



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

VLIV NUMERICKÝCH REPREZENTACÍ DNA NA ÚSPĚŠNOST MOLEKULÁRNÍ TAXONOMIE

EFFECT OF DIFFERENT NUMERICAL REPRESENTATIONS OF DNA ON MOLECULAR
TAXONOMY

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE

AUTHOR

ELIŠKA BLASCHOVÁ

VEDOUCÍ PRÁCE

SUPERVISOR

Ing. DENISA MADĚRÁNKOVÁ

BRNO 2014



VYSOKÉ UČENÍ
TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor

Biomedicínská technika a bioinformatika

Studentka: Eliška Blaschová

ID: 147461

Ročník: 3

Akademický rok: 2013/2014

NÁZEV TÉMATU:

Vliv numerických reprezentací DNA na úspěšnost molekulární taxonomie

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte literární rešerši na téma molekulární taxonomie a numerických reprezentací DNA. 2) V libovolném programovém prostředí naprogramujte alespoň 3 metody numerických reprezentací, které se jeví jako použitelné pro porovnávání sekvencí. 3) Z dostupných databází vyberte sekvence jednoho genu pro skupinu organismů s dobře známou taxonomií. Pro tato data vytvořte numerické reprezentace všemi vybranými metodami. 4) Pro jednotlivé čeledě organismů vytvořte zástupné referenční signály vhodným zpracováním numerických signálů. 5) Proveďte porovnávací klasifikaci nezávislého souboru dat na základě podobností s referenčními signály a vyhodnoťte úspěšnost pro jednotlivé metody numerické reprezentace.

DOPORUČENÁ LITERATURA:

[1] SCHWARTZ, Jeffrey. Molecular Systematics and Evolution. Encyclopedia of Molecular Cell Biology and Molecular Medicine, 2nd ed. Wiley: Weinheim, 2005, 696 s. ISBN: 3-527-30550-2.

[2] BLAXTER, Mark. The promise of a DNA taxonomy. Phil. Trans. R. Soc. Lond. B., 2004, vol. 359, pp. 669-679.

Termín zadání: 10.2.2014

Termín odevzdání: 30.5.2014

Vedoucí práce: Ing. Denisa Maděránková

Konzultanti bakalářské práce:

prof. Ing. Ivo Provazník, Ph.D.

Předseda oborové rady

UPOZORNĚNÍ:

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

ABSTRAKT

Práce seznamuje s klasickou a molekulární taxonomií, které se používají ke klasifikaci organismů. Představuje směr DNA barcodingu jako možnost jak přiřadit neznámý vzorek organismu k známému druhu. Získané sekvence mitochondriální DNA při DNA barcodingu je třeba zpracovat vhodnou metodou numerické reprezentace, která bude správně informovat o příslušnosti neznámého organismu k druhům a dokáže jej tedy správně zařadit. V této práci jsou použity k porovnání sekvencí organismů celkem tři metody numerické reprezentace, a to 1. a 4. kvadrant, EIIP hodnoty a třívektorová reprezentace. V praktické části je naprogramována identifikační analýza, která přiřazuje testovaným organismům referenční organismus. Testování je provedeno na 4 referenčních souborech a 5 analyzovaných souborech. Práce shrnuje úspěšnost přiřazení správné reference pro vybrané metody numerické reprezentace.

KLÍČOVÁ SLOVA

Molekulární taxonomie, mitochondriální DNA, DNA barcoding, numerické reprezentace

ABSTRACT

This paper introduces the classical and molecular taxonomy used for classification of organisms. It represents the direction of DNA barcoding as a possibility to identify unknown organism. Acquired mitochondrial DNA sequences in DNA barcoding can be transformed by suitable method to numerical representation that will properly inform about the organism's relationship to other taxa. In this work, we consider three methods of numerical representation: the 1st and 4th quadrant EIIP values and 3D numerical representation reduction. The practical part is programmed identification analysis, which assigns the test organisms reference organism. Testing is performed at 4 reference files and analyzed 5 files. The work summarized the success assigning the correct reference for the selected method of numerical representation.

KEYWORDS

Molecular taxonomy, mitochondrial DNA, DNA barcoding, numerical representation

BLASCHOVÁ, E. *Vliv numerických reprezentací DNA na úspěšnost molekulární taxonomie*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2014. 55 s. Vedoucí bakalářské práce Ing. Denisa Maděránková.

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma Vliv numerických reprezentací DNA na úspěšnost molekulární taxonomie jsem vypracovala 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 bakalářské práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědoma následků porušení ustanovení § 11 a následujících zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne

.....

(podpis autora)

PODĚKOVÁNÍ

Děkuji Ing. Denise Maděránkové, vedoucí mé bakalářské práce za precizní výklad metodiky, neochvějnou pedagogickou i odbornou pomoc a za účelné rady při zpracování mé bakalářské práce.

V Brně dne

.....

(podpis autora)

OBSAH

Seznam tabulek	vi
Seznam obrázků	vii
Úvod	1
1 Taxonomie	2
1.1 Taxonomie i systematika	2
1.2 Klasická taxonomie.....	3
1.2.1 Taxonomický kruh dedukce.....	4
1.3 Molekulární taxonomie.....	6
1.3.1 Výhody a nevýhody molekulárních znaků	7
1.4 Druh	7
1.5 Variabilita znaků.....	8
2 Mitochondrie	9
2.1 Mitochondriální DNA.....	9
2.1.1 Mitochondriální genom u metazoa	11
2.1.2 Mitochondriální genom u rostlin	11
2.1.3 Mitochondriální genom v houbách	12
2.1.4 Mutace	12
3 Biologická identifikace pomocí DNA barcodingu	15
3.1 Přístup DNA barcodingu	15
3.2 Standardní postup analýzy	16
3.3 Organizace pro DNA barcoding	18
3.4 DNA barcoding založený na znacích DNA.....	19
4 Fylogenetické stromy	21
5 Numerické reprezentace	25

5.1	2D reprezentace v 1. a 4. kvadrantu.....	25
5.2	EIIP hodnoty	26
5.3	Třívektorová reprezentace	27
6	Praktická část	29
6.1	Programové řešení	29
6.2	Identifikační analýza	32
6.3	Fylogenetické stromy.....	34
6.4	Analýza a vyhodnocení výsledků	37
7	Závěr	42
	Literatura	44
	Seznam symbolů, veličin a zkratk	48

SEZNAM TABULEK

Tabulka 1: Hodnoty EIIP pro jednotlivé nukleotidy (převzato z [1]).....	27
Tabulka 2: Princip zarovnání sekvencí	31
Tabulka 3: Souhrn počtu druhů, rodů a vyššího taxonu pro jednotlivé referenční soubory	33
Tabulka 4: Souhrn počtu druhů, rodů a vyššího taxonu pro jednotlivé testované soubory	33
Tabulka 5: Početní zastoupení testovaných sekvencí, ke kterým existuje reference.....	33
Tabulka 6: Procentuální úspěšnost taxonomického určení pro druhy	34
Tabulka 7: Úspěšnost taxonomického zařazení pro rody	34
Tabulka 8: Úspěšnost kondenzace fylogenetických stromů pro taxonomické jednotky	37
Tabulka 9: Úspěšnost kondenzace fylogenetických stromů pro taxonomické jednotky	37
Tabulka 10: Průměrná úspěšnost správného přiřazení testovaného souboru pro druhy a rody	37
Tabulka 11: Úspěšnost taxonomického přiřazení pro kombinaci srovnání pomocí minima a euklidovské vzdálenosti	40
Tabulka 12: Průměrné hodnoty testovaných souborů pro fylogenetické stromy	41

SEZNAM OBRÁZKŮ

Obrázek 1: Odhad druhů v taxonomických úrovních (převzato z [19])	3
Obrázek 2: Taxonomický kruh dedukce (převzato z [8])	5
Obrázek 3: Stavba mitochondrie (převzato z [2]).....	9
Obrázek 4: Mitochondriální genom (převzato z [24])	11
Obrázek 5: Ukázka bodové mutace (převzato z [20])	13
Obrázek 6: Zásah HSP do genové exprese (převzato [12])	14
Obrázek 7: Postup jakým by měly být získávány sekvence z jednoho druhu (převzato z [4]).....	18
Obrázek 8: Fylogenetický strom života (převzato z [11])	22
Obrázek 9: Ukázka metody outgroup (převzato z [11])	23
Obrázek 10: Příklad kladogramu vidíme vlevo (první dva) a fylogramu vpravo	24
Obrázek 11: Reprezentace nukleotidů v 1. a 4. kvadrantu kartézského souřadného systému (převzato z [16]).....	26
Obrázek 12: Nukleotidový čtyřstěn (vlevo) a umístění nukleotidového čtyřstěnu v RGB prostoru (vpravo) převzato z [16]	27
Obrázek 13: Zjednodušené blokové schéma programu	32
Obrázek 14: Zjednodušené blokové schéma pro tvorbu fylogenetického stromu.....	35
Obrázek 15: Princip kondenzace fylogenetického stromu pro druh.....	36
Obrázek 16: Kondenzovaný fylogenetický strom souboru subtropických ryb (FCFP) metodou K2P - NJ	36
Obrázek 17: Podobnost testovaného druhu <i>Phyllostomus hastus</i> k referenci <i>Anoura latidens</i> namísto k referenci <i>Phyllostomus hastus</i>	38
Obrázek 18: Podobnost testovaného druhu <i>Xenops minutus</i> k referenci <i>Hypocnemis cantor</i> namísto k referenci <i>Xenops minutus</i>	39

ÚVOD

V poslední době vzrostl zájem o systematiku a taxonomii. Jedním z hlavních důvodů bylo zjištění, že systematika je limitována vymíráním organismů. Je zcela pravděpodobné, že některé organismy mohou vymřít dříve, než se je podaří správně katalogizovat. Tenhle případ se málem stal u druhu kukuřice *Zea diploperennis*. V 70. letech byl objeven tento druh kukuřice v Mexiku. Tato plodina je odolná vůči nemocem běžně pěstované kukuřice *Zea mays* a navíc se jedná o trvalku. Bohužel, než byly tyto informace zjištěny, nový druh kukuřice rostl vlivem devastace krajiny pouze na ploše o rozloze 10 ha [9]. Tato příhoda dokazuje, že objevení nových druhů může mít spoustu kladných využití, stejně jak tomu bylo u druhu kukuřice *Zea mays*.

V roce 2013 byl v České republice zveřejněn článek s předpokládaným odhadem druhů a jejich doposud popsaným počtem. Odhad byl stanoven na 8,7 mil druhů (připustili chybu $\pm 1,3$ mil). Biologové dosud popsali 1,25 mil druhů a každý rok k tomuto číslu přidají 15 tisíc nových druhů. [19]

Je tedy zcela pravděpodobné, že nestihneme popsat všechny druhy za jejich života. Přičemž správné určení u vymřelých organismů je založeno na klasické taxonomii, tedy znakové metodě. Na kolik bude správný výsledek za použití znakové metody u vymřelých dosud neznámých organismů? Mohli bychom jako lidstvo vynalézt nebo dát základ rychlejší a přesnější taxonomické metodě? Tato práce se pokusí nastínit současné bioinformatické možnosti taxonomického určení organismů.

Mezi současné nadějné bioinformatické nástroje pro klasifikaci organismů, v některých případech dokonce přesnější a rychlejší, bude představen směr DNA barcodingu. Nyní se tato metoda nachází na pomezí oblíbenosti a kritiky. Práce představí základní princip metody, jeho výhody, nevýhody a v praktické části se bude věnovat úspěšné klasifikaci organismů za pomoci DNA barcodingu.

TEORETICKÁ ČÁST

1 TAXONOMIE

Výraz taxonomie pochází ze dvou řeckých slov *tax* (srovnat, uspořádat) a *nomia* (zvyk, zákon). Jedná se tedy o soubor principů a metod k uspořádání či zařazování organismů. Taxonomie bývá často zaměňována se systematikou, a to i ve vyšších kruzích zkušených biologů. Systematika je brána jako vědní obor, zabývající se studiem původu a organizací biologické diverzity. Teprve ta se dělí na taxonomii, studium fylogeneze a biosystematiku. Může se zdát, že se významově jedná o totéž, avšak podle některých a mezi ně patří i Pavel Drozd, se pojem systematika a taxonomie nekryje. Protože se ale tato práce nezabývá rozdíly mezi systematikou a taxonomií, dále jsou oba dva pojmy na stejné úrovni. [9]

1.1 Taxonomie i systematika

Taxonomie nám slouží k hierarchickému zařazení organismů. Každá skupina organismů, která má společné znaky nebo jejich kombinaci, utváří taxonomickou skupinu. Podle počtu společných znaků se organismy v jedné taxonomické skupině odlišují od ostatních jedinců či taxonomických skupin a tím vytváří taxonomickou jednotku. Například se jedná o rod. Avšak menší rozdílné odlišnosti znaků či jejich kombinace lze najít i mezi stejnou taxonomickou skupinou. Tyto odlišnosti nám umožňují další rozřazení např. do druhů. Současná taxonomie má několik hierarchických úrovní, jak napovídá obrázek 1 pro taxonomii živočichů. Taxonomii dále dělíme na klasickou a molekulární. [9]



Obrázek 1: Odhad druhů v taxonomických úrovních (převzato z [19])

1.2 Klasická taxonomie

Klasické taxonomické určení se zakládá na znakových informacích. Jedná se o informace morfologické, behaviorální, reprodukční, ekologické a geografické. Na základě těchto informací jsou systematici schopni zařadit daný organismus do jednotlivých taxonů.

Znaky jsou odlišnosti, kterými se organismy či jednotlivé taxony navzájem liší. Znaky mohou nabývat různých hodnot a dokonce i více stavů. Znakem je pak určena jakákoli vlastnost mající dva stavy. Rozeznáváme znaky kvantitativní. Ty jsou buď spojité, nabývají reálné číselné hodnoty, nebo diskrétní, jejich hodnotou jsou přirozená čísla. Dále pak rozeznáváme znaky kvalitativní, jejich stav nelze vyjádřit číslem (např. barva očí – hnědá). [11]

Znaky mohou pocházet od společného předka, pak se jedná o tzv. homologické znaky. Tyto znaky dále dělíme na znaky synplesiomorfni, synapomorfni a autapomorfni. Synplesiomorfni znaky jsou znaky, kde se jedná o původní formu znaku původního předka, která se později při větvení taxonů může či nemusí změnit. Někteří potomci budou mít znaky v nezměněné podobě a někteří naopak v podobě pozměněné. Navzájem si potomci s nezměněnými znaky nejsou příbuzensky bližší než ti, co své znaky mají pozměněné. Všichni potomci jsou mezi sebou na stejném stupni příbuznosti k předkovi, a tudíž tyto znaky nemají význam při rekonstrukci fylogeneze neboli vývojové příbuznosti mezi organismy. Naopak synapomorfni znak je významnou informací pro fylogenetickou rekonstrukci, protože ukazuje na existenci společného předka, u kterého tento znak vznikl. Znak autapomorfni vzniká pouze u jednoho taxonu a nemá význam pro fylogenezi, ale pro anagenezi, tedy vývoj vlastností. Pokud znaky

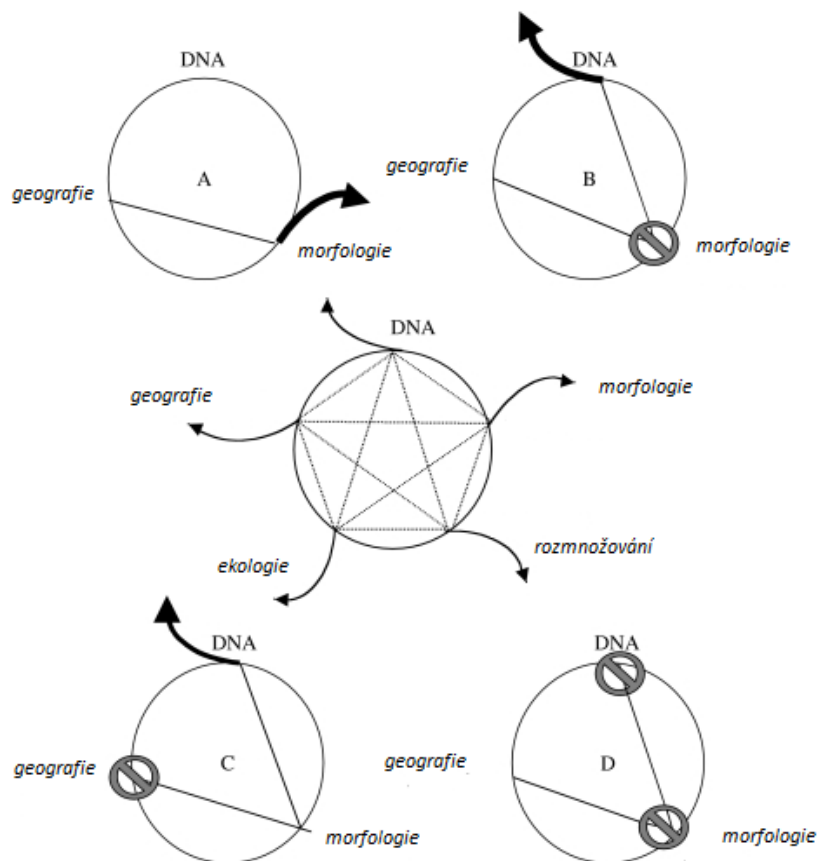
nemají společný původ, jedná se o homoplazické znaky. Mezi ně patří paralelismus (několikrát projevená skrytá vloha), konvergence (podobnost, vzniklá různým vývojem) a analogie (povrchní podobnost, ovlivněná funkcí). [11]

Taxon je formální jednotka představující jistou úroveň v zařazení organismů. Do taxonu řadíme organismy, které mají společné znaky. Tyto znaky přiřazujeme živočichům subjektivně. Tudíž není stanovena ani hranice, která ohraničuje námi vybraný taxon. Subjektivnost se zvětšuje s přesnějším zařazením. Podle vazby předek – potomek se rozlišují tři druhy taxonů. Monofyletický taxon neboli klad (ang. clade, také překládán jako vývojová větev) obsahuje společného předka a všechny jeho potomky. Jeho členové jsou si vzájemně více příbuzní, než je kterýkoliv z nich příbuznější k jinému druhu. Tento taxon je rozšířenější. Dále “nepřípustný” polyfyletický taxon, který neobsahuje svého předka (mnohdy i většinu potomků). Neboť právě tento společný předek neměl většinu znaků, aby byl zařazen ke svým potomkům. Právě počet těchto polyfyletických taxonů se v rámci taxonomie snažíme snížit. Čímž se snažíme o zjištění příbuznosti jednotlivých organismů. A poslední parafyletický taxon, často nazýván vývojovým stupněm, obsahuje všechny předky bez všech potomků. Právě tento poslední druh taxonu je nejčastějším. Tento předpoklad nejčastějšího taxonu (parafyletického) vychází z dosaženého počtu popsanych druhů 1,25 z 8,7 mil (k roku 2013). [9]

1.2.1 Taxonomický kruh dedukce

Mezi práci taxonoma patří kruh dedukce. Analyzovaný organismus je určen na základě taxonomického kruhu dedukce a to jeho prolomením (obrázek 2 pro klasickou taxonomii bez položky DNA). Na začátku se stanoví hypotéza, která se prolomením kruhu potvrdí. Taxonomická hypotéza je stanovena na základě podobnosti v určité geografické lokalitě. Překřížením kruhu se dostáváme k morfologickým znakům, které potvrzují hypotézu podobnosti na daném geografickém území, a tím se stávají diagnostickými charakterystiky, a umožňují prolomení kruhu dedukce (obrázek 2 – panel A). Pokud nedokážeme potvrdit morfologickou podobnost, tak překřížujeme kruh dedukce jinou cestou, dokud se nám jej nepodaří prolomit. Samozřejmě se může stát, že taxonom může uvíznout v kruhu dedukce. Proto bychom mohli zvážit začlenění DNA (tedy molekulární taxonomie) do tohoto kruhu pro přesnější analýzu. V jistých příkladech by se práce taxonomů mohla výrazně zjednodušit a to především u neznámých organismů, které čekají na objevení (obrázek 2 – panel B). V tomto případě morfologická podobnost nemůže být potvrzena, a tím pádem nemůže být potvrzena ani hypotéza, a taxonom nemůže opustit kruh. Proto se obrátí na zkoumání DNA sekvencí. V případě

shody se taxonom dostává z kruhu a tím dojde k potvrzení hypotézy. Pokud máme skupinu sympatrických druhů (příbuzných) vycházíme z morfologických znaků. Bohužel neexistuje žádný geografický vzor k potvrzení hypotézy, proto musíme použít DNA sekvence k prolomení kruhu (obrázek 2 – panel C). V případě (obrázek 2 – panel D), že DNA sekvence a ostatní překřížení kruhu selhaly, taxonom není schopen potvrdit hypotézu. V tomto případě můžeme říci, že se jedná o dosud neznámý druh organismu.[8]



Obrázek 2: Taxonomický kruh dedukce (převzato z [8])

Na základě tohoto začlenění DNA sekvence do kruhu dedukce bylo rozpoznáno několik nových druhů, které se do té doby považovaly jako součást jiného druhu či naopak. Například se jednalo o výzkum domnělého nového druhu muntjaka (drobného jelínka podobného srně v jihovýchodní Asii). Sekvence domnělého nového druhu v porovnání s ostatními druhy ve skupině byla unikátní. Avšak po začlenění muzejních vzorků se objevila shoda se vzorky nasbíraných před sedmdesáti lety v Laosu. Domnělý nový druh se stal druhem již existujícím *Muntiacus rooseveltorum*. [8]

Se zahrnutím DNA sekvence do kruhu dedukce musely být zváženy níže uvedené předpoklady, aby došlo ke správnému určení nového, případně již existujícího, druhu. Aby DNA data měla potenciální význam pro využití v objevování nového druhu, musí být ve skupině zahrnuty všechny druhy testovaného rodu k analýze. Navíc počet jedinců pro každý druh musí být dost velký, aby byl zástupcem variability (proměnlivosti) v daném genovém regionu pro druh jako celek. Navíc genový region musí být dostatečně variabilní, aby odhalil i drobné rozdíly mezi druhy. Avšak s velkým množstvím druhových vzorků se může stát, že variabilita genového regionu bude příliš velká. Toto povede k nesprávnému určení druhu organismu a jeho reprezentativnosti.[8]

1.3 Molekulární taxonomie

Oproti klasické taxonomii využívá molekulární taxonomie znaky, které mají stejnou důležitost. V klasické taxonomii se přiřazuje morfologickým znakům různě velká důležitost. Například se jedná o důležitou anatomii kostry a méně důležitý pigment. V molekulární taxonomii jsou molekulární znaky uloženy v sekvenci informačních molekul. Tyto informační molekuly nám mohou lépe ozřejmit evoluční novinky, které vznikají mutacemi na sekvenční úrovni DNA, RNA a proteinů. Tyto mutace, uloženy v jednotlivých lokusech DNA nebo proteinu, nás informují o evoluční historii, identitě jedince a příbuznosti druhové či populační.

Fylogenetika (nauka o vývoji organismů) si postupem času získala významné příznivce, kteří prosazovali názor, že organismy by se měly třídit pomocí příbuznosti. S myšlenkou příbuzenského vztahu jako obecného znaku se do popředí dostala molekulární fylogenetika, která rekonstruuje fylogenezi (vývojovou příbuznost organismů) s ohledem na jejich vývoj vlastností, tedy anagenezi, a také s ohledem na evoluční pořadí větvení taxonů, tedy kladogenezi. [9]

V molekulární taxonomii najdeme dva směry využívající molekulární znaky pro rekonstrukci fylogeneze. Kladistický a fenetický. Vývojově první byl přístup fenetický.

Fenetický přístup zkoumá podobnost mezi organismy tím, že do své analýzy zahrnuje všechny znaky ze všech kategorií znaků (popsaných výše u rozdělení znaků). Již v 60. letech minulého století byla na tomto přístupu založena numerická taxonomie. Za její zakladatele jsou považováni Sneath a Soukal, jejichž publikace Principy numerické taxonomie vyšla v roce 1963. Numerická taxonomie byl první pokus o objektivizaci taxonomie. Numeričtí taxonomové kladli důraz na použití velkého množství dat a vyvinuli matematické postupy, kterými vypočítali celkovou podobnost (nebo naopak odlišnost – distanci) mezi taxony. [11]

Kladistický přístup využívá pouze synapomorfních znaků pro rekonstrukci fylogeneze. Tedy toho znaku potomků, který vede ke společnému předkovi, u kterého daný znak vznikl. V praxi tento znak narazil na problém. Synapomorfní znak nelze tak snadno odlišit od jiných znaků homoplazií. Proto kladistická metoda založená na maximální parsimonii řeší problém obdobným způsobem, jaký najdeme ve fenetickém přístupu. [11]

1.3.1 Výhody a nevýhody molekulárních znaků

Mezi hlavní výhody patří to, že molekulární znaky jsou genetické a v hojném počtu. Jsou selekčně neutrální, takže dokážeme rozeznat homologii a homoplazii. Jsou použitelné od určení již vymřelých druhů až po porovnání současných jedinců. Můžeme je jednoznačně popsat (moc možností není, máme 4 nukleotidy a 20 aminokyselin). [11]

Mezi nevýhody patří vyšší finanční náklady oproti klasické taxonomii. Dále neposkytují informaci o anagenezi, protože se s velkou pravděpodobností neprojeví ve fenotypu.

1.4 Druh

Druh je definován jako samostatný taxon, který je řazen nejbližší pod taxon rodu. Rod je vysvětlován jako soubor druhů. Samotná definice druhu má několik koncepcí. Zde jsou uvedeny ty nejpodstatnější. Podle taxonomické koncepce pojetí druhu tvoří samostatný druh jedinci, kteří mají morfologicky shodné znaky. Z této koncepce vychází klasická taxonomie, která navíc přidává další kritéria, jako je geografické místo a čas. [9]

Za biologickou koncepcí druhu je považována definice z roku 1942 pronesená biologem Ernstem Mayerem. „Druh je skupina aktuálně nebo potenciálně se křížících populací, které jsou reprodukčně izolovány od jiných takových skupin.“ Biologickou koncepcí aplikoval v jeho době na lokální flóru v Severní Americe. Koncepce byla použitelná pro 93,5 % taxonů. [9]

Tyto koncepce nereflektovaly časový vývoj jednotlivých druhů, proto vznikla i evoluční koncepce druhu. Poprvé byla navržena Georgem G. Simpsonem. Koncepce vznikla na podkladě rozvoje paleontologie a evoluční biologie. Podle ní je druh jediné pojetí předka a potomka, které má své vlastní evoluční tendence a historii. [9]

Koncepce, která sjednocuje všechny hlediska dosud popsaných pojetí druhu, je fylogenetická. Podle ní je druh nejmenší možná skupina pohlavně se rozmnožujících organismů, které vlastní alespoň jeden diagnostický znak, jenž je přítomný u všech

členů skupiny, ale chybí u všech blízkých příbuzných této skupiny. [9]

1.5 Variabilita znaků

Variabilitou znaků je nazývána různorodost projevu znaku. Pro taxonomické určení organismů je variabilita znaků velmi podstatná. U jedinců každého organismu existuje vnitrodruhová variabilita. Tato vlastnost ukazuje, jak moc si jsou jedinci stejného druhu podobní mezi sebou. V kontrastu s touto vlastností je mezidruhová variabilita, příkladem toho, že každý druh se navzájem liší od ostatních druhů.

Vnitrodruhová neboli intervariabilita může být genetická nebo negenetická. Toto rozdělení se odlišuje podle původu znaků organismu. Genetická variabilita je dána genotypem. V jádrech buněk najdeme chromozomy, které jsou získány kombinací (crossing-over) z matčiny a otcovy strany. Touto kombinací je zajištěna proměnlivost potomků. [9]

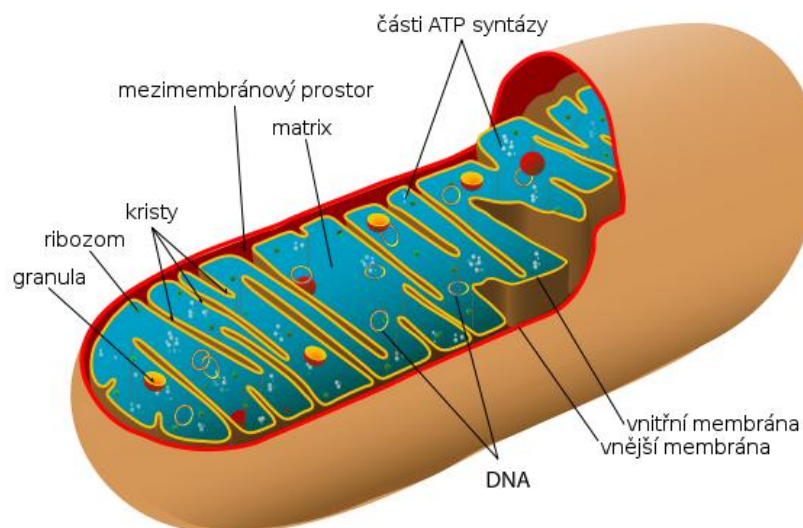
Negenetická variabilita je dána spoluúčastí prostředí a několika faktorů. Jedinci jsou variabilní v čase, každý dva jedinci se mezi sebou mohou odlišovat, protože každý z nich se nachází v jiném vývojovém stádiu. Mohou nastat i periodicky se opakující situace, kdy se jedinci navzájem od sebe liší (příkladem může být květenství stromů). Avšak odlišné situace mohou být podmíněny i prostředím. Organismy reagují na změny prostředí fenotypovým přizpůsobením. [9]

Určitou odchylkou od variability vzniklé evolučním vývojem nebo pouhým dospíváním organismu jsou například znaky, které se objevují v přechodné formě. Z hlediska genetiky mohou nastat poruchy vývoje, které jsou nejčastěji zapříčiněny určitým typem mutace (více v kapitole Mutace). [9]

2 MITOCHONDRIE

Mitochondrie jsou velmi malé orgány od 0,5 do 2 μm . Jsou tvořené dvojitou membránou bohatě obsahující transmembránové komplexy dýchacího řetězce, syntézu ATP a membránové přenašeče. Vnitřek mitochondrie je členěn v krysty a vyplněn matrix, ve které najdeme i volně plovoucí mitochondriální DNA. Stavbu mitochondrie představuje obrázek 3. [24]

První hypotézou původu mitochondrií je endosymbiontní hypotéza. Podle této hypotézy byly mitochondrie původně volně žijícími organismy bakteriálního typu, pro které se postupem času ukázalo efektivnější spolupracovat s jinou buňkou v rámci symbiózy. Buňka poskytovala mitochondriím ochranu a na oplátku mitochondrie poskytovala buňce produkty oxidativní fosforylace a to ATP, které je zdrojem energie pro buňku. Objevením DNA mimo jádro hlavní buňky se do popředí dostala další evoluční teorie, která říkala, že se původně struktura mitochondrie oddělila od jádra.



Obrázek 3: Stavba mitochondrie (převzato z [2])

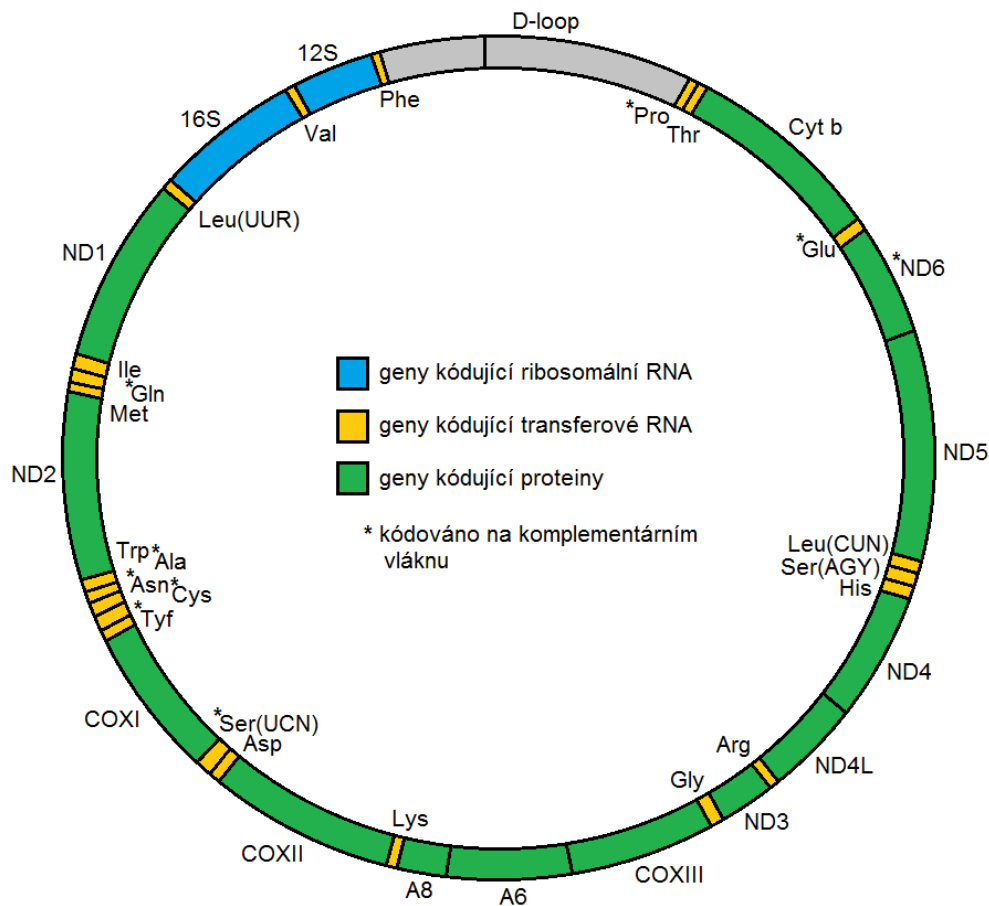
2.1 Mitochondriální DNA

Při studiích mitochondrií elektronovým mikroskopem se podařilo izolovat neporušenou a čistou mitochondriální DNA (mtDNA). Tato mtDNA byla většinou kruhového tvaru, jak ukazuje obrázek 4, ale můžeme ji také najít ve formě spojujících se dimerů nebo

vyšších oligomerů, s délkou odpovídající přibližně 16 000 nukleotidů u vícebuněčných organismů. První mtDNA byla prozkoumána u lidí a u živočichů, teprve později se začalo se zkoumáním mitochondriálního genomu rostlin. První kompletní mitochondriální rostlinný genom, a to porostnice mnohotvárné *Marchantia polymorpha*, byl zveřejněn teprve po roce 1992. [24]

Mitochondriální DNA se dědí převážně od matky, to zaručuje hromadění mutací v mitochondriálním genomu u každé nově vzniklé linie. Původní předpoklad výhradní mateřské dědičnosti vylučoval možnost rekombinace otcovské a mateřské mtDNA. Avšak v roce 1992 bylo dokázáno, že se u myši dědí mtDNA i z otcovy strany, a to pomocí bičíku spermie. Po tomto důkazu byla výhradnost přenosu mtDNA po mateřské linii odvolána. [17]

Mutační rychlost v mtDNA je asi 5 až 10 krát větší než u jaderného genomu [9]. Už ve studii publikované roku 1979 W. Browne, M. George a A. Wilson objevili, že v mtDNA bylo více odlišností mezi vzorky člověka a vzorky opic starého světa (oddělení úzkonosí) než bylo mezi samotnými primáty v jejich jaderné DNA. U rostlin najdeme přibližně 20 krát delší nukleotidovou sekvenci mtDNA než u živočichů a hub. U živočichů byla délka sekvence mtDNA zredukována na minimum. Zatímco u hub a rostlin najdeme mnoho neživočišných sekvencí genů tvořících introny, které dokonce obsahují úseky ORF, což jsou úseky DNA vymezené iniciačním a terminačním kodonem kódujícím polypeptidový řetězec. Pokud nejsou brány v úvahu intergenové sekvence, pak je genetická informace podobná u většiny organismů. To může být díky tomu, že ztráta genů z původního symbionta byla v podstatě dokončena dříve, než došlo k rozsáhlému větvení fylogenetického stromu. [24]



Obrázek 4: Mitochondriální genom (převzato z [24])

2.1.1 Mitochondriální genom u metazoa

Velikost rozsahu mtDNA nalezené u mnohobuněčných živočichů je poměrně úzká, přibližně 16,5 kb (kilobází), s některými výjimkami v rozmezí od 14 kb pro háďátka obecné z kmene hlístic, až po 42 kb pro mořskou hřebenatku *Placeopecten megallanicus*. Všechny mají tvar jednotlivých kruhových mtDNA až na jednu lineární mtDNA skládající se ze dvou molekul, nalezenou u nezmara štíhlého. Živočišná mtDNA kóduje dvě ribosomální RNA, 22 tRNA (s výjimkami), a 13 proteinů. [24]

2.1.2 Mitochondriální genom u rostlin

Je z pravidla označován za nejzajímavější z hlediska molekulární biologie. U tohoto genomu najdeme nejvíc mutací, které se podílejí na růstu délky sekvence mtDNA, ale i na nepoužitelnosti určitých úseků sekvence. Dosahuje délky až 2 400 kb, bez toho aniž bychom brali v potaz rozsah prokaryotického genomu. MtDNA se u rostlin nachází buď v kruhové anebo lineární podobě, ale ani pomocí elektronového mikroskopu se

nepodařilo zjistit, zda se nejednalo u lineární mtDNA pouze o prasknutí kruhové mtDNA. Genové uspořádání v mtDNA u rostlin je velmi variabilní, a to dokonce i mezi blízce příbuznými rody. [24]

2.1.3 Mitochondriální genom v houbách

Velikost mitochondriálních genomů hub je typicky mezi 40 a 60 kb, s extrémy 20 kb pro *Schizosaccharomyces pombe* (také známé jako štěpné kvasinky) a 170 kb pro pečárku opásanou, kterou najdeme na člověkem ovlivněných místech. Velikost genomu hub je 3x až 4x větší než u živočichů. [24]

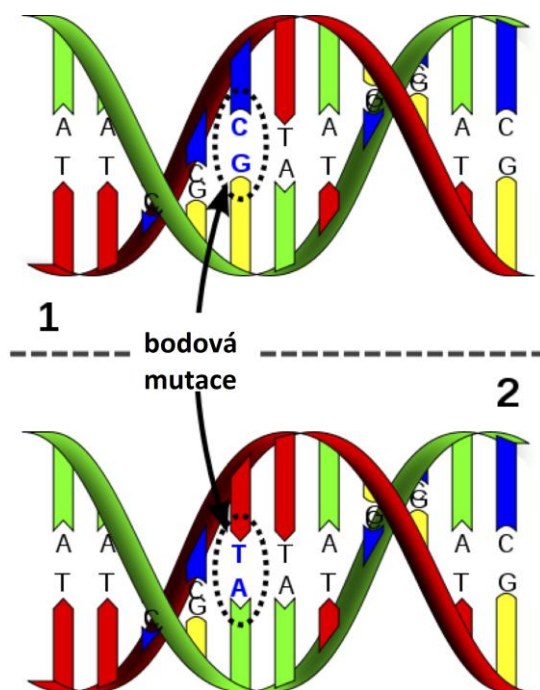
2.1.4 Mutace

Mutace v genomu mění smysl nukleotidů. Existují mutace na několika úrovních podle fyzické povahy, a to na úrovni bodové, úsekové, chromozomální a genomové. Vzhledem k tomu, že se tato práce zabývá nukleotidy, bude se dále soustředit na bodové mutace. Mutace mohou nastat jak v jaderné DNA, tak i v DNA jiných organel. Ukázku bodové mutace představuje obrázek 5. [10]

Většina replikací buněčné DNA se děje s velkou přesností, avšak vždy může docházet k chybám. Nejčastěji se jedná o bodové mutace, kdy dojde k inzerci (vlození), k delecí (vymizení) anebo k substituci jednoho nukleotidu. Nejčastější inzercí a delecí je jeden nukleotid, ovšem objevují se i dinukleotidové inserce a delece. Z hlediska chemické stavby dělíme nukleotidy na puriny a pirimidiny. Adenin (A) a guanin (G) je řazen mezi puriny a mezi pirimidiny patří cytosin (C), thymin (T) a uracil (U). V DNA je uracil nahrazen thyminem a v RNA zase naopak. Pokud dojde v rámci substituce k výměně purinu za purin nebo pirimidinu za pirimidin jedná se o tranzici. V případě změny pirimidinu za purin a naopak se jedná o substituci transverzní. Velká část těchto bodových mutací je detekována systémem oprav DNA a následně opravena. První velkou kontrolu zajišťují enzymy DNA polymerázy, které opravují přibližně 99% vzniklých chyb. Po replikaci dochází ke druhé kontrole, která snižuje zbylou 1% chybovost. Bohužel ani tento sofistikovaný postup na opravu mutací není dokonalý a k mutacím tedy dochází. [10]

Jakmile jednou buňka neopraví vzniklou mutaci, stává se tato mutace po dalším dělení buňky již mutací trvalou. Což znamená, že znova už tato mutace není detekována k opravě. Rozložení mutací v DNA není stejnoměrné. Obvykle se liší druh od druhu a také organela od organely. Ukazuje se, že existují náchylnější lokalizace ke vzniku mutací, a tím jsou horká místa (ang. hot spots). Také bylo dokázáno, že v genomech samců dochází k mutacím s větší frekvencí. U člověka je poměr rychlostí mutační

změny u samců 6x vyšší než u samic. V myší populaci byl stanoven poměr samců k samicím 2:1. Naopak u vášek rozdíl v mutační rychlosti patrně není vázán na pohlaví. Mutační rychlost není vázána všeobecně na pohlaví, ale záleží konkrétně na testovaném druhu. [7]



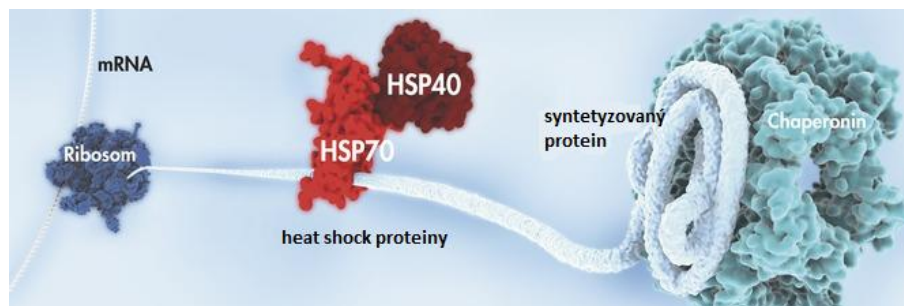
Obrázek 5: Ukázka bodové mutace (převzato z [20])

Určitý vliv na urychlení evolučního vývoje můžeme přiřadit k proteinům teplotního šoku (heat shock proteiny, HSP). Tyto proteiny najdeme jak u prokaryot, tak i u eukaryot. Proteiny teplotního šoku se podílí na stresové odpovědi buňky. V normálním stavu homeostázy jsou zodpovědné za správné funkční vytvoření proteinu. Ovšem ze strany prostředí na ně může být vyvíjen nátlak (tím dochází k narušení homeostázy) takového rázu, že dojde k překročení meze, kterou je schopen organismus zvrátit. [10]

Díky překročení meze dojde k produkci velkého množství HSP. Ta se může podílet na způsobení genetické změny, která se vlivem tlaku nekontroluje a neopravuje při replikaci a transkripci, vedoucí k produkci proteinů. Tudíž může dojít k aktivaci jiných genů než by mělo být podle kontroly genové exprese a vzniklý organismus může být rezistentní na tlak okolního prostředí. [17]

Obrázek 6 ukazuje zásah HSP do genové exprese. Podle tohoto odstavce se může zdát, že mutace by mohly být indukované buňkou. Tento závěr ovšem není pravdivý,

protože bylo prokázáno, že mutace jsou vždy spontánní, náhodné a nikdy ne indukované. [10]



Obrázek 6: Zásah HSP do genové exprese (převzato [12])

3 BIOLOGICKÁ IDENTIFIKACE POMOCÍ DNA BARCODINGU

DNA barcoding je molekulární a bioinformatický nástroj pro identifikaci biologických druhů. DNA barcoding není nová záležitost v oblasti molekulární taxonomie. Avšak zmínky o tomto přístupu k taxonomické identifikaci nenajdete v žádné české literatuře. Myšlenka DNA barcodingu by se dala přirovnat k čárovému kódu, kterým jsou označeny všechny výrobky v obchodě. Každý výrobek je jiný, a proto má každý svůj specifický čárový kód. Více výrobků stejného druhu má obvykle shodný čárový kód. To by se dalo přirovnat k jednotlivým taxonomickým úrovním v identifikaci jedince. U pokladny je pak tento čárový kód přečten, a pokud je výrobek načten správně, my se pak dozvídáme, například, jaký chleba jsme si koupili. Druh chleba by pak odpovídal výsledné katalogizaci organismu. U modernějších pokladních zařízení se obvykle načtou i další informace, což je nejen cena, ale i váha a firma, ve které byl daný výrobek udělán. Lze tento způsob aplikovat i na identifikaci jedinců?

3.1 Přístup DNA barcodingu

Přístup identifikace jedinců pomocí DNA barcodingu se poprvé dostal do popředí díky vědeckému článku kanadského profesora Paula Heberta publikovaného v lednu 2003. Profesor Hebert ovšem není považován za zakladatele tohoto přístupu k identifikaci jedinců, protože metoda používající sekvenci DNA k identifikaci byla popsána už dříve v roce 1982 (Nanney, 1982). Profesor Hebert ve svém článku zveřejnil svůj návrh výběru univerzální mitochondriální sekvence a to konkrétně *co1* nebo *cox1* (část genu cytochromu c oxidázy podjednotky I.). Svůj výběr tohoto úseku genu *co1* podložil praktickým testem na druhu motýlu *Lepidopteras*, kde měl 100% úspěšnost identifikace. Touto senzací si získal mnoho příznivců, kteří nadšeně praktikují právě metody DNA barcodingu i na jiné druhy, ale i mnoho odpůrců. DNA barcoding je analýzou, která odráží zastoupení intravariability a intervariability druhů, která je oddělena prahovou hodnotou. [13]

Ačkoliv původní myšlenkou byla aplikace DNA barcodingu na říši živočichů použitím právě markeru *co1*, později došlo k rozšíření i na kvetoucí rostliny a houby. Před pár lety se začalo zkoušet použití DNA barcodingu i na ostatní větve evolučního stromu, tedy i na bakterie a archea, pomocí multi - lokusové sekvence. [5]

Největší výhodou DNA barcodingu je identifikace organismů ve všech životních stádiích a dokonce se nemusíme obávat ani pohlavního dimorfismu, protože sekvence daného druhu jsou natolik stejné, že daný organismus je většinou správně zařazen. Rovněž lze dospět k správnému zařazení i z části kusu nebo plátku dotazovaného vzorku (např. u muzejních exemplářů). DNA barcoding může jednak provést identifikaci stávajícího druhu, ale navíc může i snadněji objevit druh nový. [5]

Mitochondriální genom obsahuje méně intronů (nekódujících sekvencí, jejichž význam není zcela znám) než v jaderném genomu. Také tento genom není vystaven velkému množství rekombinací a je haploidně přenášen převážně po mateřské linii. Navíc 13 genů kódujících protein v živočišné mtDNA obsahuje méně delecí, které jinak vedou k posunutí čtecího rámce. Neexistuje přesvědčivý důvod, proč by se analýza soustředila na specifický gen, ale *co1* má dvě důležité výhody. První je, že všeobecné primery pro gen jsou velmi odolné a umožňují tak opětovné užití 5' konce z většiny zástupců kmene živočichů, pokud ne ze všech. Druhou výhodou je, že *co1* vlastní větší škálu fylogenetického signálu než jakýkoliv jiný mitochondriální gen. *Co1* společně s dalšími protein-kódujícími geny, jejíž třetí pozice nukleotidů ukazuje vysokou četnost báзовých substitucí vedoucích k rychlosti molekulární evoluce, která je asi 3x větší než u velké nebo malé podjednotky ribozomální DNA (12S nebo 16S), je při evoluci tohoto genu dostatečně rychlá, aby rozlišila nejen blízce příbuzné druhy, ale i fylogenetické skupiny v rámci samostatných druhů. [3]

Existují studie používající k určení jiné úseky genů. Například úsek genu *cytochrom b* nebyl sledován dostatečným, protože se tento gen mění rychleji v rámci své aminokyselinové sekvence.

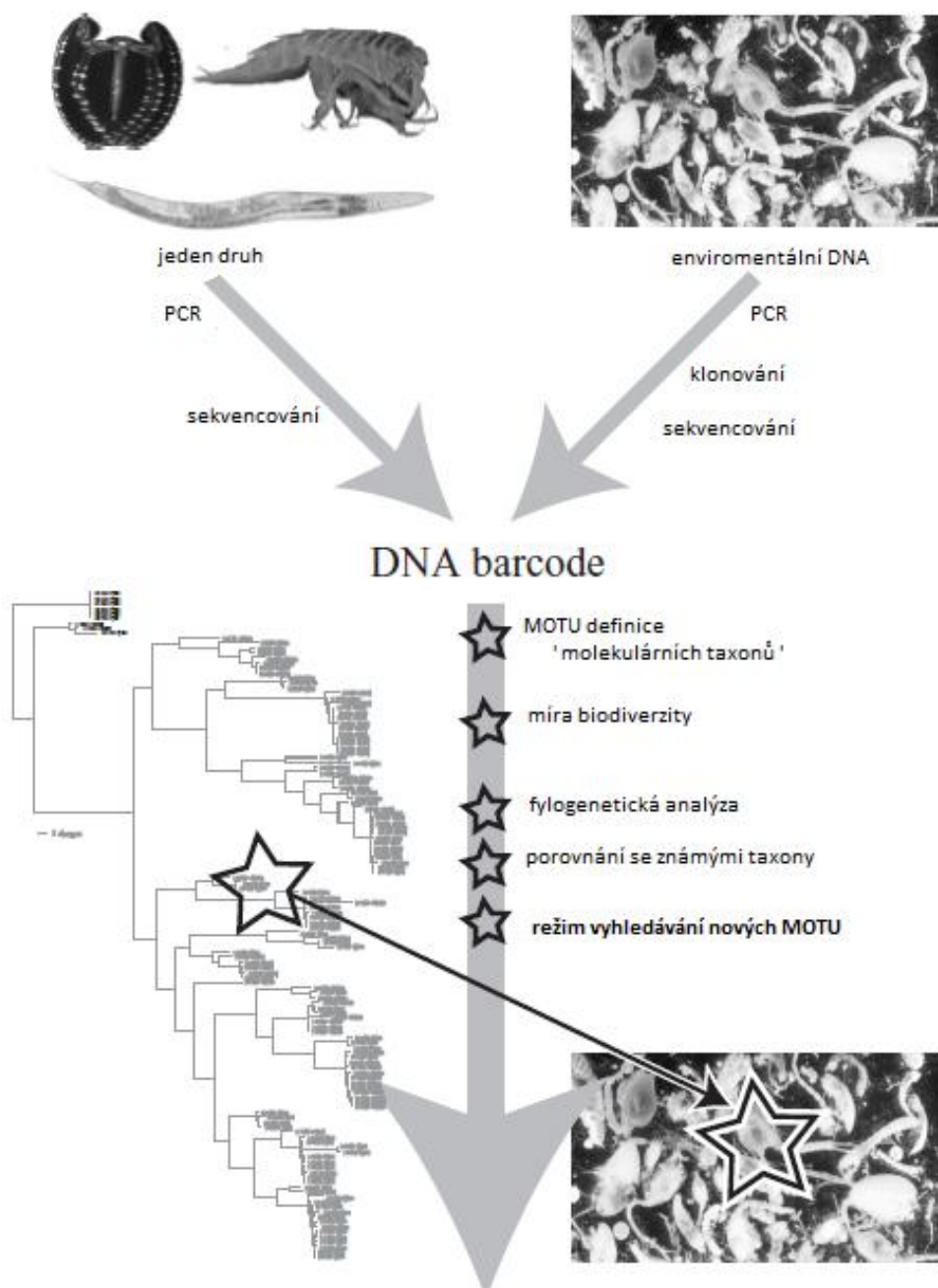
Postupem času se vytvořilo více referenčních sekvencí, které se liší podle daného organismu a jeho rychlosti divergence v evoluční linii. Protože pokud nemá úsek genu dostatečně rychlou divergenci, tak nedokáže oddělit a určit nové vývojové větve u daného druhu organismu. V současnosti se po snaze sjednocení definitivně používá pro metazoa (živočichy) *co1* marker, pro houby *ITS* (úsek RNA cistronu) a pro rostliny kombinace dvou lokusů *rbcL* a *matK*. Rovněž byla stanovena i základní délka sekvence. Pro *co1* nebo *rbcL* (chloroplasty ribulózy bisfosfátů karboxylázy velké podjednotky) nebo *matK* (úsek rostlinného plastidového genu) musí být sekvence dlouhá minimálně 500 pb (párů bází), zatímco pro *ITS* stačí alespoň 100 pb. [5]

3.2 Standardní postup analýzy

Nejprve proběhne získání extraktu z DNA, poté je provedena polymerázová řetězová

reakce (PCR) sekvence. Polymerázová řetězová reakce je biochemickou reakcí, využívající DNA polymerázu ke kopírování DNA. Využívá se k zmnožení DNA geometrickou řadou. Výsledné produkty PCR jsou sekvence, které se osekvenují, tj. získá se posloupnost jednotlivých nukleotidů. Takto získané sekvence jsou ukládány do databází, např. BOLD. DNA barcodingové sekvence v databázi BOLD neprochází žádnou nezávislou kontrolou. Sekvence v této databázi jsou uspořádané do projektů, což jsou soubory sekvencí určité skupiny organismů či všech organismů z jedné lokality. Sekvence v projektech lze analyzovat některými nástroji, které databáze poskytuje. Velká část prací vychází z toho, že použije jednoduchou analýzu matice vzdáleností (p-distance) s korekcí Kimurova dvouparametrického modelu. Tento model je nejčastějším a byl stanoven jako nejlepší korekční model DNA na nízkou substituční genetickou vzdálenost. Nakonec dojde k vytvoření fylogenetického stromu používající nejčastěji Neighbour - joining algoritmus. Celý postup získání barcode sekvence neboli čárového kódu sekvence reprezentuje obrázek 7. [5]

V případě, že chceme analyzovat sekvenci, která by již mohla být uložena v databázi, nahrajeme analyzovanou sekvenci do databáze BOLD. Díky vložené sekvenci dochází k vyhledání podobných sekvencí a to BLAST algoritmem. Po prohledání databáze dostáváme výpis sekvencí s nejnižší hodnotou podobnostního skóre. Ovšem ani nejnižší hodnota podobnostního skóre neznamená skutečně správně přiřazenou sekvenci s názvem druhu. Chyba může nastat v případě, že hledaný druh ještě nebyl zadán do databáze BOLD a samotný algoritmus BLAST nemusí nalézt nejpodobnější sekvenci. [16]



Obrázek 7: Postup jakým by měly být získávány sekvence z jednoho druhu (převzato z [4])

3.3 Organizace pro DNA barcoding

Současná správa DNA barcodingu je zastřešená organizací Consortium for the Barcode of Life (CBoL; <http://barcoding.si.edu/>). Nynějším cílem je zavedení zobecnění v taxonomii, které by umožnilo vědcům mnoha specializovaných oblastí pracovat ve společném rámci a sdílet data. To vedlo k vytvoření univerzálního systému pro uložení

sekvencí živých organismů a to BOLD systému. Ten do února 2010 získal 790 000 sekvencí, což pokrývá 67 000 formálně popsanych druhů. BOLD systém ukládá také podrobnosti, jako je morfologie, zeměpisné rozšíření, fotografie, sběrná místa a další. Dalším nástrojem je BioBarcode, která je zaměřena na sběr asijských organismů. [5]

3.4 DNA barcoding založený na znacích DNA

Oproti analýzám založeným na vzdálenostech, kde jsou porovnávány různé sekvence v jednom celku a následně jsou konstruovány stromy na základě celkových podobností, znakově založené modely hledají diagnostické, diskrétní znaky nebo jejich kombinace. Tyto znaky by měly být v DNA jedinečné, jedná se o tzv. jedinečné nukleotidové polymorfismy (JNP). Jednotlivé JNP jsou potom spojeny přes algoritmus a přiřazeny k znakově založenému barcode neboli čárového kódu, který je specifický pro každý taxon. Mezi tyto znaky se řadí kromě JNP DNA i znaky aminokyselin, expresních znaků a dalších atributů. Celkově tyto znakově založené čárové kódy vytvoří řídicí strom. Samotný algoritmus, systém charakteristických znaků (systém je znám pod názvem CAOS), je založen na základním konceptu, že členové dané taxonomické skupiny sdílí znaky (např. polymorfismy), které se nevyskytují v jiných skupinách. Algoritmus CAOS identifikuje specifický barcode, tedy znaky označované jako charakteristický atribut (CA), které jsou jedinečné pro každou vývojovou větev a každý větvící uzel uvnitř řídicího stromu. Řídicí strom je používán CAOSEm pouze jako prostředek k identifikaci diagnostických znaků. Což nemusí nutně představovat domnělé fylogenetické vztahy. [3]

Charakteristické znaky jsou stavy znaků, které jsou přítomny pouze v jednom souboru vývojově příbuzných biologických druhů, ale ne v alternativní skupině, která také sestupuje do uzlu. Charakteristické znaky jsou rozděleny na čtyři typy. Jednoduché čisté charakteristické znaky, vyskytující se pouze v určité vývojové větvi a ne v jiné. Jednoduché soukromé charakteristické znaky jsou sdílené pouze několika jedinci ve vývojové větvi. Složené čisté charakteristické znaky jsou kombinace stavů znaků, které se vyskytují pouze v rámci jedné vývojové větve a nikdy v celé kombinaci jiných vývojových větví a složené soukromé charakteristické znaky jsou stavy znaků nacházejících se jen u některých jedinců ve vývojové větvi, ale nikdy ne v ostatních vývojových větvích. [3]

CAOS při testování identifikace vážek na jednom genovém fragmentu bral v úvahu pouze jednoduché čisté a soukromé charakteristické znaky, které nalezneme alespoň u 80 % jedinců v taxonu. Z 833 vážek bylo správně zařazeno 22 rodů z 25 a na úrovni

druhů 54 z 64. Z této studie vyplývá, že DNA barcoding založený na znacích má velký potenciál. Je ovšem důležité si uvědomit, že u identifikace nových druhů nemůžeme slepě věřit DNA barcodingu. Výsledky by se měly potvrdit klasickými taxonomickými postupy. Pokud je ovšem už daná sekvence katalogizována a ověřena, je velmi jednoduché identifikovat organismus pomocí DNA barcodingu založeného na znacích.[3]

DNA barcoding si svou jednoduchostí našel široké uplatnění. Přes použití ke kontrole druhů prodávaných ryb na rybím trhu [26] až po detekci patogenů tuberkulózy, Q-horečky a dalších onemocnění díky barcode neboli díky čárovému kódu jejich přenašečů. [3]

4 FYLOGENETICKÉ STROMY

Fylogenetický strom je grafickým vyjádřením příbuzenských vztahů mezi jednotlivými taxonomickými jednotkami. Předpokladem pro vytvoření fylogenetického stromu je příbuznost druhu organismů. Tyto organismy sdílejí společnou evoluční historii, tedy mají společného předka. Vědním oborem popisující vývoj organismů je fylogenetika. Evoluční vývoj je časově náročný a jen stěží může být zkoumán pouze několika generacemi vědců. Oproti tomu ontogeneze zkoumá vývoj organismu od narození do dospělosti. Existují tři směry zabývající se konstrukcí fylogenetických stromů. [9, 11]

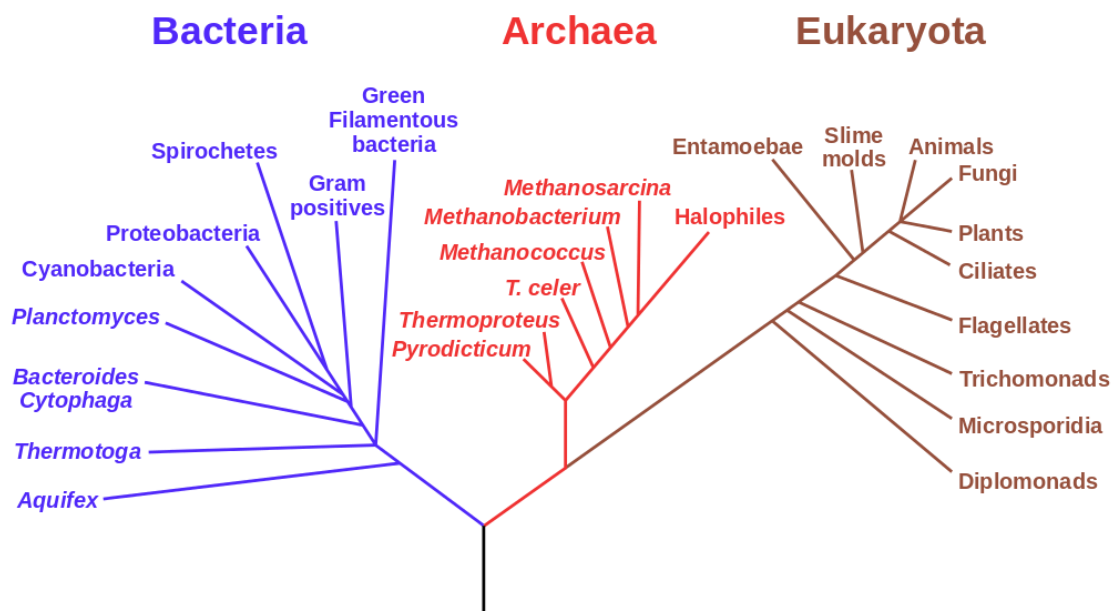
Jedním směrem fylogenetiky je fenetika. Často označována jako numerická taxonomie. Tento směr konstruuje fylogenetické stromy především podle stupně jejich celkové podobnosti nebo odlišnosti. Znaky se hodnotí bez ohledu na to, jakým způsobem vznikaly. Tento bod přístupu fenetiky je často kritizován, protože ne všechny shodné znaky jsou odvozeny od jednoho společného předka.[10, 11]

Dalším směrem, který se zabývá tvorbou fylogenetických stromů, je fylogenetická systematika neboli kladistika a posledním je evoluční taxonomie. Kladistická disciplína popisuje vznik a vývoj jednotlivých linií organismů na základě rozluštění genetického kódu. Předpokládá, že vývojově nové druhy vznikají vidličnatým dělením z původní linie druhu. Vznikají dva nové druhy, ale původní druh zaniká. Poprvé byl tento směr řazení organismů popsán v práci německého biologa Willi Henninga „Grundzüge einer Theorie der phylogenetischen Systematik“ v roce 1950. Tímto dílem byl položen základ vysoce objektivní metody pro klasifikaci organismů.[10, 11]

Evoluční taxonomie staví na základech kladistiky. Tento směr, ale respektuje rozdílnost znaků u skupiny organismů podle evolučního stupně vývoje. V tomto přístupu jsou zahrnuty podobnosti, které byly získány z molekulární taxonomie. Jedná se o podobné oblasti nukleotidů nebo aminokyselin pro příbuzné druhy. Zároveň jsou zde k porovnávání příbuznosti jedinců použity algoritmické postupy, které se postupem času vyvíjejí. Za nejvýznamnější představitele v tomto směru byli považováni Ernst Mayr a George Gaylord Simpson. [11]

Pro laickou veřejnost je pojem fylogenetický strom znám spíše jako „strom života“, který je na obrázku 8, nebo také „rodokmen“. Příbuzenské vztahy jsou zjišťovány pomocí morfologických nebo genetických aspektů. Dřívější fylogenetické stromy, a to až do padesátých let minulého století, byly konstruovány na základě morfologické podobnosti organismů. S postupem molekulární taxonomie do popředí se začaly hledat

objektivní způsoby konstrukce fylogenetických stromů z molekulárních dat. V současné době se využívají k porovnávání příbuznosti sekvence bází v genomech jednotlivých biologických druhů nebo lze také použít informace vycházející z proteinových sekvencí.[11]



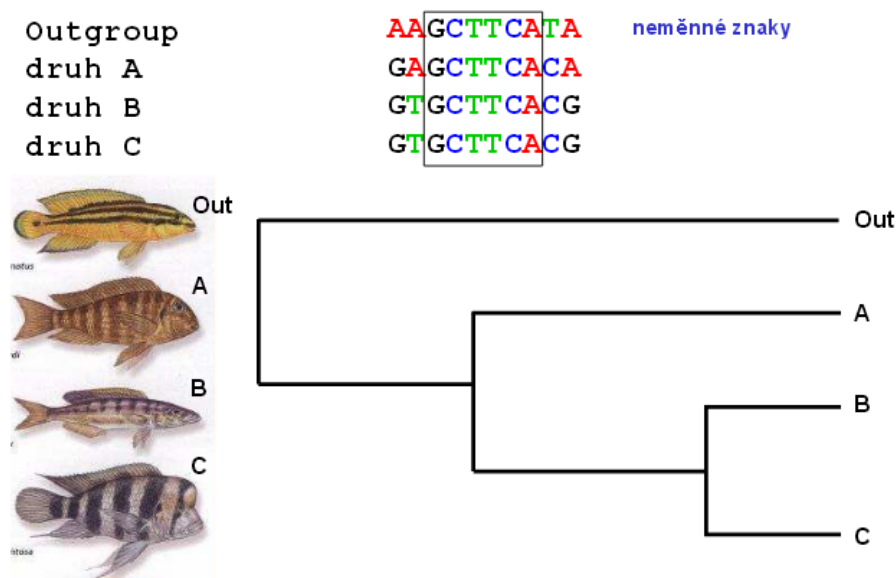
Obrázek 8: Fylogenetický strom života (převzato z [11])

Slovo strom zde představuje acyklický graf. Jednotlivé organismy představují větve stromu, které končí terminálním uzlem (názvem testovaného jedince nebo skupiny jedinců). V případě, že se strom rozvětňuje pouze do dvou větví, jedná se o bifurkaci. Pokud je počet terminálních uzlů větší než dva už nejde o případ bifurkace, ale multifurkace. Vrcholy nebo uzly které jsou spojeny s dalšími dvěma či více vrcholy nazýváme vnitřní vrcholy. Pokud je vrchol spojen pouze s jedním dalším vrcholem nazýváme ho listem. Každý vrchol u fylogenetického stromu představuje určitou taxonomickou jednotku. Hrana mezi vrcholy reprezentuje příbuzenský vztah mezi jednotlivými taxonomickými skupinami. U některých metod konstrukce fylogenetických stromů je možné určit délku hrany a tím i podobnost v závislosti na procentu odlišných bází mezi sekvencemi. Zároveň může délka této hrany poukázat na délku vývoje příslušné taxonomické skupiny. V závislosti na tom, co má délka hrany ukazovat, musíme zvolit příslušnou metodu vytvoření fylogenetického stromu. Existuje mnoho metod ke zkonstruování fylogenetických stromů. Metody se dělí podle dat, která jsou používána na testování. Pro znaková data jsou vhodné metody maximální

parsimonie a maximum likelihood. Pro data distanční se nejčastěji používají metody Neighbor - joining a UPGMA. Tato práce dále využívá právě ke konstrukci fylogenetických stromů pro referenční sekvence metodu Neighbor - joining (překlad spojování sousedů) podle hodnot distancí mezi sekvencemi s aplikací Kimurova dvouparametrického evolučního modelu, proto bude věnována pozornost právě této metodě konstrukce fylogenetického stromu.[10, 11]

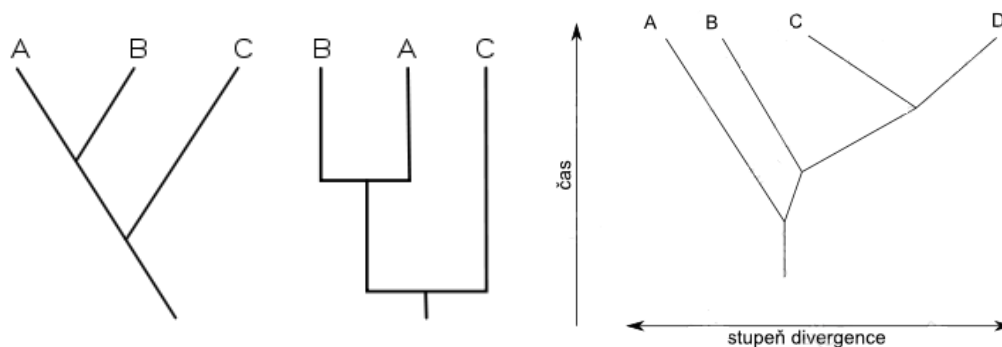
Fylogenetické stromy můžeme rozdělit na kořenové a nekořenové. Nekořenový fylogenetický strom porovnává příbuznost taxonomických jednotek, aniž by se specifikoval společný předek. U kořenových fylogenetických stromů je společný prapředek, který tvoří kořen stromu. Ke kořeni jsou v závislosti na příbuznosti zařazeny jednotlivé taxonomické jednotky. U kořenového stromu můžeme rozlišovat jednotlivé znaky potomků na pleziomorfní (znaky pocházející od předka) a apomorfní (odvozená forma znaku). [11]

Nejčastější metodou zakořeňování stromů je outgroup. Tato metoda spočívá v tom, že vezmeme blízce příbuzný taxon, který se odštěpil v dřívější době od skupiny testovaných taxonů. Po konstrukci fylogenetického stromu by se měl vybraný outgroupový taxon objevit jako společný prapředek testovaných souborů, jak je vidět na obrázku 9. Této metody budeme využívat i v praktické části. [11]



Obrázek 9: Ukázka metody outgroup (převzato z [11])

Dále fylogenetické stromy dělíme na základě informativnosti délky větví. Pokud délka větví ani jejich úhel nenese žádnou informaci o vývoji, jedná se o dendrogram (nebo kladogram, který zobrazuje genealogické vztahy). V případě, že délka větví koresponduje s divergencí a úhel vyjadřuje rychlost evoluce jednotlivých organismů, jedná se o fylogram (nebo fenogram pro fenotypovou podobnost). Příklad jednotlivých fylogenetických stromů je na obrázku 10.



Obrázek 10: Příklad kladogramu vidíme vlevo (první dva) a fylogramu vpravo

Kimurův dvouparametrický model je využíván převážně proto, že rozlišuje dva typy substitucí u nukleotidů. Těmito substitucemi se mohou sekvence lišit a jsou tedy důležitým parametrem při porovnávání na základě procenta odlišných bází mezi sekvencemi. Substituce, které model rozlišuje, jsou tranzice a transverze. Tranzice je výměna purinu za purin nebo pyrimidinu za pyrimidin. K transverzi dochází při změně purinu na pyrimidin nebo naopak. Detailnější popis je uveden v kapitole Mutace. Konstrukce metodou Neighbor - joining na počátku vytvoří jeden hvězdovitý strom. Jednotlivé listy reprezentující taxonomické jednotky se sbíhají do vnitřního uzlu. Při samotném porovnávání příbuznosti jednotlivých taxonomických jednotek se pak strom rozkládá tak, aby v každém kroku byla co nejmenší vzdálenost celkové délky stromu. [11]

5 NUMERICKÉ REPREZENTACE

Pomocí numerických reprezentací dokážeme vhodně předzpracovat genomická a proteomická data. V dalších odstavcích se zaměříme pouze na genomická data. Vstupní parametr numerických reprezentací se skládá ze sekvence nukleotidů $s = s[n]$, která je tvořena bázemi adenin, guanin, thymin a cytosin, a délky sekvence značené N . Poté, co je známa sekvence, je zvolena metoda analýzy, která převádí sekvenci do numerického formátu. Tímto procesem je vytvořena numerická mapa. Sekvence má informativní charakter různých vlastností, které by měli být zahrnuty v numerickém formátu.

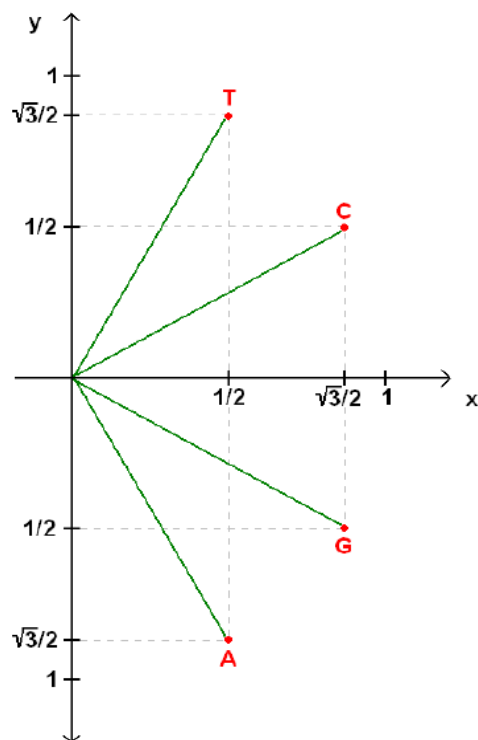
Existuje mnoho druhů numerických map. Podle typu konečného zpracování je volena odpovídající mapa. Mapa může zvýraznit požadované vlastnosti sekvence a potlačit zavádějící informace. Tedy ty informace, které pro analýzu nejsou podstatné. V tomto případě se jedná o tzv. redukované formy numerických map. Také jsou druhy map, které zohledňují všechny vlastnosti. Níže jsou vysvětleny tři vybrané metody numerické reprezentace. [16]

5.1 2D reprezentace v 1. a 4. kvadrantu

Výsledkem této metody je zobrazení DNA v 1. a 4. kvadrantu kartézského souřadného systému. Každému nukleotidu jsou přiřazeny souřadnice na osách x a y [3]:

$$\begin{aligned}\left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right) &\rightarrow A & \left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right) &\rightarrow C \\ \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right) &\rightarrow G & \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right) &\rightarrow T\end{aligned}$$

Výhodou této metody je, že lze vizuálně kontrolovat data. Zvolili jsme kumulovanou verzi této metody. Ta spočívá v tom, že každá další hodnota následujícího nukleotidu je připočtena k předcházejícím hodnotám předešlého nukleotidu. Tedy u této verze metody nedochází k tvorbě smyček, a tím ke ztrátě informace bez možnosti zpětné rekonstrukce sekvence ve správném pořadí. Metoda pomáhá při rozpoznávání velkých rozdílů mezi podobnými sekvencemi DNA. Na následujícím obrázku (obrázek 11) vidíme reprezentaci pyrimidinových sekvencí v 1. kvadrantu a purinových sekvencí ve 4. kvadrantu. [16]



Obrázek 11: Reprezentace nukleotidů v 1. a 4. kvadrantu kartézského souřadného systému (převzato z [16])

5.2 EIIP hodnoty

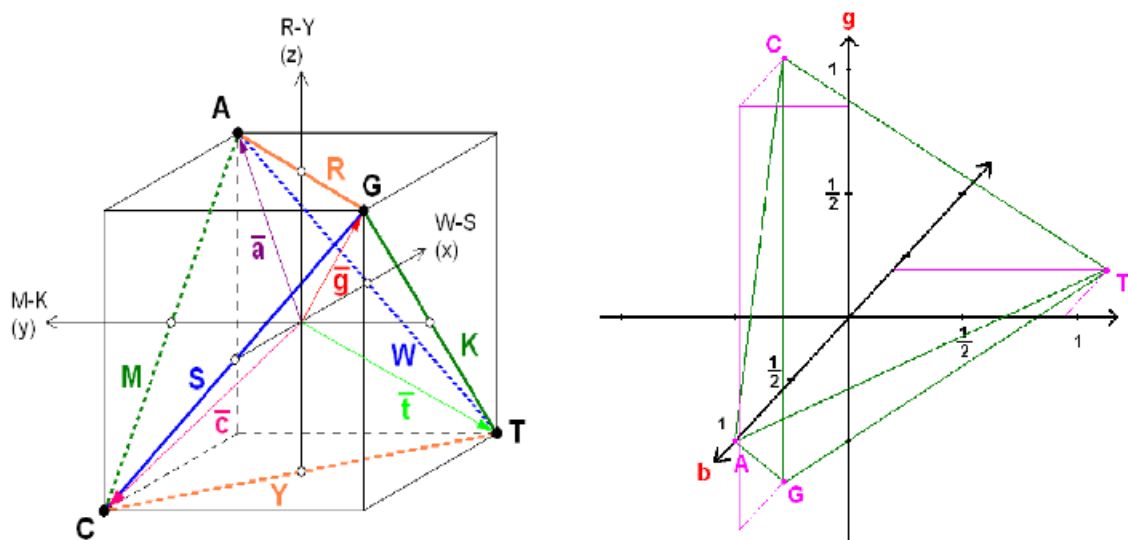
Elektron – iont interakční potenciál (EIIP) je numerická hodnota. Tato hodnota odpovídá jednotlivým nukleotidům v sekvenci a zároveň i jejich průměrné energii valenčních elektronů. Přehled hodnot ukazuje tabulka 1. Výsledkem této numerické mapy nukleotidů je číselný vektor, který reprezentuje DNA sekvenci. V některé literatuře se dočteme, že EIIP hodnoty se uvádějí v rydbezlích (Ry). V ostatních literaturách je tato numerická hodnota bezrozměrná. Tento druh numerické reprezentace se často používá v kombinaci s diskretní Fourierovou transformací pro převedení do frekvenčního spektra. Zde se pak vyhodnocují jednotlivé společné složky mezi sekvencemi. Metoda EIIP hodnot se používá k detekci kódujících a nekódujících oblastí a také k specifikaci účinků léčiv. [1]

Tabulka 1: Hodnoty EIIP pro jednotlivé nukleotidy (převzato z [1])

Nukleotid	EIIP
A	0,1260
C	0,1340
G	0,0806
T	0,1335

5.3 Třívektorová reprezentace

Jedná se o metodu, která vznikla redukcí nukleotidového čtyřstěnu (obrázek 12 vlevo). Jednotlivým nukleotidům byly přiřazeny souřadnice v 3D prostoru. Oproti čtyřstěnu sice jde o redukcí z 4D prostoru na 3D prostor, ale vše je bez ztráty informace. [7], [15]



Obrázek 12: Nukleotidový čtyřstěn (vlevo) a umístění nukleotidového čtyřstěnu v RGB prostoru (vpravo) převzato z [16]

Nukleotidový čtyřstěn byl zredukován na pravidelný čtyřstěn, který vidíme na obrázku 12 vpravo. Jednotlivým nukleotidům umístěných ve vrcholu byly přiřazeny nové souřadnice (1) [3].

$$\begin{aligned}
A &= (a_r, a_g, a_b) = (0, 0, 1) \\
C &= (c_r, c_g, c_b) = \left(-\frac{\sqrt{2}}{3}, \frac{\sqrt{6}}{3}, -\frac{1}{3}\right) \\
G &= (g_r, g_g, g_b) = \left(-\frac{\sqrt{2}}{3}, -\frac{\sqrt{6}}{3}, -\frac{1}{3}\right) \\
T &= (t_r, t_g, t_b) = \left(\frac{2\sqrt{2}}{3}, 0, -\frac{1}{3}\right)
\end{aligned} \tag{1}$$

Výsledkem této metody je zpracování sekvence na tři vektory x_r , x_g a x_b . Rovnice (2) se skládají z indikačních vektorů u_A , u_C , u_G a u_T , které ukazují přítomnost jednotlivých nukleotidů v sekvenci [3]:

$$\begin{aligned}
x_r[n] &= \frac{\sqrt{2}}{3} (2u_T[n] - u_C[n] - u_G[n]) \\
x_g[n] &= \frac{\sqrt{6}}{2} (u_C[n] - u_G[n]) \\
x_b[n] &= \frac{1}{3} (3u_A[n] - u_T[n] - u_C[n] - u_G[n])
\end{aligned} \tag{2}$$

Písmeno n v rovnici značí pořadí jednotlivých nukleotidů. Tento výsledek se používá k zobrazení spektrogramu sekvence v RGB spektru.

6 PRAKTICKÁ ČÁST

V programovém prostředí Matlab byly vytvořeny funkce pro převod DNA sekvencí do zvolených numerických reprezentací, pro vytvoření referenčních signálů pro druh a také pro následnou druhovou identifikaci dalších sekvencí. Převod sekvencí využívá funkcí, které jsou popsány v kapitole Numerické reprezentace (*kvadrant_repre* pro numerickou reprezentaci 1. a 4. kvadrantu, *eiip* pro EIIP hodnoty, *eiipkum* pro kumulované EIIP hodnoty a *rgb_repre* pro třívektorovou reprezentaci). Popis numerických metod je detailně popsán v kapitole numerických reprezentací. Program je tvořen stěžejní hlavní funkcí, která si volá pomocné funkce. Reference jsou vytvářené ze sekvenčních souborů obsahující druhy netopýrů z Guyany (název souboru je BCBNC), chrostíků (CUCAD), neotropických ptáků (MNCN) a subtropických ryb (FCFP). Přiřazení se děje na základě jejich největší podobnosti v rámci získaného úseku mitochondriální DNA. Pro přiřazení reference k testovanému druhu je možné vybírat z metod pro numerické zpracování. Tyto metody si volí uživatel sám v programovém prostředí Matlab.

6.1 Programové řešení

Program je určen pro programové prostředí Matlab, které obsahuje i pomocný bioinformatický toolbox, jenž je programem využíván. Nutnost tohoto toolboxu je podmíněna funkcí *fastaread*, která je potřeba pro načtení souborů z FASTA formátu. FASTA formát je uveden popisnou hlavičkou, ve které najdeme mimo jiné i jméno druhu jedince, a sekvenci. Hlavička obsahuje jméno souboru, kódované informace, jméno druhu organismu a místo původu DNA sekvence. Pro program je prioritní získat z hlavičky jméno druhu organismu, ke kterému patří sekvence mtDNA. Jméno druhu se v hlavičce obvykle nachází na čtvrtém místě. Jednotlivé kódy a jména jsou odděleny tímto znakem „|“. Pro správnou funkci programu je hlavička upravena funkcí s názvem *prejmenovani*, tak aby obsahovala pouze jméno druhu organismu.

Po úpravě hlavičky organismu je dále potřeba zkontrolovat sekvenci nukleotidů, zda neobsahuje nechtěné znaky „N“ a „-“. Znak „N“ vyjadřuje neupřesněný nukleotid a znak „-“ vyjadřuje mezeru v sekvenci v případě, že je sekvence zarovnána. Pokud by se v sekvenci tyto znaky vyskytly, budou smazány.

Z takto načtených a upravených souborů je potřeba vytvořit reference pro jednotlivé druhy organismů. Pro reference byly záměrně vybrány soubory, které obsahují větší počet sekvencí z části mitochondriálního genu *cox1* pro jednotlivé druhy.

Pro každý druh organismu jsou sekvence převedeny na numerickou reprezentaci podle zvolené metody zpracování (1. a 4. kvadrant, EIIP hodnoty, kumulované EIIP hodnoty a třívektorová reprezentace). Jednotlivé sekvence pro stejný druh jsou zarovnány, tak aby se překrývaly v nejpodobnějším úseku vzhledem k nejdelší sekvenci druhu. Zarovnání je realizováno *for* cyklem, kde posunujeme menší sekvenci vzhledem k druhé delší sekvenci a hledáme právě tu pozici, ve které bude vypočtený rozdíl nebo euklidovská vzdálenost sekvencí nejmenší. Takto zarovnané numericky reprezentované sekvence jsou zprůměrovány. Výsledný průměr je uložen jako referenční sekvence pro daný druh do proměnné *matice* v programovém prostředí Matlab.

U numerických reprezentací, které se nacházejí v kumulované formě, je potřeba odstranit trend tak, aby postup na ose x byl roven kroku jedna. Jedná se o metody 1. a 4. kvadrantu a kumulované EIIP hodnoty. Kumulovaná forma znamená, že další hodnota je připočtena k předešlému výstupu (hodnotě). V případě neodstranění trendu by docházelo i vlivem malé záměny nukleotidů k velkému číselnému rozdílu, a to i v případě shodných nukleotidů (protože závisí na předchozí numerické hodnotě). Navíc vlivem kumulace by tento rozdíl dále narůstal a hrozilo by špatné přiřazení referenčního druhu k testovanému. K odstranění trendu je využita funkce *polyfit*, kdy se průběh kumulované numerické reprezentace aproximoval polynomem třetího stupně. Pro zjištění hodnot je použita funkce *polyval*. Ze zjištěných hodnot se sestavily rovnice pro přímku, díky kterým se získali hodnoty osy x , tak aby výpovědní hodnota numerické reprezentace kumulované formy byla zachována a krok na ose x se rovnal jedné.

Samotné přiřazení reference k testovanému druhu se skrývá pod funkcí s názvem *druhy_soubor*. V této funkci se postupně přiřazuje jméno referenčního druhu k testovanému druhu. Sekvence testovaného druhu je převedena na numerickou reprezentaci a zkontrolována, jestli neobsahuje nechtěné znaky. Samotná testovaná numerická reprezentace je dále porovnána s každou referenční reprezentací. Dvojice vektorů převedených sekvencí je zarovnávána tak, aby se překrývaly v té nejpodobnější oblasti. Princip zarovnání je zobrazen v tabulce 2, k výpočtu byly použity vzorce (3) a (4). Nejpodobnější oblast zjistíme pomocí euklidovské vzdálenosti (D_E) nebo pomocí rozdílu (D). Ve vzorci (3) jsou uvedené hodnoty na osách x a y , které jsou potřeba k výpočtu euklidovské vzdálenosti nebo k výpočtu rozdílu (4).

$$D_E = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (3)$$

$$D = \sum_{i=1}^n |x_i - y_i| \quad (4)$$

Pro souřadnice bodu je použita osa x a y . (v našem případě převedené sekvence na číselnou podobu) a n označuje délku sekvence

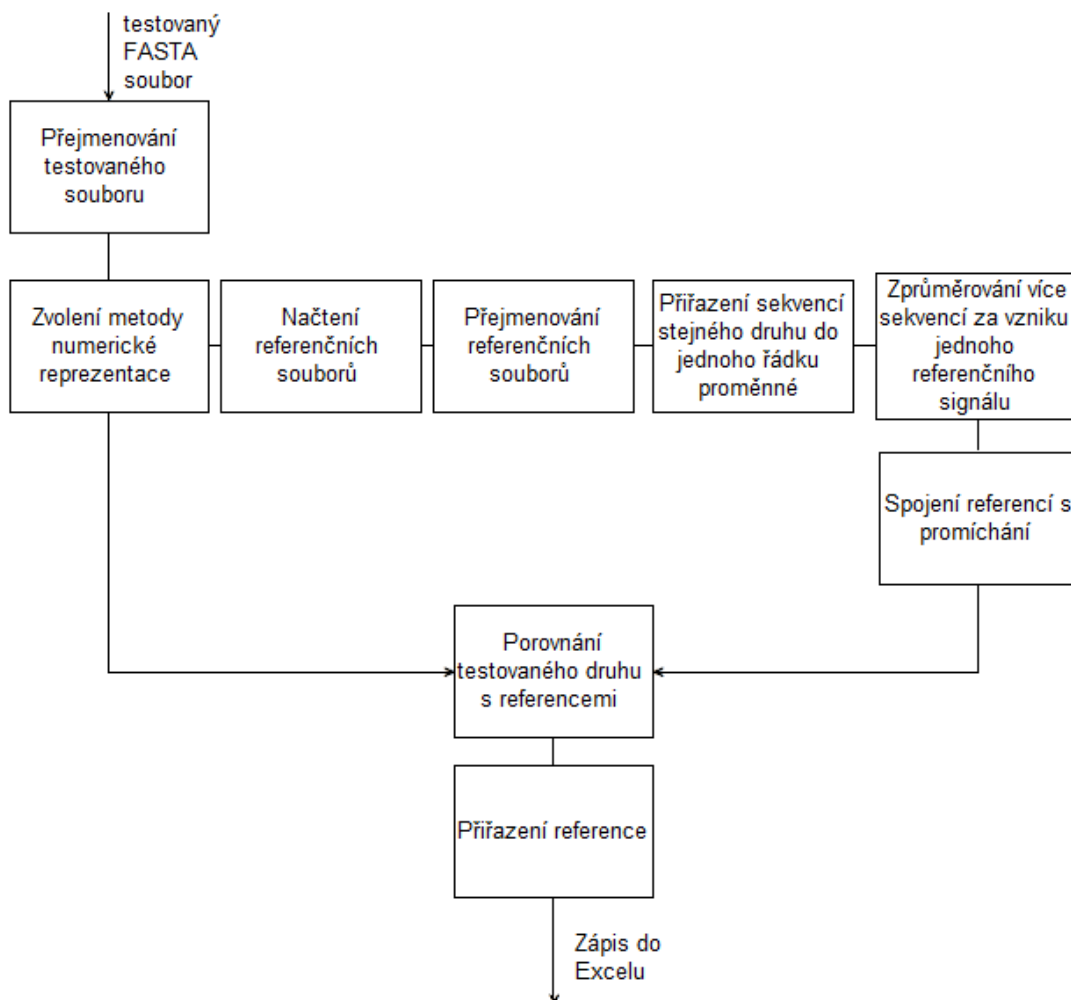
Tabulka 2: Princip zarovnání sekvencí

Sek 1	ACACCC	Numerická reprezentace (převod do EIIP hodnot)	sek 1 (0,126; 0,134; 0,126; 0,134; 0,134; 0,134)	
Sek 2	ACCC		sek2 (0,126; 0,134; 0,134; 0,134)	
Cyklus zarovnání:			Srovnání numerických reprezentací sekvencí (počítáme ty úseky, které se překrývají):	
1)	ACACCC ACCC - -	Euklidovskou vzdáleností:	$D_E = 0,008$	
		Rozdílem:	$D = 0,008$	
2)	ACACCC -ACCC -	Euklidovskou vzdáleností:	$D_E = 0,011$	
		Rozdílem:	$D = 0,016$	
3)	ACACCC ---ACCC	Euklidovskou vzdáleností:	$D_E = 0$	
		Rozdílem:	$D = 0$	
Pro zarovnání vybíráme minimum, tedy v tomto případě $D/D_E=0$:				
ACACCC - - ACCC				

Stejným způsobem, pomocí euklidovské vzdálenosti nebo rozdílu, je pak počítána i podobnost mezi dvěma vektory – převedené testované a referenční sekvence. Vypočtená numerická hodnota podobnosti se ukládá k příslušné srovnávané referenci. Po srovnání testované sekvence se všemi referenčními sekvencemi je vybrána právě ta reference, která má nejmenší hodnotu podobnosti. Minimální hodnota ukazuje na nejlepší možnost shody mezi danými vektory (sekvencemi) a tím i mezi druhy, případně rody či vyššími taxonomickými jednotkami.

Výše popsaným způsobem jsou porovnány všechny sekvence z testovaného souboru se všemi referencemi. Výsledek je zapsán do tří sloupců v Excelu. První sloupec obsahuje jména testovaných druhů, ve druhém je jméno druhové reference, ke které byla sekvence přiřazena a ve třetím sloupci je hodnota minimální euklidovské vzdálenosti nebo rozdílu dvou sekvencí. Pro lepší představitelnost programu je na

obrázku 13 uvedené zjednodušené blokové schéma pro analýzu organismu.



Obrázek 13: Zjednodušené blokové schéma programu

6.2 Identifikační analýza

Pro identifikační analýzu byly jako referenční druhy vybrány soubory s největším množstvím variabilních sekvencí pro jeden určitý druh. Tyto soubory zdaleka nemohou obsahovat všechny druhy daného rodu natož řádu. Z těchto důvodů je omezena úspěšnost klasifikace pouze na ty testované druhy, které mají svou referenci. Druhy pro referenční sekvence byly získány ze souborů obsahující druhy netopýrů z Guyany (název souboru je BCBNC), chrostíků z Churchillu (CUCAD), neotropických ptáků (MNCN) a subtropických ryb (FCFP). Tyto soubory jsou získány z databáze BOLD. Počet jednotlivých taxonů je shrnut v tabulce 3.

Testované soubory oproti referenčním souborům obsahují více jednotlivých druhů

organismů. V testovaném souboru pro neotropické ptáky (BRAS) nenajdeme skoro žádnou variabilitu, protože jeden druh má pouze jednu sekvenci. Kromě tohoto souboru program testuje soubory netopýrů (BCDR), chrostíků (DSTRI) a dva soubory subtropických ryb (FCFPS a FCFPW). Zastoupení taxonomických jednotek je uvedeno v tabulce 4. Tabulka 5 ukazuje početní zastoupení testovaných souborů majících referenci pro druhy a rody. Procentuální úspěšnost přiřazení správné reference k testovanému druhu je pro jednotlivé soubory dána v následující tabulce 6 a úspěšnost taxonomického určení pro rody testovaných organismů je v tabulce 7.

Tabulka 3: Souhrn počtu druhů, rodů a vyššího taxonu pro jednotlivé referenční soubory

<i>reference</i>	<i>počet</i>		
soubory	druh	rod	podčeleď/nadčeleď/řád
BCBNC	87	50	13
CUCAD	53	25	7
MNCN	43	36	6
FCFP	48	36	10

Tabulka 4: Souhrn počtu druhů, rodů a vyššího taxonu pro jednotlivé testované soubory

<i>testované</i>	<i>počet</i>		
soubory	druh	rod	podčeleď/nadčeleď/řád
BCDR	68	41	13
BRAS	431	270	23
DSTRI	22	13	6
FCFPS	55	47	17
FCFPW	49	42	13

Pozn. Početní zastoupení jednotlivých taxonů bylo získáno pomocí funkce z předmětu – Praktika z bioinformatiky.

Tabulka 5: Početní zastoupení testovaných sekvencí, ke kterým existuje reference

<i>soubory</i>	<i>počet sekvencí s referencí u testovaných souborů</i>	
	druh	rod
DSTRI	276	489
BRAS	85	145
BCDR	4108	4126
FCFPS	137	156
FCFPW	134	157

Tabulka 6: Procentuální úspěšnost taxonomického určení pro druhy

<i>soubory</i>	<i>úspěšnost určení v %</i>					
	Třívektor. reprezentace	1. a 4. kvadrant - rozdíl	1. a 4. kvadrant - e. v.	EIIP	kumul. EIIP - rozdíl	kumul. EIIP - e. v.
DSTRI	100,00	75,36	99,64	71,74	96,74	94,57
BRAS	85,88	96,47	35,29	92,94	56,47	24,71
BCDR	99,42	97,15	97,44	84,06	92,67	92,70
FCFPS	98,54	99,27	97,81	99,27	97,08	97,08
FCFPW	100,00	100,00	100,00	100,00	100,00	100,00

V tabulce *rozdíl* znamená numerický rozdíl mezi dvěma numerickými reprezentacemi, které jsou použité pro nalezení nejbližšího podobného druhu organismu a *e. v.* je euklidovskou vzdáleností, podle vybraného způsobu jsou porovnávány sekvence mezi sebou.

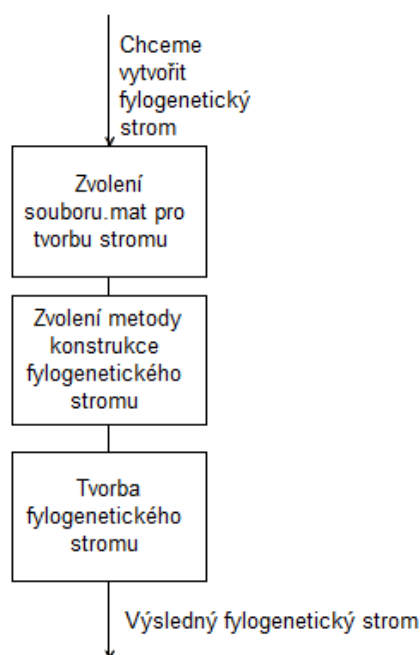
Tabulka 7: Úspěšnost taxonomického zařazení pro rody

<i>soubory</i>	<i>úspěšnost určení v %</i>					
	Třívektor. reprezentace	1. a 4. kvadrant - rozdíl	1. a 4. kvadrant - e. v.	EIIP	kumul. EIIP - rozdíl	kumul. EIIP - e. v.
DSTRI	96,11	99,39	92,23	91,00	57,26	55,62
BRAS	91,03	94,48	44,14	85,52	44,83	26,21
BCDR	99,88	99,90	99,88	99,52	98,59	97,62
FCFPS	98,08	99,36	98,08	98,72	92,95	88,46
FCFPW	100,00	94,90	94,27	94,90	94,27	94,90

6.3 Fylogenetické stromy

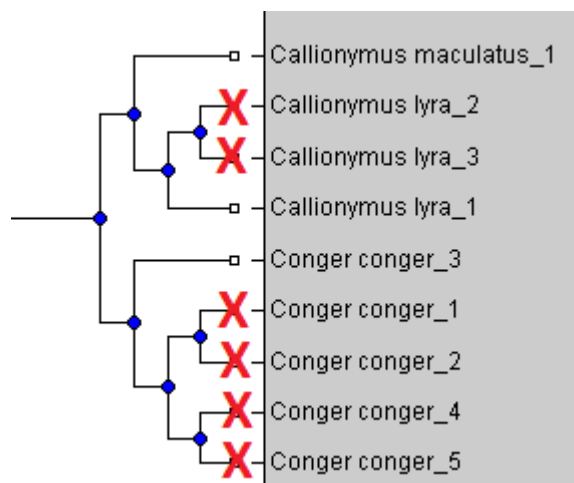
Aby bylo možné porovnávat úspěšnost správného určení druhů v rámci čeledí, byly pro sekvence v souborech netopýrů (BCBNC), chrostíků (CUCAD), neotropických ptáků (MNCN) a subtropických ryb (FCFP) vytvořeny fylogenetické stromy metodami třívektorové reprezentace, 1. a 4. kvadrantu srovnávaného pomocí euklidovské vzdálenosti, EIIP hodnot a kumulovanou metodou EIIP hodnot srovnávanou pomocí rozdílu sekvencí. Princip vytvoření stromu je uveden v zjednodušeném blokovém schématu na obrázku 14. Pro srovnání výsledků se použily hodnoty z vytvořených

stromů pomocí Kimurova dvouparametrického modelu spojovaného metodou spojování sousedů (Neighbor - joining). Takto vytvořené stromy se kondenzovaly na druhy, rody a vyšší taxonomické jednotky, které budou mít největší výpovědní hodnotu (obrázek 16). Kondenzace spočívá v tom, že v případě spojení větví stejných druhů organismů do jednoho uzlu se ostatní smažou tak, aby zůstala právě jedna. Příklad je ukázán na obrázku 15, kde červený křížek je u druhů, které se budou mazat pro kondenzaci stromu. Odpovídající numerickou hodnotu pro porovnávání metod jsme dostali příkazem *get*. Hodnota udává počet listů v kondenzovaném stromu. Listy zastupují reálný počet taxonomických jednotek. Funkce pro výpočet srovnávacích hodnot metodou K2P - NJ byly vytvořeny v předmětu - Praktika z bioinformatiky. Tyto funkce byly využity k očíslování jednotlivých druhů z FASTA souboru, porovnání každé dvojice sekvencí a získání matice pro metodu Kimurova dvouparametrického modelu konstruovaného spojováním sousedů.

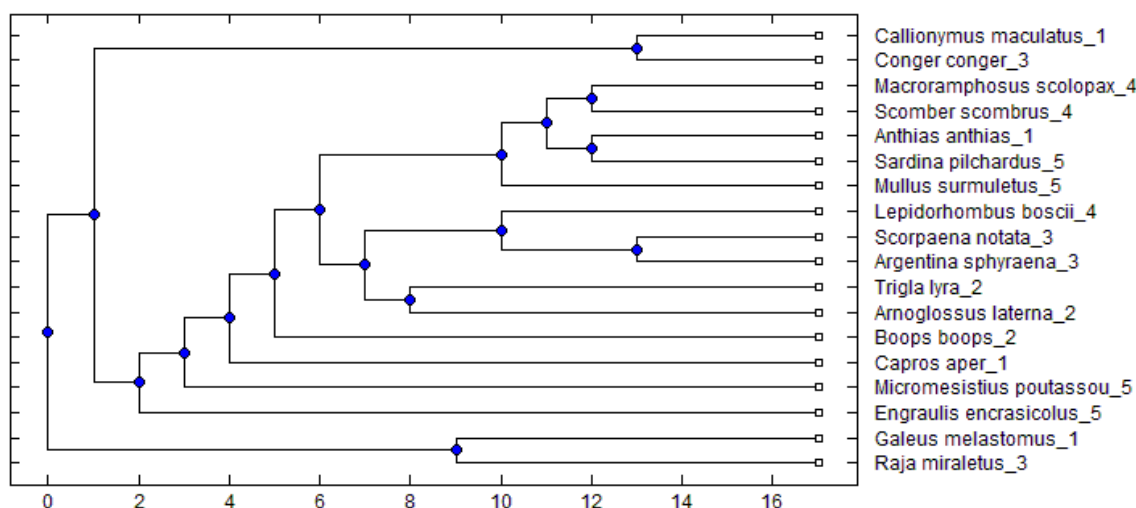


Obrázek 14: Zjednodušené blokové schéma pro tvorbu fylogenetického stromu

Po získání srovnávacích hodnot K2P - NJ byly testovány naprogramované metody numerické reprezentace (1. a 4. kvadrant, EIIP hodnoty, kumulované EIIP hodnoty a třívektorová reprezentace). S tím rozdílem, že hodnotou vzdálenosti pro spojování podobných druhů do jedné větve byla právě vypočtená hodnota minima, což je hodnota nejpodobnějšího úseku sekvencí z testovaného souboru. Na základě těchto vypočtených hodnot, různých podle zvolené metody numerické reprezentace, byl sestaven fylogenetický strom. Tento vytvořený strom byl opět zkondenzován.



Obrázek 15: Princip kondenzace fylogenetického stromu pro druh



Obrázek 16: Kondenzovaný fylogenetický strom souboru subtropických ryb (FCFP) metodou K2P - NJ

V průběhu sestavení stromu jsou porovnávány vždy dvojice sekvencí. Porovnávaná sekvence se přiřadí do uzlu k nejpodobnější sekvenci a vytváří tak další větev. V tabulce 8 a 9, která je rozdělená, jsou uvedeny úspěšnosti sestavení fylogenetických stromů pro jednotlivé soubory. Ideální metoda pro správně vytvořený strom, a tedy i pro správně zařazené sekvence stejného druhu do jednoho uzlu, by měla být rovna jedné. Hodnoty větší než jedna dosahují menší přesnosti. Jako porovnávací metoda, ke které jsou srovnány úspěšnosti naprogramovaných numerických reprezentací, je brána metoda Kimurova dvoupametrického modelu (K2P).

Tabulka 8: Úspěšnost kondenzace fylogenetických stromů pro taxonomické jednotky

<i>soubory</i>	<i>K2P</i>			<i>Třívektorová reprezen.</i>			<i>1.a 4.kvad.-e.v.</i>		
	druh	rod	podčeleď/ nadčeleď/ řád	druh	rod	Podčeleď/ nadčeleď/ řád	druh	rod	podčeleď/ nadčeleď/ řád
BCBNC	1,24	1,38	2,62	1,22	1,28	2,85	1,25	1,28	2,85
CUCAD	1,08	1,28	1,14	1,28	1,64	2,43	1,43	1,60	2,86
MNCN	1,00	1,03	1,17	1,16	1,19	1,83	1,81	1,86	2,83
FCFP	1,02	1,03	1,80	1,02	1,03	1,60	1,02	1,06	1,70

Tabulka 9: Úspěšnost kondenzace fylogenetických stromů pro taxonomické jednotky

<i>stromy</i>	<i>EIIP</i>			<i>EIIP-kumul. rozdíl</i>		
soubory	druh	rod	podčeleď/ nadčeleď/řád	druh	rod	podčeleď/ nadčeleď/řád
BCBNC	1,34	1,58	4,00	1,52	2,18	6,38
CUCAD	1,19	1,52	1,71	1,19	1,68	3,43
MNCN	1,28	1,31	3,17	2,74	3,03	4,33
FCFP	1,17	1,00	2,10	1,02	1,08	3,00

6.4 Analýza a vyhodnocení výsledků

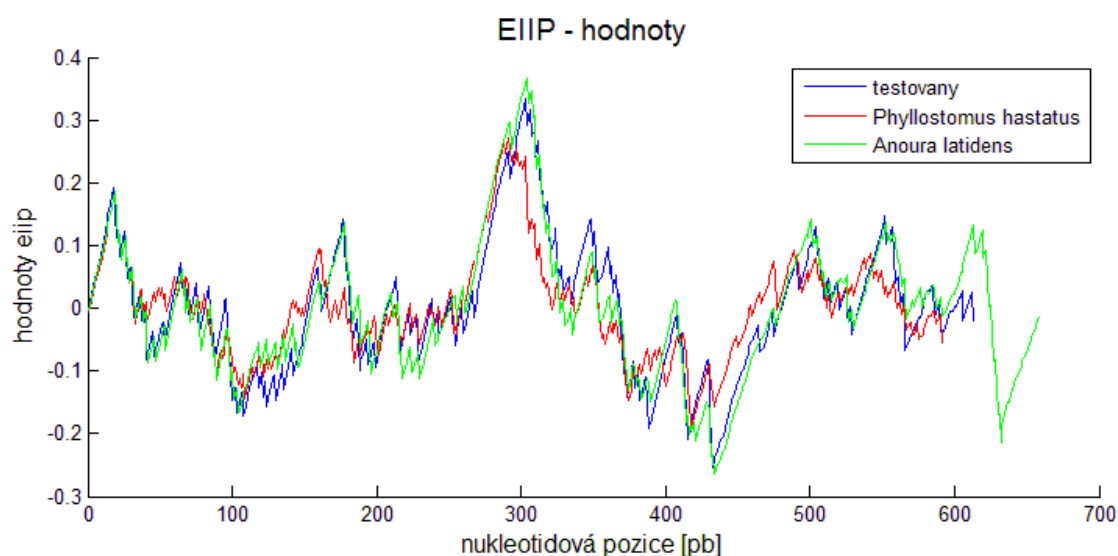
Pro druhovou identifikaci se nejvíce osvědčily metody třívektorové reprezentace a 1. a 4. kvadrantu se srovnáním numerických reprezentací sekvencí pomocí rozdílu. Obě tyto metody dosáhly průměrné úspěšné klasifikace větší než 93 %, jak je vidět v tabulce 10.

Pro identifikaci do rodů se osvědčily metody EIIP hodnot, třívektorové reprezentace a 1. a 4. kvadrantu se srovnáním sekvencí pomocí minima. Tyto metody rovněž přesahují 93 % úspěšnost ve správném taxonomickém určení.

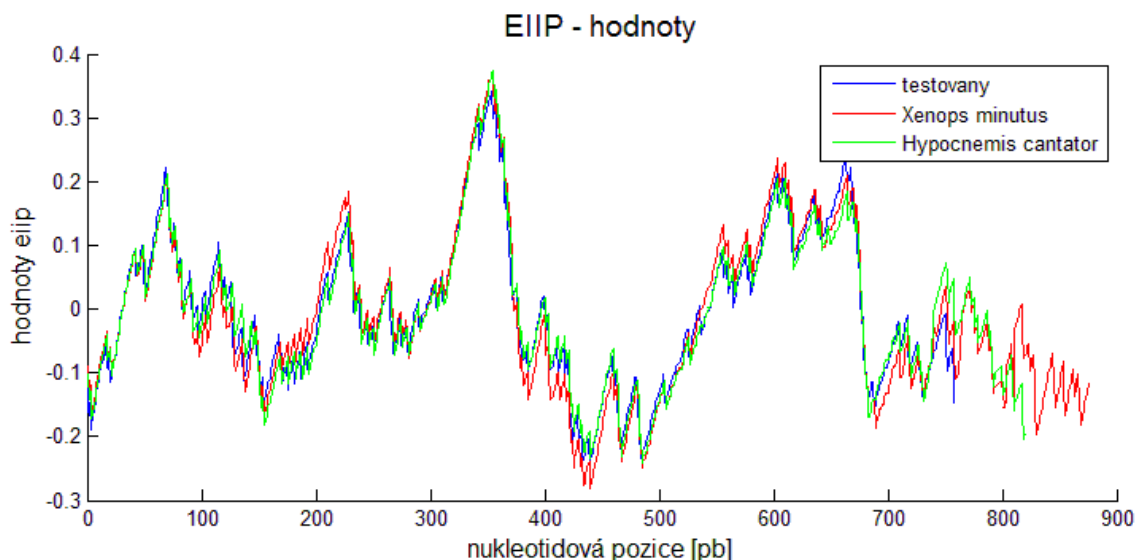
Tabulka 10: Průměrná úspěšnost správného přiřazení testovaného souboru pro druhy a rody

	<i>Třívektor. Reprezentace</i>	<i>1. a 4. kvad. - rozdíl</i>	<i>1. a 4. kvadrant -e.v</i>	<i>EIIP</i>	<i>kumul. EIIP - rozdíl</i>	<i>kumul. EIIP - e.v.</i>
druh	96,77	93,65	86,04	89,60	88,59	81,81
rod	97,02	97,61	85,72	93,93	77,58	72,56

Nejmenší úspěšnost byla dosažena u metody kumulovaných EIIP hodnot, kde jsou sekvence srovnávány pomocí euklidovské vzdálenosti. Mezi problémovými testovanými druhy byly i organismy *Phyllostomus hastus* ze souboru BCDR a *Xenops minutus* ze souboru BRAS, u kterého vychází nejmenší úspěšnost. *Phyllostomus hastus* měl v testovaném souboru několik sekvencí a k většině byl přiřazen jako referenční druh *Anoura latidens*. Tato reference byla k testované sekvenci bližší než správný referenční druh *Phyllostomus hastus*. Podobnost sekvencí je vidět na obrázku 17. V referenčním souboru je druh *Phyllostomus hastus* zastoupen pouze dvěma sekvencemi, z kterých je počítána reference. Oproti tomu druh *Xenops minutus* obsahuje v referenčním souboru 23 sekvencí, ze kterých je počítána reference. Přesto, že byla testovanému druhu přiřazena nesprávná reference, je vidět, že se sekvence od sebe tolik neliší. Testovaný druh byl v testovaném souboru obsažen jedenkrát a byl přiřazen k referenci *Hypocnemis cantor*, jak vidíme na obrázku 18.



Obrázek 17: Podobnost testovaného druhu *Phyllostomus hastus* k referenci *Anoura latidens* namísto k referenci *Phyllostomus hastus*



Obrázek 18: Podobnost testovaného druhu *Xenops minutus* k referenci *Hypocnemis cantator* namísto k referenci *Xenops minutus*

Z těchto obrázků můžeme usoudit, že záleží na variabilitě jednotlivých druhů v rámci referenčních souborů. Lépe budou přiřazeny druhy, které v referenčních souborech mají více sekvencí, než ty druhy, které tam mají pouze jednu.

Z tabulek úspěšnosti určení také vyplývá, že metody 1. a 4. kvadrantu a kumulované EIIP hodnoty srovnávané pomocí minima jsou úspěšnější než tytéž metody srovnávané pomocí euklidovské vzdálenosti. Princip metod je vysvětlen na následujícím příkladu. Pokud se vezme vektor 1 s čísly (1, 4, 8) a vektor 2 (2, 6, 10) a srovná se pomocí minima (sečtou se rozdíly v absolutní hodnotě):

$$|(1-2)| + |(4-6)| + |(8-10)| = 5$$

Výsledkem je hodnota 5. Pokud však vektory srovnáme pomocí eukleidovské vzdálenosti, naprogramované v programu (rozdíly jsou umocněny na druhou, sečtou se a pak odmocní):

$$\sqrt{(1-2)^2 + (4-6)^2 + (8-10)^2} = 3$$

Výsledkem je mnohem menší číslo než u srovnání pomocí minima. Protože referenční soubory neobsahují velký počet sekvencí pro jednotlivé druhy, jsou v tomto případě výhodnější větší rozdíly, tedy srovnání pomocí minima. Je možné, že u souborů obsahujících větší počet sekvencí pro jeden druh a zároveň i větší počet druhů pro určitý

rod, by bylo srovnání pomocí euklidovské vzdálenosti vhodnější. Srovnávaly by se menší rozdíly. Tato hypotéza by se musela otestovat.

Pokud ovšem je brána k porovnání sekvencí kombinace euklidovské vzdálenosti a rozdílu v programu, je dosaženo zlaté střední cesty.

Euklidovské vzdálenost je užita pro zarovnání sekvencí tak, aby se překrývaly nejvíce shodné úseky. Tedy ty úseky, které mají oblasti shodných nukleotidů. Zarovnání euklidovskou vzdáleností je konkrétně použito při tvorbě referenčního signálu a při zarovnání testovaného signálu s referenčními.

Naopak rozdíl je použit k samotnému porovnání testované sekvence s referencemi a následnému výběru minimální hodnoty. Průměrná úspěšnost správného přiřazení se ve většině případů shoduje s největší dosaženou úspěšností pro metody numerické reprezentace EIIP kumulované hodnoty a 1. a 4. kvadrantu. Srovnání uvádí tabulka 11, kde KOM znamená kombinace.

Tabulka 11: Úspěšnost taxonomického přiřazení pro kombinaci srovnání pomocí minima a euklidovské vzdálenosti

soubory	<i>úspěšnost určení v %</i>			
	druhy		rody	
	eiip-kumul-KOM	1a4-KOM	eiip-kumul-KOM	1a4-KOM
DSTRI	96,74	75,36	57,26	99,39
BRAS	56,47	96,47	44,83	94,48
BCDR	92,67	97,25	98,59	99,85
FCFPS	97,08	99,27	92,95	99,36
FCFPW	100,00	100,00	94,27	94,90
průměr	88,59	93,67	77,58	97,59

Při tvorbě fylogenetických stromů se nejvíce přiblížily hodnoty vybraných numerických reprezentací k hodnotám metody K2P - NJ u metod třívektorové reprezentace 1. a 4. kvadrantu. Pro vyšší taxonomickou jednotku než rod, dosáhla metoda třívektorové reprezentace lepšího výsledku než metoda Kimurova dvouparametrického modelu, a to pro soubor subtropických ryb (FCFP). Z průměrného hlediska se numerické reprezentace přibližují úspěšnosti Kimurovu dvouparametrickému modelu tvořeného spojováním sousedů (K2P-NJ), ale zdaleka se nestávají výhodnějším obzvláště pro taxonomické určení čeledí nebo řádů. Metoda Kimurova dvouparametrického modelu bohužel také nedosahuje ideální hodnoty a tou je jedna. Průměrné hodnoty kondenzovaných stromů ukazuje tabulka 12.

Tabulka 12: Průměrné hodnoty testovaných souborů pro fylogenetické stromy

	<i>druh</i>	<i>rod</i>	<i>vyšší taxon</i>
K2P-NJ	1,08	1,18	1,68
Třívektorová reprezent.	1,17	1,29	2,18
1. a 4. kvadrant - min	1,24	1,29	2,17
1. a 4. kvadrant - e.v.	1,38	1,45	2,56
EIIP	1,24	1,35	2,75

Je ovšem důležité zdůraznit, že na správnost úspěšného přiřazení na základě metod numerických reprezentací mají velký vliv soubory, které budou považovány za referenční. U těchto souborů záleží nejen na počtu sekvencí pro jeden druh, ale i na jejich variabilitě, která je udána rychlostí evolučních změn, a na počtu druhů, které se podaří získat pro daný rod (případně počtu rodů pro čeleď, atd.). Rychlost evolučních změn není u většiny organismů rovnoměrná. Bylo prokázáno, že organismy mají období rychlejšího vývoje anebo pomalejšího vývoje. Tyto období jsou nahodilé, ale mohou se prokazovat zpětně.[5, 11]

Metody numerických reprezentací by se nejlépe ověřovaly na souborech organismů, které budou mít nízkou vnitrodruhovou variabilitu a vysokou mezidruhovou variabilitu. Bohužel tohoto ideálu k testování nemůžeme dosáhnout, protože nové druhy se vyvíjejí postupně. Z hlediska evoluční teorie jsou si nově vznikající druhy mnohem podobnější s druhy původními než již nově vzniklé druhy. Dá se předpokládat, že většina jedinců má tím vyšší mezidruhovou variabilitu, čím větší je jejich evoluční vzdálenost a proto úspěšnost pro každý vybraný soubor organismů je odlišná.

7 ZÁVĚR

Tato bakalářská práce se zabývá vědním oborem taxonomie. Taxonomie je dále představena ve dvou metodách, které se používají k identifikaci druhů, a to v klasické a v molekulární taxonomii. Obě tyto metody se výrazně liší, neboť každá stojí na jiných principech. Dále práce naznačuje, že obě tyto metody identifikace druhů by měly být brány v úvahu, když katalogizujeme neznámý druh.

Dále byly popsány tři vybrané metody numerické reprezentace genomických dat, které byly použity k analýze druhů. Převod symbolické reprezentace DNA sekvencí do numerických reprezentací, metodami 1. a 4. kvadrantu kartézského souřadného systému, EIIP hodnotami a třívektorovou reprezentací, byl naprogramován pomocí samostatných funkcí v programovém prostředí Matlab. U metod 1. a 4. kvadrantu a EIIP hodnot byla navíc použita kumulovaná metoda. Ta spočívá v tom, že vypočítaná hodnota je přičtena k hodnotě předešlé. U všech těchto metod nedochází ke ztrátě informace.

Funkce vybraných metod byly otestovány na souborech sekvencí získaných z databáze BOLD. Při analýze byla sledována úspěšnost jednotlivých metod v zařazení testovaných sekvencí ke správnému druhu, rodu či čeledi. Za úspěšné metody pro dané testované soubory lze považovat metody numerických reprezentací pro 1. a 4. kvadrant a třívektorovou reprezentaci. Pro druhové určení dosáhla metoda třívektorové reprezentace úspěšnosti 96,77 % a pro rodové taxonomické správné určení metodou 1. a 4. kvadrantu byla tato úspěšnost 97,61 %. Průměrově vybrané metody numerické reprezentace pro taxonomické určení vyšších jednotek (např. řády a čeledě) nebyly nejvhodnější. Naopak taxonomické zařazení bylo přesnější pomocí běžně v bioinformatice používaného Kimurova dvouparametrického modelu spojovaného sousedů (K2P-NJ) u tvorby fylogenetických stromů. Kimurův dvouparametrický model dosáhl v zařazení vyšších taxonomických jednotek lepších výsledků, a to průměrně 1,68 na kondenzovaných stromech, než hodnotově nejbližší metoda třívektorové reprezentace 2,18 či 2,17 u metody 1. a 4. kvadrantu. Tímto byly splněny všechny body zadání.

K vyhodnocení musíme zdůraznit, že testování bylo prováděno na omezeném počtu souborů, jak referenčních, tak testovaných. Na správnost úspěšného přiřazení na základě metod numerických reprezentací mají velký vliv soubory, které budou považovány za referenční. U těchto souborů záleží nejen na počtu sekvencí pro jeden druh, ale i na jejich variabilitě pro tento druh a na počtu druhů, které se podaří získat pro daný rod

(případně počtu rodů pro čeleď, atd.).

Zároveň můžeme předpokládat, že numerické reprezentace budou lépe přiřazovat sekvence, u kterých je menší vnitrodruhová variabilita než variabilita mezidruhová. Mutační rychlost organismů je rozdílná. Lze předpokládat, že organismy vývojově mladší se budou lišit od vývojově starších, právě díky změnám v sekvencích získaných během vývoje. Proto, abychom mohli zvýšit úspěšnost numerických reprezentací v taxonomickém určení u vývojově mladších organismů, je třeba vytvořit databázi jednotlivých referenčních druhů. Tyto referenční signály pro jednotlivé druhy musí být vytvořeny z dostatečného množství sekvencí. Čím více referenčních druhů bude daná databáze obsahovat, tím přesnějšího určení můžeme dosáhnout.

Při klasickém taxonomickém určování nového druhu organismu by bylo vhodné zapojit i metody molekulární taxonomie (např. i numerické reprezentace), protože může být docíleno spolehlivějších a reprezentativnějších výsledků. Druhy, co jsou určováním klasické taxonomie považovány za nové, mohou být z hlediska molekulární taxonomie už dávno známé a pouze na příslušném geografickém území nečekané. Avšak může tomu být i naopak. Se spojeným přístupem klasické a molekulární taxonomie bychom mohli objevit mnohem rychleji nové druhy organismů a nepřijít tak o jejich existenci.

LITERATURA

- [1] A. S. Nair, S. P. Sreenadhan, "A Coding Measure Scheme Employing Electron-Ion Interaction Pseudopotential (EIIP)," *Bioinformation*, vol. 1, no. 6, pp. 197- 202, 2006.
- [2] Animal_mitochondrion_diagram_cs. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2014-04-27]. Dostupné z: http://upload.wikimedia.org/wikipedia/commons/3/37/Animal_mitochondrion_diagram_cs.svg
- [3] BERGMANN, T., H. HARDYS, G. BREVES a B. SCHIERWATER. Character-based DNA barcoding: a superior tool for species classification. *BERLINER UND MUNCHENER TIERARZTLICHE WOCHENSCHRIFT*; 122, 11/12; 446-450: *Blackwell Publishing Ltd*. 2009, č. 1, s. 5. DOI: 10.2376/0005-9366-122-446. Dostupné z: <http://www.ncbi.nlm.nih.gov/pubmed/19999380>
- [4] BLAXTER, M. L. The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences* [online]. 2004-04-29, vol. 359, issue 1444, s. 669-679 [cit. 2013-11-12]. DOI: 10.1098/rstb.2003.1447. Dostupné z: <http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2003.1447>
- [5] CASIRAGHI, M., M. LABRA, E. FERRI, A. GALIMBERTI a F. DE MATTIA. DNA barcoding: a six-question tour to improve users' awareness about the method. *Briefings in Bioinformatics* [online]. 2010-07-17, vol. 11, issue 4, s. 440-453 [cit. 2013-11-12]. DOI: 10.1093/bib/bbq003. Dostupné z: <http://bib.oxfordjournals.org/cgi/doi/10.1093/bib/bbq003>
- [6] COSIC, I. Macromolecular bioactivity: is it resonant interaction between macromolecules?-theory and applications. *IEEE Transactions on Biomedical Engineering* [online]. vol. 41, issue 12, s. 1101-1114 [cit. 2013-11-16]. DOI: 10.1109/10.335859. Dostupné z: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=335859>
- [7] CRISTEA, P.D. Getting DNA to numbers: Conversion of nucleotides sequences into genomic signals. *J.Cell.Mol.Med.*[online]. 2002, roč. 6, č. 2, s. 25 [cit. 2013-11-16]. Dostupné z: <http://www.ncbi.nlm.nih.gov/pubmed/12169214>
- [8] DESALLE, R., M. G. EGAN a M. SIDDALL. The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal*

- Society B: Biological Sciences* [online]. 2005-10-29, vol. 360, issue 1462, s. 1905-1916 [cit. 2013-11-12]. DOI: 10.1098/rstb.2005.1722. Dostupné z: <http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2005.1722>
- [9] DROZD, Pavel. *Principy systematiky a taxonomie*. Vyd. 1. Ostrava: Ostravská univerzita, 2004, 89 s. ISBN 80-704-2995-X.
- [10] FLEGR, Jaroslav. *Úvod do evoluční biologie*. Vyd. 1. Praha: Academia, 2007, 544 s. Galileo, sv. 13. ISBN 978-802-0015-396.
- [11] HAMPL, Vladimír. Molekulární taxonomie. *Přednášky z molekulární taxonomie* [online]. 2012, č. 1 [cit. 2014-05-16]. Dostupné z: <http://web.natur.cuni.cz/~vlada/moltax/>
- [12] Heat Shock Proteins & Chaperones. In: *QIAGEN* [online]. 2013 [cit. 2014-04-27]. Dostupné z: http://www.qiagen.com/products/genes%20and%20pathways/complete%20biology%20list/heat%20shock%20proteins%20and%20chaperones/~media/NextQ/Image%20Library/ILLU/03/02/ILLU_0302_GG_Heat_Shock_Proteins/3_10.ashx?la=en
- [13] HEBERT, P. D. N., A. CYWINSKA, S. L. BALL a J. R. DEWAARD. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences* [online]. 2003-02-07, vol. 270, issue 1512, s. 313-321 [cit. 2013-11-12]. DOI: 10.1098/rspb.2002.2218. Dostupné z: <http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2002.2218>
- [14] CHAO, Kun-Mao a Louxin ZHANG. *Sequence comparison: theory and methods*. London: Springer, c2009, xx, 209 p. ISBN 18-480-0320-X.
- [15] LIAO, Bo a Tian-Ming WANG. New 2D graphical representation of DNA sequences. *Journal of Computational Chemistry* [online]. 2004, vol. 25, issue 11, s. 1364-1368 [cit. 2013-11-16]. DOI: 10.1002/jcc.20060. Dostupné z: <http://doi.wiley.com/10.1002/jcc.20060>
- [16] MADĚRÁNKOVÁ, D., PROVAZNÍK, I. Motive representation in nucleotide densities of bird's mitochondrial gene COX1. *ACM Digital Library: Proceedings of 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL '11)*, Barcelona (Španělsko), 2011
- [17] MEYERS, Ed. by Robert A. *Encyclopedia of molecular cell biology and molecular medicine*. 2. ed. Weinheim: Wiley-VCH-Verl, 2005. ISBN 35-273-0550-5.
- [18] PAZDERA, Josef. Biomarkery – nový pojem v posuzování stavu naší

- kůže. *Biomarkery – nový pojem v posuzování stavu naší kůže* [online]. 2006, č. 1, 2006-3-3 [cit. 2013-11-12]. Dostupné z: <http://www.osel.cz/index.php?clanek=1739>
- [19] Podle nového odhadu žije na zemi 8,7 milionu druhů. Ale... *Rozmanitost života: Podle nového odhadu žije na zemi 8,7 milionu druhů. Ale...* [online]. 2011, č. 1, s. 1, 2011-8-24 [cit. 2013-11-12]. Dostupné z: http://www.tyden.cz/rubriky/veda/priroda/podle-noveho-odhadu-zije-na-zemi-8-7-milionu-druhu-ale_210324.html#.UoJIW_kmbW-
- [20] Point mutation. In: *ROSALIND* [online]. 2012 [cit. 2014-04-27]. Dostupné z: http://rosalind.info/media/problems/hamm/point_mutation.png
- [21] PRAY, L. DNA replication and causes of mutation. *Nature Education* [online]. 2008, č. 1 [cit. 2013-12-06]. Dostupné z: <http://www.nature.com/scitable/topicpage/dna-replication-and-causes-of-mutation-409>
- [22] RANDIĆ, Milan, Marjan VRAČKO, Nella LERŠ a Dejan PLAVŠIĆ. Novel 2-D graphical representation of DNA sequences and their numerical characterization. *Chemical physics letters* [online]. 2003, č. 368, s. 1-6, 2013 [cit. 2013-11-16]. Dostupné z: <http://www.sciencedirect.com/science/article/pii/S0009261402017840>
- [23] RUBINOFF, DANIEL. *Essays: Utility of Mitochondrial DNA Barcodes in Species Conservation*. 2. ed. Weinheim: Wiley-VCH-Verl, 2005. ISBN 10.1111/j.1523-1739.2006.00372.x. Dostupné z: <http://doi.wiley.com/10.1111/j.1523-1739.2006.00372.x>
- [24] SCHEFFLER, Immo E. *Mitochondria*. 2nd ed. Hoboken, N.J.: Wiley-Liss, c2008, xviii, 462 p., [12] p. of plates. ISBN 04-700-4073-4.
- [25] SCHWARTZ, Jeffrey H. a Bruno MARESCA. Do Molecular Clocks Run at All? A Critique of Molecular Systematics. *Biological Theory* [online]. 2006, vol. 1, issue 4, s. 357-371 [cit. 2013-11-16]. DOI: 10.1162/biot.2006.1.4.357. Dostupné z: <http://www.mitpressjournals.org/doi/abs/10.1162/biot.2006.1.4.357>
- [26] TAYLOR, H. R. a W. E. HARRIS. An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources* [online]. 2012, vol. 12, issue 3, s. 377-388 [cit. 2013-11-16]. DOI: 10.1111/j.1755-0998.2012.03119.x. Dostupné z: <http://doi.wiley.com/10.1111/j.1755-0998.2012.03119.x>
- [27] YAU, S. S. -T. DNA sequence representation without degeneracy. *Nucleic Acids*

Research [online]. 2003-06-15, vol. 31, issue 12, s. 3078-3080 [cit. 2013-11-16].

DOI: 10.1093/nar/gkg432. Dostupné z:

<http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkg432>

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

BOLD	Barcode of Life Database, databáze pro čárové kódy
CAOS	algoritmus systému charakteristických znaků
D	rozdílová vzdálenost
D_E	Euklidovská vzdálenost
DNA	deoxyribonukleová kyselina
EIIP	Elektron – iontová interakce potenciálu
HSP	heat shock protein, protein teplotního šoku
K2P	Kimurův dvouparametrický model
K2P - NJ	Kimurův dvouparametrický model s metodou Neighbor - joining
kb	kilobaze
mtDNA	mitochondriální DNA
NJ	Neighbor - joining (překlad spojování sousedů)
pb	pár baze
RGB spektrum	red green blue spectrum, červeno – zeleno-modré spektrum
RNA	ribonukleová kyselina
tRNA	transferová RNA