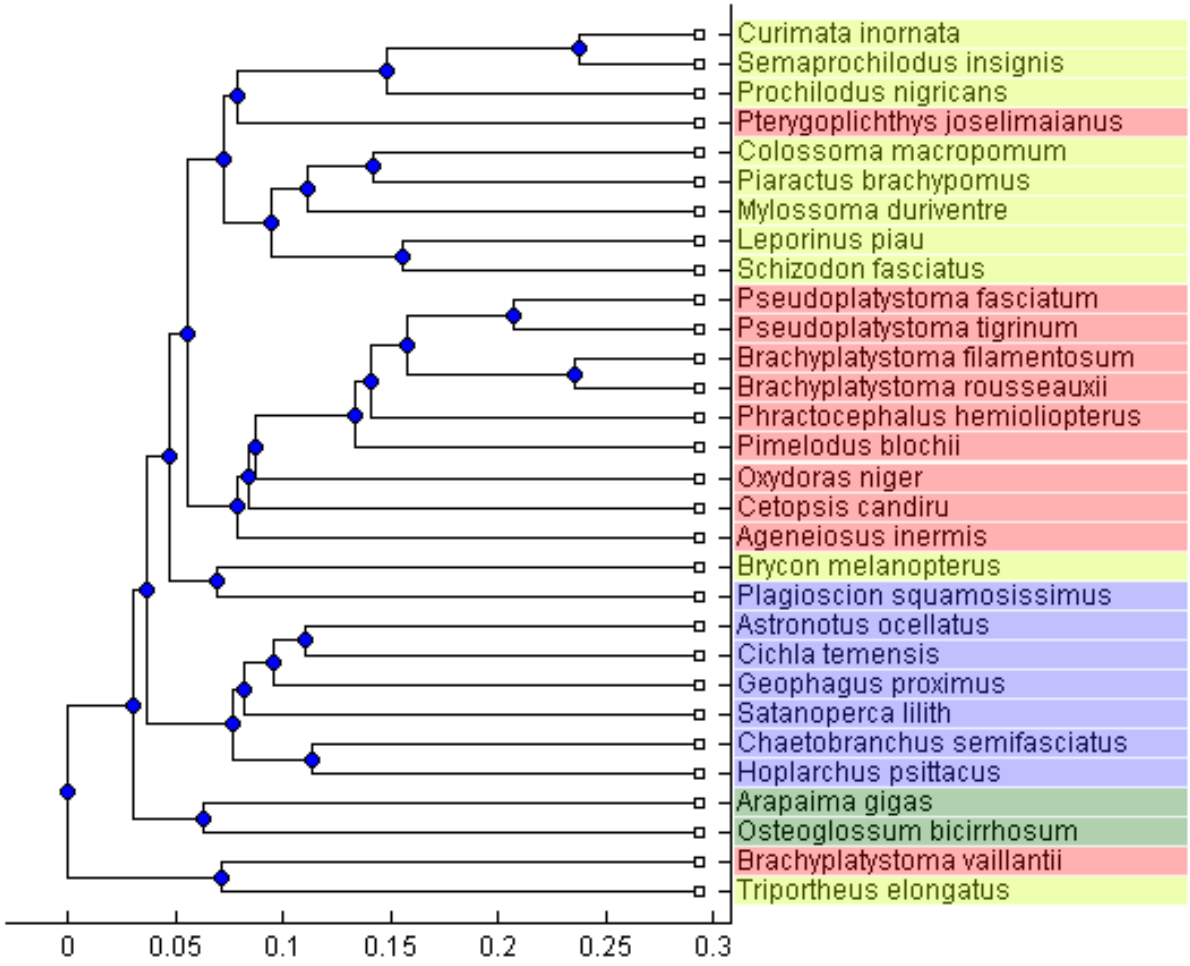


Obrázek 23.: Dendrogram - délka okna 25, metoda sestavení complete



Obrázek 24.: Dendrogram - délka okna 25, metoda sestavení single

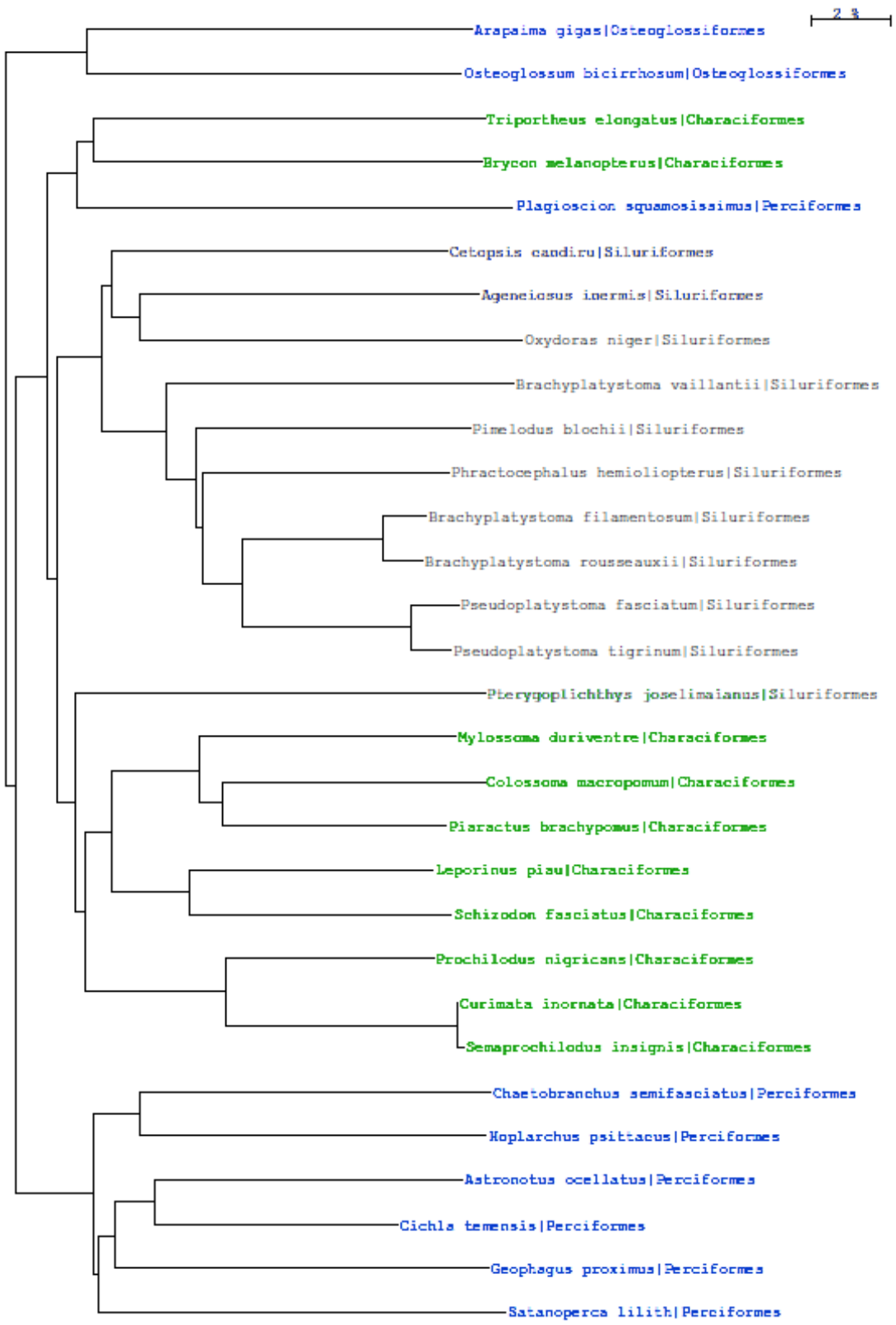
Pro délku okna 25 je rozdělení na první pohled o něco horší než pro délku okna 15. Objevují se zde ale podobně chybná umístění druhů. Druh Pterygoplichthys joselimaianus je zde při metodách average a complete chybně zařazen mezi druhy řádu Osteoglossiformes. A druh Brycon melanopterus je zde opět z neznámého důvodu odloučen od ostatních druhů řádu Characiformes.



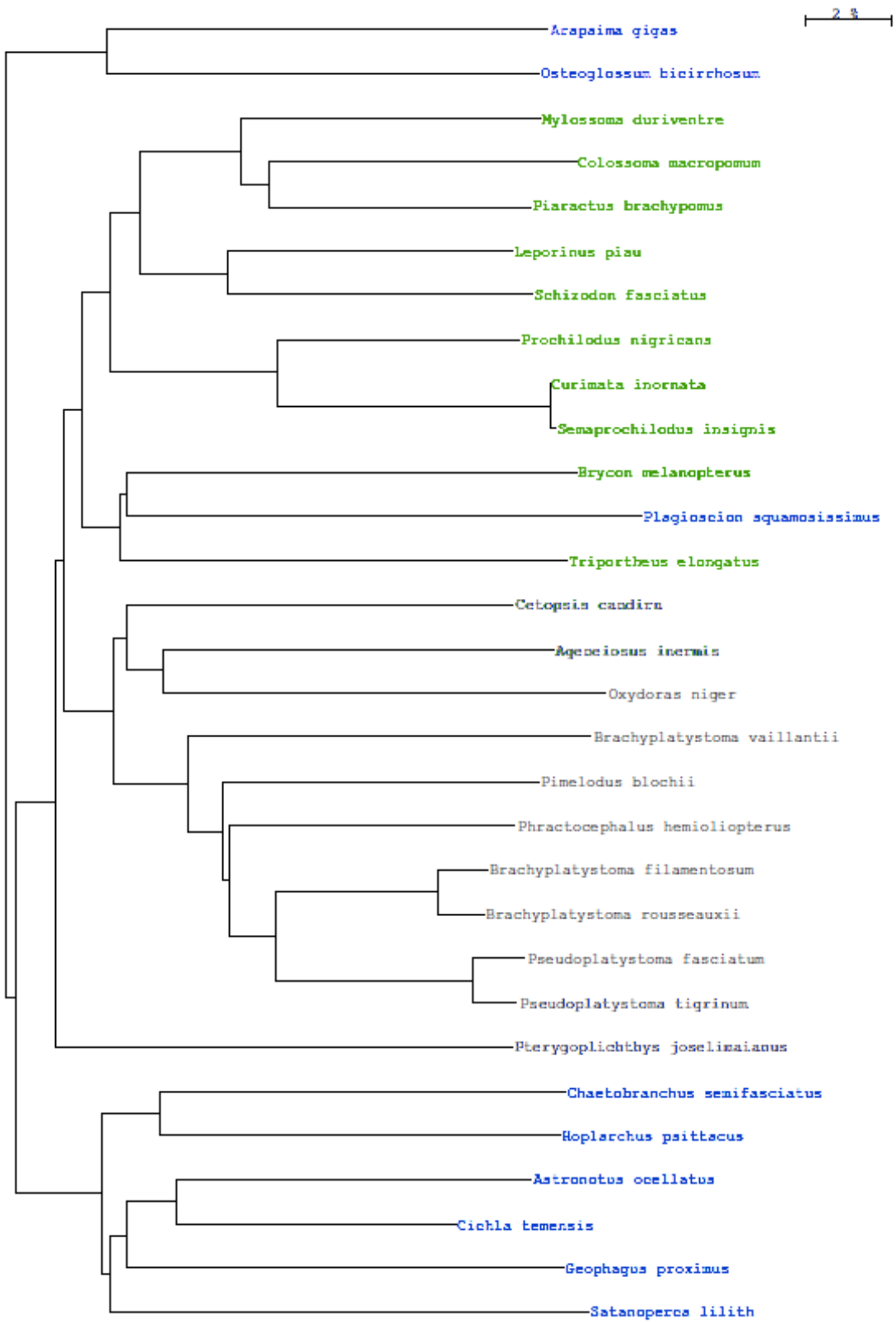
Obrázek 25.: Dendrogram - metoda výpočtu Jukes-Cantor, metoda sestavení average

Na obrázku 25 je dendrogram sestavený pomocí interních funkcí Matlabu, a to metodou výpočtu Jukes-Cantor, zobrazení pak metodou average. Tento dendrogram je až na vzdálenost uzlů identický s dendrogramem sestaveným z denzit s délkou okna 5 a metodou sestavení complete. Proto ho nebudeme znovu podrobněji rozebírat.

Na závěr jsou na obrázku 26 a 27 zobrazeny fylogenetické stromy vytvořené přímo na webu BOLDsystems. Zde se opět vyskytují podobné odchylky v rozdělení jednotlivých řádů jako v předchozích dendrogramech.



Obrázek 26.: Fylogenetický strom vytvořený na webu Boldsystems metodou Jukes-Cantor



Obrázek 27.: Fylogenetický strom vytvořený na webu Boldsystems metodou Kimura 2 Parameter

8.4. Zhodnocení výsledků

Z analýzy dendrogramů je zřejmé, že námi navržená metoda jejich sestavení s využitím denzit sekvencí má pro různá nastavení různé výsledky. Ve většině případů však přiřazuje druhy ze stejných řádů ke stejné větvi dendrogramu. Ke stejným větvím, bezprostředně k sobě, byly také přiřazovány nejpodobnější druhy – ty které mají stejné rodové jméno a liší se jen druhovým. Nejlepších výsledků bylo dosaženo pro délku okna 15, což je v našem případě, v povolených mezích délky okna 5 – 1/20 délky sekvence, přibližně průměr. V dendrogramech se bohužel vyskytuje i neprávné přiřazení do řádů. Tyto chyby mohou být způsobeny například špatným, či rozdílným sekvenováním DNA, kdy mají sekvence různou délku. V našem souboru se vyskytují sekvence o délce 605 – 672 bází. Tyto rozdíly se pak projeví při zarovnání přidáním mezer a ty pak mají vliv na výslednou denzitu. Dalším zdrojem chybných přiřazení mohou být mutace v mitochondriální DNA, kterých se zde oproti jaderné vyskytuje více.

9. Závěr

Úkolem této bakalářské práce na téma „Komparační analýza genomických dat pomocí grafické reprezentace“ bylo sestavení rešerše o taxonomické metodě DNA barcodingu. V popisu této metody se můžeme seznámit s její historií, se způsobem analýzy vzorků ale i se způsobem získáváni genomických dat. Z přehledu informací je zřejmé, že DNA barcoding je metoda poměrně mladá, avšak do budoucna by mohla mít široké uplatnění. V práci je také pojednáno o buněčné organele mitochondrii, která s touto metodou úzce souvisí. Po té jsou shrnuty informace o databázích barcode a práci s nimi. Jsou zde také informace o formátu FASTA, který je využíván ve většině databází pro práci s biologickými daty.

Dalším úkolem, teď už praktickým, bylo porovnání denzit nukleotidů. K tomuto účelu byl vytvořen program v prostředí MATLAB, mezi jehož hlavní funkce pro práci s genomickými daty patří výpočet denzity, euklidovských distancí a zobrazení dendrogramů.

Tento program pro svou činnost využívá 2 funkce. První funkce ze zadané sekvence mitochondriálního genu CO1 vypočítá denzitu DNA. Tato funkce funguje správně a vypočítá požadovaný výstup.

Druhá funkce slouží k porovnání podobnosti/rozdílnosti dvou sekvencí, a to výpočtem distancí mezi sekvencemi z denzit v načteném souboru dat. Tato funkce také funguje podle očekávání. Program počítal nejprve denzity a distance z nezarovnaných sekvencí. Později byl, z důvodu lepších výsledků, aktualizován pro počítání s globálně zarovnaným souborem dat. Program umožňuje pro vytvoření dendrogramů volbu různých metod sestavení.

Pro analýzu byl zvolen soubor 30 druhů ryb (Amazon Fishes) a jejich mitochondriální DNA, konkrétně gen CO1. Ve vytvořeném programu byla nejprve vypočítána denzita sekvence jednotlivých druhů. Z těchto denzit pak byla mezi druhy spočítána jejich euklidovská vzdálenost a tím vytvořena distanční matice jednotlivých druhů. Z této distanční matice pak byly vytvořeny dendrogramy s využitím různých metod sestavení a vstupních dat.

Tyto dendrogramy až na výjimky přiřazovaly druhy ze stejných řádů k sobě. To ovšem neznamená, že můžeme vytvořené dendrogramy srovnávat například s fylogenetickými stromy a určovat tím evoluční vývoj jednotlivých druhů. K přiblížení k fylogenetickým stromům by bylo pro tuto analýzu vhodnější využít například místo mitochondriální DNA jadernou a dále srovnávat sekvence o stejné délce.

10. Seznam použité literatury

[1] MCBRIDE HM, NEUSPIEL M, WASIAK S. Mitochondria: more than just a powerhouse. Curr Biol., červenec 2006, roč. 16, čís. 14, s. R551–60. Dostupné online:10.1016/j.cub.2006.06.054. PMID 16860735

[2] KUBIŠTA, Václav. Buněčné základy životních dějů. Praha : Scientia, 1998. ISBN 80-7183-109-3. S. 210.

[3]VŠCHT PRAHA, Ústav organické technologie. Mitochondriální dědičnost. [online]. [cit. 12. prosince 2010]. Dostupné na WWW: <http://www.vscht.cz/kot/resources/studijni-materialy/bc-skripta/kapitola04.pdf>

[4]ŠTEFÁNEK, Jiří. Medicína, nemoci, studium na 1. LF UK *[online]. [s.n.]. [cit. 11.02 2010]. <http://www.stefajir.cz>.*

[5]Genetika - Váš zdroj informací o genetice : Mitochondrie a jejich genetická informace [online]. 26. 3. 2005. 26. 3. 2005 [cit. 2011-03-18]. http://genetika.wz.cz. Dostupné z WWW: <http://genetika.wz.cz/clanky/clanek2.php>.

[6] EFENBERK, Aleš. MIMOJADERNÁ DĚDIČNOST U ČLOVĚKA. [s.l.], 2008. 32 s. MASARYKOVA UNIVERZITA; Přírodovědecká fakulta; Ústav experimentální biologie; Oddělení genetiky a molekulární biologie. Vedoucí bakalářské práce prof. RNDr. Jiřina Relichová, CSc.

[7] KHALIMONCHUK, O.; RÖDEL, G. (2005). "Biogenesis of Cytochrome c Oxidase". Mitochondrion 5 (6): 363–383.

[8] SILVEIRA PC, STRECK EL, PINHO RA. (2005). "Cellular effects of low power laser therapy can be mediated by nitric oxide.". Lasers Surg Med. 36 (4): 307–14.

[9] Barcode of Life [online]. 2010 [cit. 2010-12-27]. www.barcodeoflife.com. Dostupné z WWW: <http://www.barcodeoflife.org/content/about/what-dna-barcoding>.

[10] HEBERT PDN, STOECKLE MY, ZEMLAK TS, FRANCIS CM (2004) Identification of Birds through DNA Barcodes. PLoS Biol 2(10): e312. doi:10.1371/journal.pbio.0020312 Dostupné online: <http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0020312 >

[11] MORITZ, C., CICERO, C. DNA Barcoding: Promise and Pitfalls. PLoS Biology. 2004, vol. 2, no. 10, p. 1529-1531.

[12] DAVID P. CLARK, Molecular biology: Understanding the genetic revolution. Elsevier Academic Press (2005), ISBN 0-12-175551-7.

[13]RACLAVSKÝ, Vladislav. Metody molekulární genetiky [online]. Ústav biologie Lékařské fakulty Univerzity Palackého v Olomouci, 2003, [cit. 2011-03-20]. Kapitola 8. Sekvenování DNA

[14] Barcode of Life [online]. 2010 [cit. 2010-12-27]. www.barcodeoflife.com. Dostupné z WWW: <http://www.barcodeoflife.org/content/about/what-cbol>.

[15] Barcode of Life [online]. 2010 [cit. 2010-12-27]. www.barcodeoflife.com. Dostupné z WWW: <http://www.barcodeoflife.org/content/about/what-ibol>.

[16]ECBOL - European Consortium for the Barcode of Life : DNA Barcoding Databases [online]. 2010 [cit. 2011-03-18]. http://www.ecbol.org. Dostupné z WWW: <http://www.ecbol.org/Resources/dna-barcoding-databases.html>.

[17]FATIMA CVRČKOVÁ, Úvod do praktické bioinformatiky, Academia, ISBN 80-200-1360-1, 2006

[18]LABOUNEK, R. Porovnávání mitochondriální DNA pro identifikaci druhů. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2010. 54 s. Vedoucí bakalářské práce Ing. Denisa Maděránková.

11. Seznam zkratek

- ATP - adenosintrifosfát
- BOLD - Barcode of Life Data Systems, University of Guelph's Barcode of Life Database
- CBOL -The Consortium for the Barcode of Life
- CO1 - cytochrom oxydáza
- CO2 - oxid uhličitý
- DNA - deoxyribonukleová kyselina
- ECBOL - European The Consortium for the Barcode of Life
- EMBL - Evropská molekulárně biologická laboratoř
- FADH2 - Flavinadenindinukleotid
- FISH-BOL – Fish Barcode of Life
- H2O - voda
- iBOL - The International Barcode of Life project
- mtDNA - mitochondriální deoxyribonukleová kyselina
- NADH - Nikotinamid adenin dinukleotid
- PCR - Polymerase Chain Reaction
- rRNA - ribozomální ribonukleová kyselina
- tRNA - transferová ribonukleová kyselina

12. Seznam příloh

- DVD s elektronickou verzí bakalářské práce a kompletním vytvořeným programem,
- Papírová verze zdrojových kódů jednotlivých funkcí a programu

Funkce denzita_dna.m

```
function [denzita]=denzita_dna(sekvence,delkaokna)
sekvence=upper(sekvence);
y=abs(sekvence);
delkasekvence=length(sekvence);
z = zeros(1,delkasekvence+2*delkaokna);
for a=1:delkasekvence

    switch y(a)
        case 65
            z(a+delkaokna)=1;
        case 67
            z(a+delkaokna)=2;
        case 71
            z(a+delkaokna)=3;
        case 84
            z(a+delkaokna)=4;
        otherwise
            z(a+delkaokna)=0;
    end
end
denzita = zeros([4 delkasekvence+2*delkaokna]);
for radek=1:4
    for sloupec=1:delkasekvence+delkaokna
        c(radek,sloupec) = length(find
(z((sloupec+1):(sloupec+delkaokna))==radek));
      denzita(radek,sloupec)=c(radek,sloupec)/delkaokna;
    end
end
end
```

Funkce euklid.m

```
function [n]=euklid(b,d)
 if length(b)<length(d)
    n = zeros(1,length(b));
    delka=length(b);
 else
     n = zeros(1,length(d));
    delka=length(d) ;
 end
    for sloupec=1:delka
        n(sloupec) = sqrt(((b(1, sloupec)-d(1,sloupec))^2) + ((b(2, sloupec)-
d(2,sloupec))^2) + ((b(3, sloupec)-d(3,sloupec))^2) + ((b(4, sloupec)-
d(4,sloupec))^2));
    end
end
```

Zdrojový kód programu denzity_calculator

```
function varargout = denzity_calculator(varargin)
gui_Singleton = 1;
gui_State = struct('gui_Name',       mfilename, ...
                   'gui_Singleton',  gui_Singleton, ...
                   'gui_OpeningFcn', @denzity_calculator_OpeningFcn, ...
```

```
                   'gui_OutputFcn',  @denzity_calculator_OutputFcn, ...
                   'gui_LayoutFcn',  [] , ...
                   'gui_Callback',   []);
if nargin && ischar(varargin{1})
    gui_State.gui_Callback = str2func(varargin{1});
end
if nargout
    [varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
else
    gui_mainfcn(gui_State, varargin{:});
end
function denzity_calculator_OpeningFcn(hObject, ~, handles, varargin)
handles.output = hObject;
guidata(hObject, handles);
clear all
clc
function varargout = denzity_calculator_OutputFcn(~, ~, handles)
varargout{1} = handles.output;
function pushbutton1_Callback(~, ~, handles)
global Names Sequences value  pocetsekvenci overeni
[FileName,PathName]= uigetfile({'*.mat;*.m','Soubory dat'},'Vyber soubor
DAT');
set(handles.text12,'string','Načítání sekvencí' )
cela_cesta=[PathName FileName];
if ischar (cela_cesta) && exist(cela_cesta) ==2
    load ( cela_cesta , 'Names', 'Sequences');
set(handles.text2,'string',FileName )
 set(handles.listbox1,'string',Names )
 value=1;
   extsekvence=cell(1,value);
   pocetsekvenci=length(Sequences);
   overeni=0;
   set(handles.text12,'string','Sekvence načteny' )
end
function listbox1_Callback(~, ~, ~)
function listbox1_CreateFcn(hObject, ~, ~)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
function listbox2_Callback(hObject, ~, ~)
function listbox2_CreateFcn(hObject, ~, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
function pushbutton2_Callback(~, ~, handles)
global  Sequences value extsekvence
p=get(handles.listbox1,'Value');
if p>0
     extsekvence{1,value}=Sequences{1,p};
     value=value+1;
end
set(handles.text12,'string','Zobrazena vybraná sekvence' )
set(handles.listbox2,'string',extsekvence )
function pushbutton3_Callback(~, ~, handles)
```

```
global  Sequences
set(handles.text12,'string','Výpočet denzity vybrané sekvence' )
delkaokna = str2double(get(handles.edit3,'String'));
p=get(handles.listbox1,'Value');
x=Sequences{1,p};
denzita=denzita_dna(x,delkaokna);
soucet_at=denzita(1,:)+denzita(4,:);
soucet_cg=denzita(2,:)+denzita(3,:);
figure
subplot(4,1,1)
plot (denzita(1,:))
axis([0,length(denzita),-0.1,1.1])
ylabel('Denzita')
title('Adenin')
subplot(4,1,2)
plot (denzita(2,:))
axis([0,length(denzita),-0.1,1.1])
ylabel('Denzita')
title('Cytosin')
subplot(4,1,3)
plot (denzita(3,:))
axis([0,length(denzita),-0.1,1.1])
ylabel('Denzita')
title('Guanin')
subplot(4,1,4)
plot (denzita(4,:))
axis([0,length(denzita),-0.1,1.1])
title('Thymin')
ylabel('Denzita')
xlabel('Délka sekvence')
figure
subplot(3,1,1)
plot (soucet_at)
axis([0,length(denzita),-0.1,1.1])
title('Součet denzit A-T')
ylabel('Denzita')
subplot(3,1,2)
plot (soucet_cg)
axis([0,length(denzita),-0.1,1.1])
title('Součet denzit C-G')
ylabel('Denzita')
subplot(3,1,3)
plot (soucet_at)
hold on
plot (soucet_cg,'r')
axis([0,length(denzita),-0.1,1.1])
title('AT vs. CG')
ylabel('Denzita')
xlabel('Délka sekvence')
legend('A-T','C-G')
set(handles.text12,'string','Zobrazena denzita vybrané sekvence' )
function pushbutton4_Callback(hObject, eventdata, handles)
global Sequences pocetsekvenci vzdalenost overeni
set(handles.figure1,'Pointer','watch');
set(handles.text12,'string','Probíhá výpočet denzity a distancí...' );
delkaokna = str2double(get(handles.edit3,'String'));
```

```
denzity=cell(1,pocetsekvenci);
vzdalenost=cell(pocetsekvenci,pocetsekvenci);
for i=1:pocetsekvenci
 d=Sequences{1,i};
denzity{1,i}=denzita_dna(d,delkaokna);
end
for j=1:pocetsekvenci
    for k=1:pocetsekvenci
        ee=euklid(denzity{j},denzity{k});
      vzdalenost{j,k}=sum(ee);
    end
end
overeni=1;
xlswrite('testdata.xls', vzdalenost)
winopen('testdata.xls')
set(handles.figure1,'Pointer','arrow');
set(handles.text12,'string','Vypočtena denzita a distance nezarovnaných
sekvencí' )
function text2_CreateFcn(hObject, eventdata, handles)
function text4_CreateFcn(hObject, eventdata, handles)
function edit3_Callback(hObject, eventdata, handles)
global Sequences
p=get(handles.listbox1,'Value');
delkaokna = str2double(get(hObject,'String'));
dvacetinasekvence=round(length(Sequences{p})/20);
if delkaokna<5
    delkaokna=5;
    warndlg('Nepovolená délka okna - Upraveno');
elseif delkaokna>dvacetinasekvence
      delkaokna = dvacetinasekvence  ;
      warndlg('Nepovolená délka okna - Upraveno');
else
    delkaokna=delkaokna;
end
if mod(delkaokna,2)==1;
   delkaokna=delkaokna;
else
   delkaokna=delkaokna-1;
   warndlg('Nepovolená délka okna - Upraveno');
end
set(handles.edit3,'string',delkaokna);
function edit3_CreateFcn(hObject, ~, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
function text6_CreateFcn(hObject, eventdata, handles)
function pushbutton5_Callback(hObject, eventdata, handles)
global vzdalenost overeni
if overeni == 1
   [FileName, PathName] = uiputfile('*.xls','Ulož jako:');
    SaveFile = [PathName FileName];
    xlswrite(SaveFile, vzdalenost)
set(handles.text12,'string','Uložen soubor distancí:', SaveFile )
else
        warndlg('Musíte nejdříve vypočítat distanci sekvencí');
```

```
      set(handles.text12,'string','Chyba - nevypočítána distance' )
end
function pushbutton6_Callback(hObject, eventdata, handles)
global vzdalenost Names overeni
if overeni == 1
set(handles.text12,'string','Probíhá sestrojení fylogenetického stromu z
distancí...' )
val=get(handles.popupmenu1,'Value');
M = cell2mat(vzdalenost);
switch val
    case 1
     tree = seqlinkage(M,'average', Names) ;
    case 2
     tree = seqlinkage(M,'single', Names) ;
    case 3
   tree = seqlinkage(M,'complete', Names);
    case 4
       tree = seqlinkage(M,'weighted', Names) ;
    case 5
        tree = seqlinkage(M,'centroid', Names);
    case 6
        tree = seqlinkage(M,'median', Names);
end
view(tree);
set(handles.text12,'string','Sestrojen fylogenetický strom z distancí
sekvencí' )
else
      warndlg('Musíte nejdříve vypočítat distanci sekvencí');
      set(handles.text12,'string','Chyba - nevypočítána distance' )
end
function pushbutton7_Callback(hObject, eventdata, handles)
global Sequences Names
set(handles.text12,'string','Probíhá sestrojení fylogenetického stromu...' )
val=get(handles.popupmenu2,'Value');
switch val
    case 1
        D = seqpdist(Sequences, 'Method', 'jukes-cantor');
    case 2
        D = seqpdist(Sequences,'Method',  'p-distance');
    case 3
        D = seqpdist(Sequences,'Method',  'alignment-score');
end
val=get(handles.popupmenu3,'Value');
switch val
    case 1
     tree = seqlinkage(D,'average', Names) ;
    case 2
     tree = seqlinkage(D,'single', Names) ;
    case 3
   tree = seqlinkage(D,'complete', Names);
    case 4
       tree = seqlinkage(D,'weighted', Names) ;
    case 5
        tree = seqlinkage(D,'centroid', Names);
    case 6
        tree = seqlinkage(D,'median', Names);
```

```
end
view(tree);
  set(handles.text12,'string','Sestrojen fylogenetický strom' )
function popupmenu1_Callback(~, ~, ~)
function popupmenu1_CreateFcn(hObject, ~, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
function popupmenu2_Callback(~, ~, ~)
function popupmenu2_CreateFcn(hObject, ~, ~)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
function popupmenu3_Callback(~, ~, ~)
function popupmenu3_CreateFcn(hObject, ~, ~)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
function pushbutton8_Callback(~, ~, handles)
global Sequences pocetsekvenci vzdalenost overeni
set(handles.text12,'string','Probíhá výpočet denzity a distance ze zarovnaných
sekvencí...' )
set(handles.figure1,'Pointer','watch');
delkaokna = str2double(get(handles.edit3,'String'));
denzity=cell(1,pocetsekvenci);
vzdalenost=cell(pocetsekvenci,pocetsekvenci);
Sequences2=multialign(Sequences);
Sequences2=cellstr(Sequences2);
Sequences2=Sequences2';
for i=1:pocetsekvenci
 d=Sequences2{1,i};
denzity{1,i}=denzita_dna(d,delkaokna);
end
for j=1:pocetsekvenci
    for k=1:pocetsekvenci
        ee=euklid(denzity{j},denzity{k});
      vzdalenost{j,k}=sum(ee);
    end
end
xlswrite('testdata.xls', vzdalenost);
winopen('testdata.xls');
overeni=1;
set(handles.figure1,'Pointer','arrow');
set(handles.text12,'string','Vypočítána denzita a distance zarovnaných
sekvencí' );
```