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

IDENTIFIKACE ORGANISMŮ POMOCÍ ANALÝZY NUKLEOTIDOVÝCH DENZITNÍCH VEKTORŮ

IDENTIFICATION OF ORGANISMS BASED ON ANALYSIS OF NUCLEOTIDE DENSITY VECTORS

DIZERTAČNÍ PRÁCE
DOCTORAL THESIS

AUTOR PRÁCE
AUTHOR
Ing. DENISA MADĚRÁNKOVÁ

VEDOUCÍ PRÁCE
SUPERVISOR
prof. Ing. IVO PROVAZNÍK, Ph.D.

BRNO 2015

ABSTRAKT

Většina metod pro analýzu genomických dat pracuje se sekvencemi v jejich symbolickém zápisu. Genomické sekvence lze považovat za formu biologického digitálního signálu, který je možné analyzovat metodami zpracování digitálních signálů. Sekvence v symbolickém zápisu však musí být před zpracováním převedeny do vhodného numerického formátu. Tato dizertační práce představuje metodu numerické reprezentace genomických dat, nukleotidové denzitní vektory, a jejich využití k molekulární identifikaci organismů. V současnosti populární DNA barcoding je přístup molekulární identifikace organismů na základě porovnávání krátkých sekvencí určitého úseku mitochondriálního genomu. V této dizertační práci navržená metoda identifikace organismů na základě porovnávání nukleotidových denzitních vektorů byla testována na rozsáhlém souboru DNA barcodingových sekvencích. Navržený způsob identifikace do druhů byl dále rozšířen a otestován na vyšší taxonomické skupiny jako je čeleď. Dále také byly pro testovací data zkonstruovány a porovnány dendrogramy ze standardně používaných evolučních vzdáleností a vzdáleností mezi nukleotidovými denzitními vektory.

KLÍČOVÁ SLOVA

genomika, numerické reprezentace, nukleotidové denzitní vektory, DNA barcoding, molekulární taxonomie, identifikace druhů

ABSTRACT

Most methods for analysis of genomic data work with symbolic sequences. Numerically represented genomic sequences can be analyzed by signal processing methods. A new method of numerical representation of DNA sequences, nucleotide density vectors, is proposed in this thesis. Usability of this method for purposes of molecular species identification is tested on DNA barcoding sequences. DNA barcoding is modern and popular methodology based on comparison of short mitochondrial DNA sequences. Beside species identification by proposed method based on nucleotide density vectors, higher taxa rank identification (e.g. families) was also tested. Furthermore, dendrograms were constructed from standardly used evolutionary distances and distances between nucleotide density vectors and the dendrograms were compared.

KEYWORDS

genomics, numerical representations, nucleotide density vectors, DNA barcoding, molecular taxonomy, species identification

MADĚRÁNKOVÁ, D. *Identifikace organismů pomocí analýzy nukleotidových denzitních vektorů*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2015. 159 s. Vedoucí dizertační práce prof. Ing. Ivo Provazník, Ph.D.

PROHLÁŠENÍ

Prohlašuji, že svou dizertační práci na téma „*Identifikace organismů pomocí analýzy nukleotidových denzitních vektorů*“ jsem vypracovala samostatně pod vedením vedoucího dizertační práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny uvedeny v seznamu literatury na konci práce.

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

V Brně dne …… ……

Ing. Denisa Maděránková

PODĚKOVÁNÍ

Děkuji svému školiteli prof. Ing. Ivo Provazníkovi, Ph.D. za jeho odborné vedení v průběhu mého doktorského studia, čas věnovaný konzultacím a cenné připomínky k mé dizertační práci. Dále bych chtěla poděkovat rodině a přátelům za podporu.

OBSAH

ÚVOD

Bioinformatika je v současné době velmi rychle se rozvíjející vědní obor, který se zabývá metodami sběru, analýzy a vizualizace rozsáhlých souborů biologických dat, především genomických dat (sekvence DNA) a proteomických dat (sekvence aminokyselin). Tento vědní obor se prolíná s dalšími příbuznými obory jako je molekulární biologie, výpočetní biologie, systematika i biomedicínské inženýrství.

Velká část metod pro analýzu genomických a proteomických dat pracuje se sekvencemi v jejich symbolickém zápisu, např. pro DNA sekvence jde o sousledný zápis pomocí písmen A, C, G a T, které reprezentují jednotlivé nukleotidy adenin, cytosin, guanin a tymin. V pořadí písmen je zakódovaná genetická informace. Analýzou symbolických sekvencí můžeme mimo jiné provádět např.: anotaci genů, vyhledávat homologní a podobné sekvence, zarovnávat sekvence či provádět fylogenetickou analýzu.

Genomické a proteomické sekvence však můžeme považovat za formu biologického digitálního signálu, a tak se zde nachází i možnost aplikace metod zpracování digitálních signálů. Sekvence v symbolickém zápisu však musí být před zpracováním převedeny do vhodného numerického formátu. Výběr metody numerické reprezentace silně závisí na typu metody následné analýzy. Pro některé základní analýzy postačují jednorozměrné numerické reprezentace, ale některé sofistikovanější metody vyžadují numerickou reprezentaci, která vhodně reprezentuje biologický informační obsah sekvence.

Tato dizertační práce představuje metodu numerické reprezentace genomických dat, nukleotidové denzitní vektory, která průměruje zastoupení jednotlivých nukleotidů ve zvolené délce posuvného výpočetního okna. Výsledkem numerického mapování symbolické DNA sekvence jsou čtyři samostatné číselné vektory, jeden pro každý typ nukleotidu. Jde o poměrně jednoduchou numerickou reprezentaci nukleotidových sekvencí, která uchovává informaci o biochemických vlastnostech sekvencí. Tato metoda nebyla, pokud je mi známo, nikým jiným publikována. Tato numerická reprezentace umožňuje různé druhy analýz. Předběžně byla vyzkoušena např. na vyhledávání CpG ostrůvků a vyhledávání tandemových repetic. Tato práce je však zaměřena na využití nukleotidových denzitních vektorů k molekulární identifikaci organismů.

V současnosti je velmi populární DNA barcoding, což je přístup molekulární identifikace organismů na základě porovnávání krátkých sekvencí určitého úseku mitochondriálního genomu. Mitochondriální genom má oproti jadernému genomu tu výhodu, mimo jiné, že v něm během stejného časového úseku dojde k několikanásobně většímu počtu bodových mutací, na základě kterých by bylo možné odlišení i blízce příbuzných organismů, tj. organismů, k jejichž diferenciaci došlo v nedávné době. K identifikaci metazoí (vícebuněční živočichové) byl jako úsek pro DNA barcoding

zvolena část sekvence pro gen cytochrom *c* oxidázu podjednotku 1 (*coxI* nebo *coI*) o přibližné délce 650 nukleotidů. Využití tohoto úseku bylo prvně publikováno autory Paulem Hebertem *et al.* v roce 2003 na datech ze skupiny hmyzu. Autory navržený přístup byl natolik jednoduchý a efektivní, že během deseti let nabyla metoda nebývalé popularity spojené s vytvořením obrovské databáze dat. Nutno ovšem dodat, že metoda má i množství kritiků a především se ukázalo, že úsek *coxI* není tak „univerzální a efektivní" u všech skupin organismů, jak bylo předpokládáno.

V této dizertační práci navržená metoda identifikace organismů na základě porovnávání nukleotidových denzitních vektorů byla testována na rozsáhlém souboru DNA barcodingových sekvencích a představuje alternativní metodu ke standardně využívaným metodám jako je vyhledávání homologních sekvencí algoritmem BLAST.

Jelikož nejsou data použitá k testování metody sama o sobě stoprocentně druhově specifická, nemůže být ani žádný způsob identifikace bezchybný. Využití nukleotidových denzitních vektorů nabízí možnost vytvoření jednoho referenčního zástupce pro daný druh z množství jednotlivých jedinců druhů, který bude reflektovat vnitrodruhovou variabilitu druhu. Navrženy a testovány byly tři různé přístupy k výpočtu referenčních nukleotidových denzit. Samotná identifikace „neznámé" sekvence pak probíhá tak, že se nukleotidové denzitní vektory analyzované sekvence porovnávají s referenčními vektory. U sekvencí, které byly navrženou metodou chybně identifikovány, bylo provedeno kontrolní vyhledávání homologních sekvencí algoritmem BLAST v databázi GenBank, což je standardně používaný přístup k identifikaci.

Navržený způsob identifikace do druhů byl dále rozšířen a otestován na vyšší taxonomické skupiny jako jsou čeleď a řád. Dále také byly pro testovací data zkonstruovány a porovnány dendrogramy ze standardně používaných evolučních vzdáleností a vzdáleností mezi nukleotidovými denzitními vektory.

1 GENOMICKÁ DATA

Pod pojmem genomická data můžeme rozumět reprezentaci nukleotidů vyskytujících se v molekulách ribonukleových či deoxyribonukleových kyselin. Tato reprezentace nukleotidů může být symbolická (znaková), numerická, signálová a obrazová. Převážně se však pod pojmem genomická data rozumí symbolická reprezentace, tj. zápis sekvence nukleotidů pomocí sekvence znaků z vybrané abecedy. Genomická data jsou získávána např. sekvenováním, vyfocením elektroforetického gelu, skenováním mikroarray čipu apod. Takto získaná genomická data jsou pak uchovávána, dále zpracovávána a zobrazována s využitím výpočetní techniky.

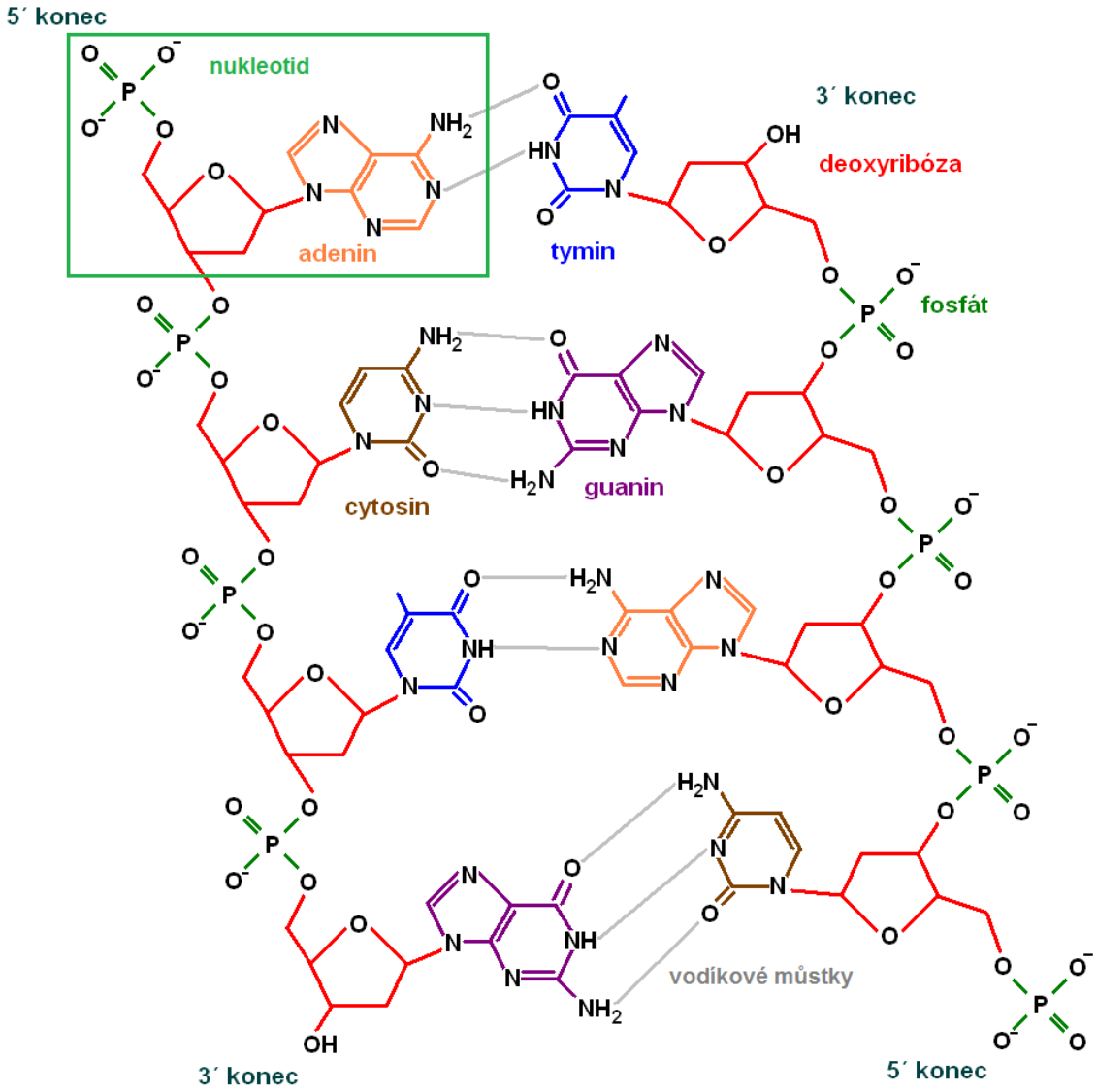
1.1 DNA

Nukleotidové sekvence rozlišujeme ve dvou typech, a to buď jako molekuly ribonukleových kyselin nebo molekuly deoxyribonukleových kyselin. Deoxyribonukleová kyselina (DNA) je velmi dlouhá molekula, která slouží jako nosič genetické informace u všech živých organismů, tj. u prokaryotických a eukaryotických buněk, u většiny virů je pak nosičem jednodušší ribonukleová kyselina (RNA). Jako příklad velikosti (jaderné) DNA lze uvést prasečí cirkovirus typu 1 s velmi krátkou DNA o 1 759 párů bází (jednotka bp) nebo jednobuněčný ameoboid *Polychaos dubium* s velmi dlouhou DNA o 670 Gbp. Délka nosiče genetické informace nekoresponduje s celkovou složitostí organismu.

DNA má strukturu dvojité šroubovice. Základem šroubovice jsou dvě vlákna tvořená střídavě cukrem (deoxyribózou) a fosfátovým zbytkem viz **Obr. 1.1** Struktura dvoušroubovice DNA.. Na těchto vláknech jsou pak navázané báze, které se v DNA vyskytují čtyři základní: adenin, cytosin, guanin a tymin. Vlákna dvoušroubovice drží u sebe vodíkové můstky mezi protějšími bázemi. Komplex báze-cukr se nazývá nukleosid, přidáme-li ke komplexu i fosfát dostaneme komplex zvaný nukleotid. Nukleotidy spojené vodíkovými můstky tvoří pár bází. V symbolické reprezentaci DNA sekvencí se pro nukleotidy volí znaky A, C, G a T pro nukleotidy obsahující bázi adenin, cytosin, guanin a tymin.

Báze se dělí do tří skupin podle jejich biochemických vlastností [1]:

1) podle molekulární struktury - báze adenin (A) a guanin (G) patří mezi puriny (R); báze cytosin (C) a tymin (T) patří mezi pyrimidiny (Y),
2) podle síly vazby mezi komplementárními vlákny – adenin a tymin tvoří vazbu ze dvou vodíkových můstků, jde o vazbu slabou (W); cytosin a guanin tvoří vazbu ze tří vodíkových můstků, jde o vazbu silnou (S),
3) podle obsahu radikálů – báze adenin a cytosin obsahují amino skupinu NH_3 (M) na atomu uhlíku C^6 u adeninu a atomu uhlíku C^4 u cytosinu; tymin a guanin obsahují na těchto atomech keto skupinu C=O (K).



Obr. 1.1 Struktura dvoušroubovice DNA.

Purinové báze tvoří vazbu (vodíkové můstky) pouze s pyrimidinovými bázemi a naopak, čímž vzniká symetrická dvoušroubovice. Této vlastnosti se říká komplementarita bází, přesněji jde o pravidlo Watsonova-Crickova párování bází [2]. Existují i jiné možnosti párování bází, jde ale o speciální případy vyžadující specifické fyzikálně-chemické podmínky.

Zastoupení jednotlivých bází v dvoušroubovici DNA určuje jednu z důležitých fyzikálně-chemických vlastností DNA a to teplotu potřebnou pro rozštěpení dvoušroubovice. Vlákna DNA drží u sebe slabé vodíkové můstky a jejich počet tedy určuje množství energie potřebné pro jejich rozštěpení. Jak bylo řečeno výše, adenin a tymin se pojí dvěma vodíkovými můstky; cytosin a guanin se pojí třemi vodíkovými můstky. Vyšší zastoupení cytosinu a guaninu tudíž zvyšuje množství potřebné energie k rozštěpení dvoušroubovice. V laboratorních podmínkách lze dvoušroubovici štěpit působením tepla, pro části molekuly s vysokým podílem cytosinu a guaninu je pak potřeba použít vyšší teplotu, na kterou se musí vzorek DNA zahřát.

1.2 RNA

Nukleotidové sekvence ve formě molekul RNA se liší tím, že v RNA se místo deoxyribózy vyskytuje ribóza, místo báze tyminu je jednodušší báze uracil a především RNA netvoří dvojitou šroubovici. Molekula RNA je tudíž oproti DNA podstatně náchylnější k chemickým změnám. U prokaryotických a eukaryotických organismů neslouží RNA jako nosič genetické informace, má ale jiné nezastupitelné a klíčové role.

Molekuly RNA se dělí na tři základní typy [5]:

1) mRNA = mediátorová RNA, přenáší genetickou informaci z DNA do místa translace (přepisu do sekvence aminokyselin),
2) tRNA = transferová RNA, jde o malé molekuly RNA, které přenáší jednotlivé aminokyseliny, účastní se procesu translace,
3) rRNA = ribozomová RNA, jde o různě velké molekuly RNA, které tvoří buněčné struktury zvané ribozomy, což jsou místa, kde probíhá translace.

Kromě těchto základních typů RNA existují ještě další typy se speciálními funkcemi.

Základní struktura RNA je jednovláknová, avšak RNA je pro zvýšení chemické stability schopná sekundární strukturalizace, jejíž forma má souvislost s funkcí molekuly. Nejjednodušší sekundární strukturou je vlásenka, kdy se úsek RNA páruje s komplementárním úsekem pomocí vodíkových můstků či struktura tRNA v podobě jetelového listu, která je esenciální pro přenos aminokyseliny a správnou vazbu tRNA na mRNA při translaci. [6]

1.3 GENETICKÝ KÓD A PŘENOS GENETICKÉ INFORMACE

Genetický kód (GK) je nepřekrývající se kód, který pomocí sledu nukleotidů určuje jednotlivé aminokyseliny, iniciaci a terminaci translace. Tento kód je tvořen sekvencí kodonů (trojicemi nukleotidů) v molekule mediátorové RNA.

Základní vlastnosti genetického kódu jsou [1]:

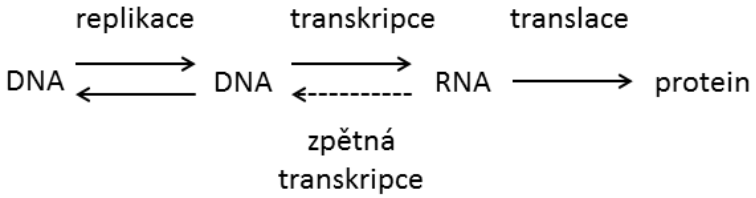
1) GK je složen z tripletů nukleotidů. Každý triplet určuje jednu aminokyselinu.
2) GK se v řetězci nepřekrývá až na výjimky, kdy se geny mohou překrývat a čtou se v různých čtecích rámcích (viz níže).
3) GK neobsahuje interpunkční znaménka, tj. v molekule mRNA se kodony čtou jeden za druhým.
4) GK je degenerovaný, tj. redundantní. Existuje 20 aminokyselin. Pomocí 4 druhů nukleotidů je možné vytvořit $2^4 = 64$ různých kodonů. Až na dvě výjimky jsou všechny ostatní aminokyseliny kódovány více než jedním kodonem.

5) GK je uspořádaný. Kodony pro jednu aminokyselinu či pro aminokyseliny podobných chemických vlastností jsou příbuzné a liší se často pouze jedním nukleotidem.

6) GK obsahuje iniciační (start) a terminační (stop) kodony.

7) GK je téměř univerzální. Až na výjimky je stejný GK u většiny živých organismů. Mírně odlišný GK existuje např. u mitochondrií.

Molekuly DNA nesou genetickou informaci, tj. předpis na výrobu všech peptidů a proteinů organismu. Přenos genetické informace se řídí centrálním dogmatem molekulární biologie, které je znázorněno na **Obr. 1.2** Schéma centrálního dogma molekulární biologie.. Toto dogma říká, že se genetická informace kódovaná v DNA může procesem replikace přenášet do další generace, nebo je genetická informace pomocí procesu transkripce přenesena na molekuly mRNA a procesem translace na konečný produkt, čímž je peptid nebo protein. Transkripce a translace slouží k vyjádření genetické informace do fenotypu organismu. Tento přenos je jednosměrný, tj. z proteinu nelze zpětně sestavit (biologicky ani počítačově) sekvenci DNA. Ze znalosti posloupnosti aminokyselin můžeme sice zrekonstruovat možnou sekvenci DNA, avšak díky degeneraci genetického kódu si nemůžeme být jisti, který kodon byl skutečně použit.



Obr. 1.2 Schéma centrálního dogma molekulární biologie.

V **Tab. 1.1** Význam kodonů v mRNA [1]. jsou uvedeny významy jednotlivých kodonů v mRNA, tj. jaké aminokyseliny jsou danými kodony kódovány. Kromě standardně kódovaných 20 aminokyselin jsou genetické kódy organismů doplněny i o další 2 aminokyseliny selenocystein a pyrolyzin, které jsou specifickou cestou kódované stop kodony UGA a UAG. K tomuto speciálnímu kódování dochází u přepisu genetické informace pro tvorbu selenoproteinů a proteinů metabolizmu produkující metan a je řízen iniciačními faktory translace. Zkratky 20 standartních a dvou doplňkových aminokyselin jsou uvedeny v **Tab. 1.2**.

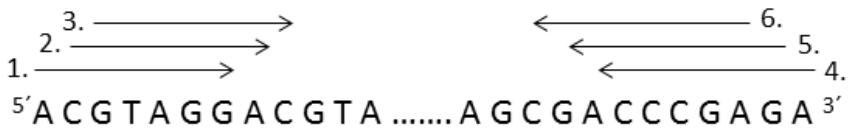
Každá sekvence DNA obsahuje 6 možných čtecích rámců (reading frame), viz **Obr. 1.3**. Jde o pozici a směr možného čtení sekvence. Otevřený čtecí rámec (open reading frame – ORF) je část obsahující možný start kodon, vnitřní kodony ale stop kodon není přítomen.

Tab. 1.1 Význam kodonů v mRNA [1].

1. nukleotid		2. nukleotid U		2. nukleotid C		2. nukleotid A		2. nukleotid G		3. nukleotid	
U		UUU	Phe (F)	UCU	Ser (S)	UAU	Tyr (Y)	UGU	Cys (C)	**U**	
		UUC		UCC		UAC		UGC		**C**	
		UUA	Leu (L)	UCA		UAA	**STOP**	UGA	**STOP**	**A**	
		UUG		UCG		UAG	**STOP**	UGG	Trp (W)	**G**	
C		CUU	Leu (L)	CCU	Pro (P)	CAU	His (H)	CGU	Arg (R)	**U**	
		CUC		CCC		CAC		CGC		**C**	
		CUA		CCA		CAA	Gln (Q)	CGA		**A**	
		CUG		CCG		CAG		CGG		**G**	
A		AUU	Ileu (I)	ACU	Thr (T)	AAU	Asn (N)	AGU	Ser (S)	**U**	
		AUC		ACC		AAC		AGC		**C**	
		AUA		ACA		AAA	Lys (K)	AGA	Arg (R)	**A**	
		AUG	Met (M) **START**	ACG		AAG		AGG		**G**	
G		GUU	Val (V)	GCU	Ala (A)	GAU	Asp (D)	GGU	Gly (G)	**U**	
		GUC		GCC		GAC		GGC		**C**	
		GUA		GCA		GAA	Glu (E)	GGA		**A**	
		GUG		GCG		GAG		GGG		**G**	

Tab. 1.2 Jednopísmenné a třípísmenné zkratky aminokyselin.

Celý název	Zkratky	Celý název	Zkratky
alanin	A (Ala)	lyzin	K (Lys)
arginin	R (Arg)	metionin	M (Met)
asparagin	N (Asn)	fenylalanin	F (Phe)
kyselina asparagová	D (Asp)	pyrolyzin	O (Pyl)
cystein	C (Cys)	prolin	P (Pro)
kyselina glutamová	E (Glu)	serin	S (Ser)
glutamin	Q (Gln)	treonin	T (Thr)
glycin	G (Gly)	selenocystein	U (Sec)
histidin	H (His)	tryptofan	W (Trp)
izoleucin	I (Ile)	tyrozin	Y (Tyr)
leucin	L (Leu)	valin	V (Val)



Obr. 1.3 Šestice možných čtecích rámců na sekvenci DNA.

Proces transkripce probíhá v jádře buňky (u eukaryotických organismů) a zjednodušeně řečeno jde o syntézu mRNA z templátu DNA. Část DNA, která je přepisována do mRNA, se nazývá transkripční jednotka (většinou odpovídá jednomu genu). Transkripce se skládá ze tří fází: 1) iniciace, 2) elongace a 3) terminace. Při iniciaci se v promotorové oblasti DNA naváže RNA polymeráza, dvojvlákno DNA je lokálně rozvinuto a vzniká transkripční bublina, kde se na templátovém vlákně DNA (komplementární vlákno ve směru 3´→ 5´) připojují transkripční faktory (speciální proteiny). Syntéza vlákna mRNA probíhá ve směru 5´→ 3´. Během iniciace se syntetizuje na templátu několik komplementárních ribonukleotidů. Pak následuje elongace, kdy je syntetizováno zbývající vlákno mRNA. RNA polymeráza se pohybuje po vlákně DNA a rozvíjí ji, syntetizují se příslušné ribonukleotidy a rozvinuté vlákno DNA je opět svinováno. Terminace (ukončení) transkripce nastává v okamžiku, kde RNA polymeráza zachytí terminační signál (specifická sekvence nukleotidů), transkripční komplex (polymeráza, DNA, mRNA, transkripční faktory) se rozpadá a vzniklá molekula mRNA je uvolněna. [1],[7]

Transkribovaná molekula mRNA obsahuje kromě kódující části i části nekódující, a proto dochází ještě k posttranskripčním úpravám. Především jde o sestřih mRNA tak, že se štěpením vyjmou introny a zůstávají pouze exony. Exony jsou většinou oblasti vlákna DNA (RNA), které kódují aminokyseliny. Mezi oblastmi exonů se u eukaryotických organismů nachází rozsáhlé oblasti intronů, které nenesou informaci pro stavbu proteinů; jejich funkce není ještě plně známa, často se zde vyskytují oblasti regulující expresi. [1],[8]

Posttranskripčně upravená molekula mRNA putuje z jádra buňky do cytoplazmy, kde probíhá proces translace, který můžeme rozdělit opět do tří fází: 1) iniciace, 2) elongace a 3) terminace. Pro translaci je nezbytná molekula mRNA nesoucí přepis genu z DNA, molekuly tRNA s jednotlivými aminokyselinami, podjednotky ribozomu a množství různých pomocných proteinů tzv. faktory. Podjednotky ribozomu utvoří po zahájení translace celý funkční ribozom, na kterém se nachází tři důležitá místa: A, P a E. Na místo A (aminoacylové místo) se váže tRNA s aminokyselinou, která bude navázána v příštím kroku. V místě P (peptidové místo) se nachází tRNA s rostoucím polypeptidem. V místě E (exitové místo) se nachází tRNA, která se oddělila od peptidového vlákna. [1],[9],[10]

Iniciace zahrnuje procesy, při nichž dochází k zahájení tvorby aminokyselinového řetězce pomocí tRNA, mRNA a ribozomů. Pro iniciaci je nezbytná přítomnost několika iniciačních faktorů. U eukaryot se iniciační komplex tvoří na 5´ konci mRNA a hledá se kodon AUG pro zahájení translace. Po nalezení iniciačního kodonu AUG opouštějí komplex iniciační faktory a komplex ribozom/mRNA/tRNA je připraven na elongaci.

Elongace (prodlužování řetězce) probíhá ve směru 5´→ 3´ po molekule mRNA. Jednotlivé aminokyseliny jsou k sobě navazovány peptidovými vazbami. Po každém navázání se ribozom posune po mRNA o jeden kodon. Elongace je proces složitý,

vyžadující součinnost několika elongačních faktorů, a přitom je to proces velice rychlý a nesmírně přesný. U bakterie *E. coli* proběhne elongace o jednu aminokyselinu za ~0,05 sekundy.

Terminace nastává v případě posunu ribozomu na jeden z terminačních (stop) kodonů na mRNA (UAA, UAG nebo UGA u standardního genetického kódu). Tyto kodony jsou rozpoznávány uvolňovacím faktorem RF. Tento faktor připojí na konec vznikajícího aminokyselinového řetězce molekulu vody. Následně dojde k uvolnění tRNA, mRNA z ribozomu a nakonec se ribozom rozpadá na podjednotky, které jsou ihned připraveny k další iniciaci.

Příklad převodu genetického kódu z DNA do AK (jednotlivé kodony jsou pro přehlednost odděleny tečkou):

vedoucí vlákno DNA:	$^{5'}$ ATG.GCC.TGG.ACT.TCA.TAG $^{3'}$
templátové vlákno DNA:	$^{3'}$ TAC.CGG.ACC.TGA.AGT.ATC $^{5'}$
mRNA:	$^{5'}$ AUG.GCC.UGG.ACU.UCA.UAG $^{3'}$
aminokyseliny:	Met.Ala.Trp.Thr.Ser

1.4 MUTACE

Mutace, tj. změny a poškození v sekvencích DNA, se týkají změn jednotlivých nukleotidů v sekvenci, částí sekvencí, chromozomů a celých genomů. Mutace patří mezi klíčové prvky evoluce [11]. Mutace mohou být vyvolány působením mutagenů, což mohou být chemické látky nebo fyzikální jevy jako je například působení ultrafialového záření (tzv. indukované mutace).

V organismu může také docházet ke spontánním mutacím – chybám při replikaci DNA. Zde se jedná o mutace jednotlivých nukleotidů, tzv. bodové mutace (SNP –single nucleotide polymorphism). Chybovost DNA polymerázy je přibližně 1 chyba na 100 000 bází [12],[13]. To se na první pohled může zdát jako nízké číslo. Když se však vezme v úvahu, že při dělení lidské diploidní buňky je replikováno přibližně 6 miliard párů bází, dojde při tomto procesu k 120 000 chybám u každé buňky v jednom dělícím cyklu. Naštěstí však existuje několik druhů výkonných opravných mechanismů, které drtivou většinu replikačních chyb odhalí a opraví [14],[15]. Neopravené replikační chyby se stanou zafixovanými mutacemi v dalším dělícím cyklu, neboť vlákno šroubovice s chybou slouží jako templát pro replikaci. Indukované mutace jsou také opravovány podobnými mechanismy jako replikační chyby. Z 1000 takových chyb se stává méně než 1 chyba fixovanou mutací, tj. předává se do další generace [12].

Bodové mutace (změna jednoho nukleotidu) mohou poškodit ORF a způsobit chyby při přepisu do proteinu. Mezi bodové mutace patří inzerce, delece a substituce. Inzerce znamená vložení nukleotidu do sekvence; delece je vynechání nukleotidu ze sekvence. Oba tyto druhy bodových mutací způsobí posunutí čtecího rámce, čímž dojde ke změně

složení kodonů. To vede k syntéze jiného řetězce aminokyselin. Pouze v případě, že inzerce či delece zahrnuje trojici nukleotidů, je kódování zachováno a výsledný protein nemusí (ale může) být poškozen [16].

Substituce je záměna nukleotidu za jiný. Dělí se na dva druhy: tranzice a transverze. Tranzice jsou substituce nukleotidů stejného typu, tj. purinový nukleotid za purinový nebo pyrimidinový za pyrimidinový. Transverze je substituce mezi nukleotidy různých typů, tj. purinový nukleotid za pyrimidinový a naopak. U substituce nedochází k posunu čtecího rámce, ale pouze ke změně jednoho kodonu. Protože je však genetický kód značně degenerován (je redundantní), nemusí substituce nutně vést ke změně kódované aminokyseliny. K substitucím nedochází se stejnou frekvencí u každého druhu nukleotidu; vždy záleží nejen na druhu nukleotidu, ale i na nukleotidech sousedních, typu genomu (jaderný, mitochondriální, plastidový) a na typu organismu.

Mutace v rámci delších částí sekvencí DNA zahrnují inzerce, delece, duplikace, inverze, translokace a transpozice. Při duplikacích dochází ke zmnožení úseku, kdy se zmnožený úsek může nacházet hned vedle původního úseku nebo je přemístěn na jiné místo. Inverze znamená, že se úsek nepřemisťuje, ale vloží se na stejnou pozici v opačné orientaci. Při translokaci si vymění pozice dva rozdílné úseky a při transpozici se přemístí pouze jeden úsek sekvence na jiné místo v sekvenci. [1]

V průběhu dělení buněk může dojít ke ztrátě či duplikaci některého chromozomu, případně celé chromozomové sady, což jsou mutace genomové. Tento typ mutací má klíčový evoluční význam u rostlin; u živočichů je tento význam omezený a většina chromozomálních mutací má na organismus značně negativní až devastující vliv. Řada těchto mutací není u živočichů vůbec slučitelná se životem.

Mutace a poškození DNA mohou být somatické a nemají tak žádný vliv na evoluci druhu. V případě výskytu mutací v germinálních tkáních (pohlavní buňky) jsou mutace přenášeny do dalších generací [14]. Mutace také rozdělujeme podle jejich biologického vlivu, který může být pozitivní, negativní nebo neutrální. Nejvíce mutací je neutrálních či mírně negativních. Užitečné mutace jsou přirozeným výběrem fixovány.

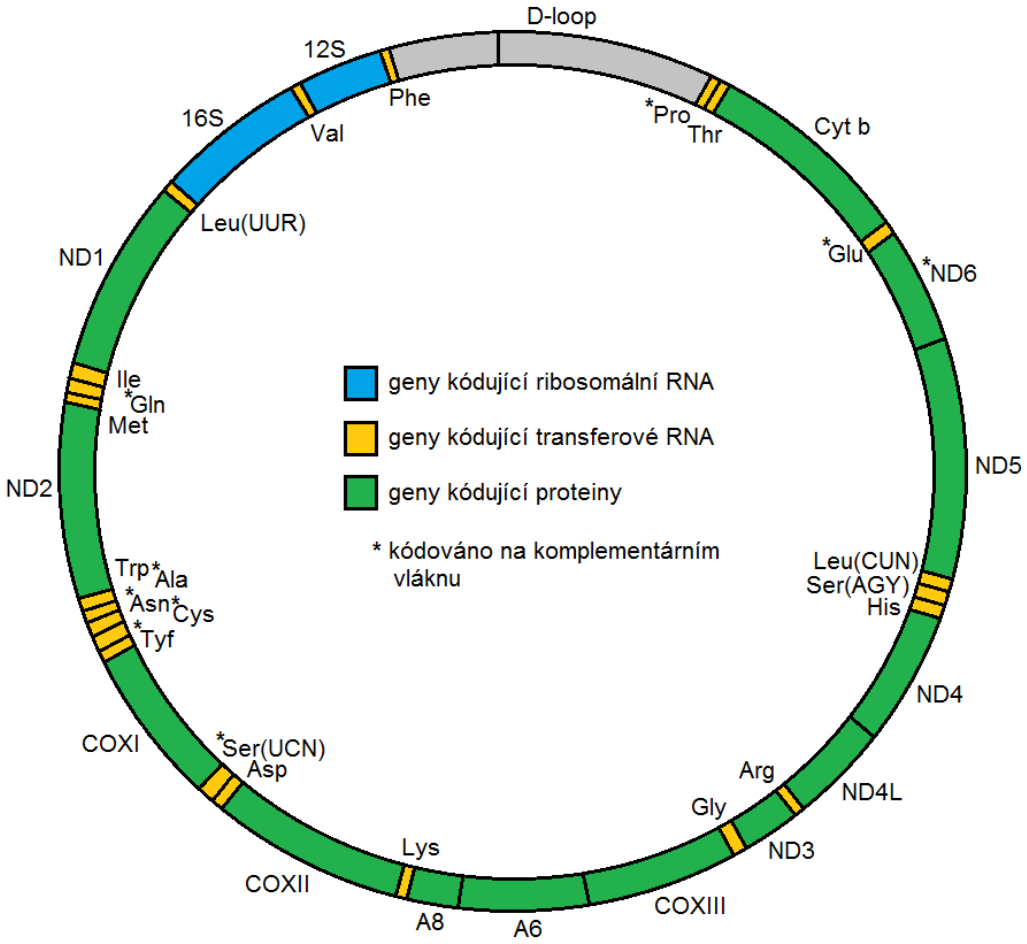
1.5 MITOCHONDRIÁLNÍ DNA

Většina eukaryotických buněk obsahuje, mimo jiné organely, mitochondrie. Mitochondrie je membránová organela oválného tvaru o velikosti v rozmezí 0,5 až 1,0 μm. Každá buňka obsahuje několik desítek až desítek tisíc mitochondrií v závislosti na energetické závislosti funkce buňky. Tato organela je specializovaná na tvorbu ATP (adenosintrifosfát), což je molekula poskytující chemickou energii pro veškeré procesy v buňce. Mitochondrie vyrábí většinu potřebného ATP. Mitochondrie se také podílí na kontrole buněčného cyklu, růstu a apoptóze (programovaná smrt buňky). Dále reguluje metabolizmus mastných kyselin, probíhá v ní Krebsův cyklus a dýchací řetězec a reguluje iontovou rovnováhu. [17],[18],[19]

Mitochondrie nese vlastní haploidní genom, tzv. mitochondriální DNA (mtDNA), který je svojí strukturou podobný bakteriálnímu genomu. Genom má formu jedné dvouvláknové kruhové sekvence DNA (bylo objeveno i množství lineárních). Každá mitochondrie obsahuje několik kopií mtDNA. Mitochondriální DNA má u většiny zvířat délku od 16 do 20 tis. párů bází. Naopak cévnaté rostliny mají mtDNA 10x až 100x delší a variabilnější. To však neznamená, že rostlinná mitochondriální DNA kóduje více genů. [17]

Genetický obsah mitochondriální DNA je u většiny organismů velmi podobný a je značně redukovaný, tj. mtDNA nekóduje všechny proteiny, které mitochondrie potřebuje ke svému fungování. Pravděpodobně došlo k přesunu většiny genů z protomitochondrie do jádra hostitelské buňky ještě před masivním fylogenetickým rozvětvováním organismů. U metazoí (vícebuněční živočichové) kóduje mtDNA 13 mRNA, které podléhají translaci, 22 tRNA a 2 rRNA [20]. Vzhledem k tomu, že předpokládaná velikost mitochondriálního proteomu je 800 – 1500 proteinů (proteom je silně vázaný na typ tkáně), kóduje mitochondriální genom jen malý zlomek proteinů a mitochondrie jsou závislé na přísunu jaderně-kódovaných proteinů z cytoplazmy [17]. Mitochondriálně kódované proteiny jsou nezbytnou součástí procesů oxidativní fosforylace. Předpokládalo se, že mtDNA byla redukována na maximální možnou úroveň, a kódovány zůstaly zachovány pouze silně hydrofobní proteiny, které by bylo obtížné transportovat do mitochondrií přes jejich dvě membrány. Avšak existuje množství jaderně-kódovaných proteinů s horšími vlastnostmi pro transport, tudíž je předcházející vysvětlení redukce mitochondriálního genomu nepravdivé. Podobně se vysvětlovalo zachování kódování molekul rRNA a tRNA, neboť transport vysoce nabitých řetězců nukleotidů přes mitochondriální membrány je obtížný. Ale na základě pozorování přenosu polynukleotidů u rostlin a kvasinek, byl i tento důvod zachování kódování zamítnut. Současně nemůže platit předpoklad, že mitochondriální genom byl redukován z původního obsahu buněčného symbionta, z kterého se mitochondrie vyvinuly, v nejvyšší možné míře, protože u mnohých rostlin a hub se vyskytují introny, repetitivní oblasti a oblasti bez funkčního významu.

Mitochondriálně kódované proteiny jsou součástí 4 komplexů: *nd1* až *nd6* a *nd4L* z komplexu I, *cyt b* z komplexu III, 3 podjednotky *cytochrom c oxidázy* z komplexu IV a 2 *ATP syntázy* z komplexu V [21]. Tyto proteiny se účastní transportu elektronů a oxidativní fosforylace. **Obr. 1.4** znázorňuje rozložení genů na lidské mtDNA. Toto rozložení se může u různých organismů lišit. U ryb, obojživelníků a savců je pořadí genů konzervované; u plazů a ptáků toto uspořádání platit nemusí [22]. Analýza uspořádání mitochondriálních genů může sloužit k fylogenetickým účelům.



Obr. 1.4 Schematické uspořádání genů na lidské mitochondriální DNA [17].

Kromě kódujících úseků obsahuje mtDNA ještě nekódující část, tzv. D-loop, která obsahuje konzervované regiony replikačních počátků a promotorů a variabilní úseky, které jsou různé i v rámci jedinců jednoho druhu. Díky geneticky neměnnému obsahu mtDNA je jednoduché určit u nově sekvenované DNA metazoí otevřené čtecí rámce. Na druhou stranu je potřeba říci, že i blízce příbuzné skupiny organismů mají samotné sekvence kódující peptidy natolik změněné, že je nelze použít jako sondy k jednoznačné identifikaci podobných sekvencí či transkriptů u jiných organismů. Mitochondriální DNA podléhá větší mutační rychlosti než jaderná DNA.

Díky vysokému počtu mutací v mtDNA se předpokládalo, že mitochondrie nemají opravné mechanismy. Avšak byly pozorovány podobné mechanismy opravy poškození způsobené oxidativním stresem jako v jaderné DNA [23],[24]. Taktéž byla pozorována oprava mismatch mutací, což jsou záměnné mutace za jiný typ nukleotidu [25].

Jelikož je mtDNA velmi kompaktní, není před každým genem přítomen úsek promotoru transkripce. Konkrétně u člověka, transkripce začíná v místě D-loop, kde jsou dva nepřekrývající se promotory vzdálené od sebe přibližně 150 bp. Jeden promotor je určen pro vlákno kódující více mRNA (HSP) a druhý pro vlákno kódující méně mRNA (LSP) a fungují nezávisle na sobě. Od LSP vzniká jeden dlouhý transkript

obsahující osm tRNA a mRNA pro podjednotku proteinu NAD6. V HSP jsou dvě iniciační místa transkripce. Z prvého vzniká transkript obsahující tRNA pro fenylalanin a valin a dvě rRNA. Tento transkript vzniká několikanásobně častěji než ostatní, aby se zajistilo dostatečné množství syntetizovaných rRNA důležitých pro translaci. Druhý typ transkriptu překrývá ten první a obsahuje všechny zbylé tRNA a mRNA [26]. Funkce D-loop je u všech obratlovců konzervována, ale různé části této oblasti se velmi liší v pořadí nukleotidů. U jiných skupin organismů je struktura D-loop rozdílná.

Vzniklé dlouhé transkripty je potřeba rozdělit na jednotlivé rRNA, tRNA a mRNA. Přesný charakter tohoto procesu ještě není plně znám. Proces translace z 13 mRNA do proteinů je částečně podobný procesu v prokaryotických buňkách, má však i svá specifika. Mitochondrie mají částečně jiný genetický kód než jaderná DNA. Především jsou zde jiné terminační kodony než standartní UAA a UAG. Jako iniciační kodon nemusí vždy sloužit jen AUG. Nelze však genetický kód mitochondrií generalizovat, protože je druhově specifický. **Tab. 1.3** shrnuje některé kodonové změny u nejprostudovanějších organismů.

Tab. 1.3 Změny v mitochondriálním genetickém kódu některých organismů [17].

Kodon	Standardní kód	Savci	*Drosophila*	Kvasinky	Rostliny
UGA	STOP	Trp	Trp	Trp	STOP
AGA, AGG	Arg	STOP	Ser	Arg	Arg
AUA	Ile	Met	Met	Met	Ile
AUU	Ile	Met	Met	Met	Ile
CUU, CUC	Leu	Leu	Leu	Thr	Leu
CUA, CUG	Leu	Leu	Leu	Thr	Leu

Ačkoliv genetický kód má 64 kodonů, mtDNA kóduje pouze 22 tRNA, tudíž by každá z těchto tRNA měla být schopna rozpoznat všechny kodony z dané čtyř-kodonové rodiny. Například u metazoí existuje jediná tRNA pro metionin (formyl-metionin) s antikodonem CAU (kodon AUG), který však musí být schopen rozpoznat i ostatní kodony z rodiny AUN ale i ostatní start kodony UUG, GUG a GUU. [17],[27]

Mitochondriální genom je populárním markerem molekulární diverzity v různých oblastech jako je populační biologie [28],[29], fylogenetika [30], fylogeografie [31] a molekulární ekologie. Popularita plyne z jednoduché extrakce mtDNA z buňky, absence intronů, malého počtu inzercí a delecí, a především krátkost a jednoduchost tohoto genomu [32]. Využití mitochondriálního genomu k určení molekulární diverzity je založeno na třech předpokladech, které jsou však v současné době podrobovány značné diskuzi [30],[33].

Prvním předpokladem je klonalita, která říká, že mtDNA je předávána pouze po mateřské linii a veškerá mtDNA od otce je zničena ve velmi krátkém čase po oplodnění vajíčka. Součástí předpokladu je také nerekombinační charakter mtDNA (vyjma hub a rostlin) [34],[35],[36]. Avšak již bylo pozorováno množství výjimek jak rekombinace, tak

zachování otcovské DNA [37],[38],[39],[40]. Omezená platnost klonality tedy musí být brána v potaz při interpretaci genealogických vazeb mezi druhy vzešlých z analýzy mitochondriálního genomu.

Neutrální mutace nebo mutace mírně delečního charakteru jsou druhým předpokladem. Adaptační mutace byly považovány za velmi vzácné [41]. Moderní výzkum odhalil, že tento předpoklad je také značně omezený [42],[43],[44].

Třetím předpokladem je konstantní evoluční rychlost. Před rokem 1979 se myslelo, že mitochondriální DNA má nízkou evoluční rychlost (menší než jaderná DNA), aby zůstala zachována konzervovanost a funkčnost kódovaných proteinů [45]. Bylo však prokázáno, že mitochondriální DNA mutuje rychleji než jaderná, což je pravděpodobně důsledek většího oxidativního stresu způsobeného funkcí mitochondrií [46],[47],[48].

I přes omezenou platnost klonality, neutrality mutací a konstantní evoluční rychlosti, je mtDNA stále používaná v množství různých studiích. Například při identifikaci současně žijících organismů nepředstavuje omezená platnost výše popsaných předpokladů závažný problém. V současnosti je populární DNA barcoding používající k identifikaci organismů část mtDNA (viz kapitola 2.2).

2 MOLEKULÁRNÍ FYLOGENETIKA

Molekulární fylogenetika (či molekulární systematika) je moderní vědní obor, který se prostřednictvím molekulárních znaků, jako jsou DNA či RNA sekvence a proteiny, snaží identifikovat druhy organismů, vysvětlit evoluční vztahy mezi organismy a také organismy taxonomicky klasifikovat.

Na Zemi existuje obrovský počet druhů organismů a to především prokaryotických. Je těžko odhadnutelné, kolik druhů bakterií se na Zemi vyskytuje, zvláště vezmeme-li v úvahu, že bakterie mají velkou schopnost adaptace a velkou evoluční rychlost. Je také složité určit, co už je a co ještě není znakem nového druhu bakterie. Naproti tomu, odhad počtu druhů eukaryotických organismů se s každým rokem zpřesňuje. Široký interval 3 – 100 milionů druhů byl v roce 2011 zúžen na 8,7 mil. ± 1,3 mil. [49]. Z tohoto počtu se odhaduje, že 86 % suchozemských a až 91 % mořských živočichů ještě nebylo popsáno (z větší části jde o organismy bezobratlé). Je to více než 250 let, co švédský biolog a lékař Carl Linné představil systém třídění a nomenklatury organismů a považuje se za otce moderní taxonomie či systematiky. Od té doby však bylo popsáno a zatříděno pouze okolo 1,2 milionu organismů.

Biologický druh je souborem populací jedinců, kteří mají shodné morfologické, etologické, ekologické a především genetické vlastnosti. Současně jsou tyto vlastnosti odlišné od vlastností ostatních druhů. Druh také musí mít soubor vlastností, podle kterých dochází k určování reprodukčního partnera v rámci druhu, a také soubor vlastností, které znemožňují rozmnožení s jinými druhy (neuvažujme mezidruhovou hybridizaci, kdy jsou potomci většinou sterilní či málo životaschopní). Jelikož evoluce druhů probíhá stále, existují i nyní přechodné druhy, jejichž vlastnosti ještě nejsou od předka plně separované; především jde o vlastnosti reprodukční. Také existují druhy, které jsou morfologicky shodné, ale v ostatních vlastnostech jsou rozdílné. Tyto druhy se označují jako kryptické a jsou odhalitelné především analýzou molekulárních znaků. [50]

Výše zmíněná definice druhu odpovídá realistickému pojetí druhu. Existuje však ještě nominalistické pojetí druhů, které říká, že reálně existují pouze jedinci, mezi nimiž je přirozená variabilita, a ostrá hranice mezi druhy je uměle vytvořená. [51]

Klasická taxonomie založená na morfologických, behaviorálních a environmentálních znacích se potýká s několika základními problémy. Ke klasifikaci organismů do druhů je potřeba znalého odborníka; automatická analýza není možná. Vzhledem k šíři a diverzitě rostlinné i živočišné říše je potřeba úzce specializovaných taxonomů. Klasická taxonomie se především opírá o morfologické znaky, u kterých je však velmi obtížné určit hranice mezi jednotlivými znaky tak, aby přesně vymezily jednotlivé druhy. Dále jsme také omezeni množstvím znaků, které můžeme zkoumat, nebo také pohlavním dimorfismem, morfologickými odlišnosti larválních stadií apod. Oproti molekulárním znakům, které mají převážně kvalitativní charakter (např. že v

proteinu je na pozici 5 valin), jsou morfologické znaky spíše kvantitativní (např. počet šupin kolem očí) nebo pravděpodobnostní či „fuzzy“ (např. asi o třetinu větší čelní výrůstek).

Molekulární fylogenetika pracuje s molekulárními znaky, kterých je možné získat obrovské množství v závislosti na velikosti genomu a proteomu organismů. K analýze není potřeba celého nepoškozeného organismu; teoreticky postačí jediná buňka. Molekulární fylogenetika je také schopná určit příbuznosti i mezi velmi odlišnými organismy, i když výsledky nemusí být vždy správné. U molekulárních znaků není potřeba stanovovat jejich relativní váhu k určení fylogenetické příbuznosti druhů. Vzájemné vážení znaků je klíčové pro klasickou taxonomii a zásadně ovlivňuje výsledky; přitom jde o subjektivní proces. Pro molekulární znaky stanovení vah buď úplně odpadá, nebo je jejich stanovení provedeno na základě nějaké přesně vymezené metodiky či algoritmu. [51]

Dalším problémem, se kterým se potýká taxonomie založená na morfologických znacích, je konvergentní a paralelní evoluce způsobená podobnými či stejnými selekčními tlaky působícími na vzájemně nepříbuzné organismy, což vede ke vzniku podobných morfologických znaků. Pomocí molekulárních znaků je možné homologní, paralelní a konvergentní znaky od sebe odlišit. Ačkoli má klasická taxonomie množství nevýhod nelze ji plně nahradit molekulární fylogenetikou, například při studiu anageneze (změny fenotypových vlastností v rámci vývojové linie či taxonu).

Fylogenetická analýza založená na jednom molekulárním znaku (jednolokusová), např. jednom genu, může dobře popsat fylogenezi daného genu, nemusí to však být plně v souladu s fylogenezí organismů. Častou příčinnou je horizontální mezidruhový přenos týkající se enzym-kódujících genů u prokaryot a jednobuněčných eukaryot. Nebo také vertikální mezidruhový přenos polymorfismu, kdy pro daný gen existuje několik alel, jejichž vzájemná divergence není v souladu s časovým vývojem jejich speciace.

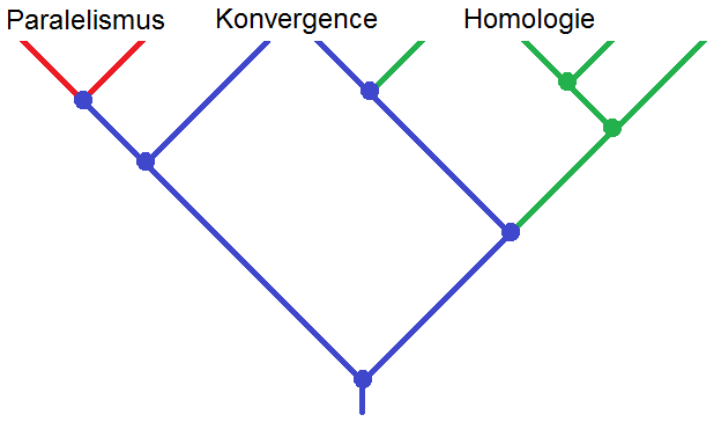
Jedním z prvních vědců, který použil molekulární znaky k určení příbuznosti organismů, byl George H. F. Nuttall. Ve své publikaci z roku 1904 „*Blood Immunity and Blood Relationship: A Demonstration of Certain Blood-Relationships Amongst Animals by Means of the Precipitin Test for Blood*“ uvedl výsledky své studie příbuznosti živočišných druhů na základě měření množství proteinového sedimentu při přidání krevního antiséra jednoho organismu do krevního séra jiného organismu. Z pokusů vyvodil, že čím jsou si organismy vývojově bližší, tím je připravené antisérum referenčního organismu reaktivnější v sérech nejpříbuznějších organismů. Mimo jiné dospěl i k výsledku, že moderní člověk si je příbuznější s hominidy Starého světa než Nového světa. [52]

2.1 FYLOGENETICKÉ A KLASIFIKAČNÍ METODY

Molekulárně-biologické metody poskytují data dvojího druhu: znaková a distanční. Znaková data jsou kvalitativní i kvantitativní; nesou informaci o vlastnostech a stavech daného zkoumaného znaku (např. výskyt aminokyseliny na určité pozici, nebo délka fragmentu z elektroforézy). Naproti tomu distanční data pouze určují míru rozdílnosti mezi znaky, nelze z nich určit konkrétní stav/hodnotu znaků.

Z molekulárních dat lze odhadnout evoluční vztahy mezi druhy organismů, tj. vytvořit hypotézu o jejich příbuznosti. Primárně se využívají znaky žijících organismů; klasická taxonomie může zahrnout i informace z fosilních nálezů, z kterých se molekulární znaky získávají velmi obtížně. Fylogenetika vyhodnocuje podobnosti znaků.

Společné znaky několika druhů zděděné od nejbližšího společného předka se nazývají homologické znaky. Homoplazické znaky jsou podobnosti mezi druhy, které vznikly nezávisle a nesdílí je jejich nejbližší společný předek. Dále se podobnosti znaků dělí na paralelní, konvergentní, analogické a reverzní. Paralelní podobnost znaků vzniká u blízce příbuzných druhů, tuto podobnost však nesdílí jejich společný předek. Konvergentní podobnost vzniká mezi druhy nezávisle na existenci společného předka. Analogická podobnost má původ např. ve vykonávání stejné funkce (podobný tvar těla tučňáků a ploutvonožců). Reverzní podobnost znamená, že znaky potomků připomínají znaky vzdáleného předka. Homoplazické podobnosti jsou pro fylogenetiku málo informativní či rovnou zavádějící. Rozlišení typu analyzovaného znaku patří mezi klíčový problém fylogenetiky. **Obr. 2.1** znázorňuje některé typy podobností.



Obr. 2.1 Dendrogram zobrazující paralelní, konvergentní a homologní podobnost.

Fylogenetické metody, např. metoda maximální úspornosti (parsimonie) a metoda maximální věrohodnosti (likelihood), pracují pouze s molekulárními daty znakovými; většina ostatních metod pracuje s daty distančními. Všechna znaková data je možné převést na distanční, naopak to však možné není.

V současnosti se k fylogenetické rekonstrukci používají sekvence DNA, RNA a proteinů. Jelikož v těchto sekvencích dochází k mutacím, musíme brát v potaz jejich vliv. K mutacím nedochází ve všech místech DNA sekvence se stejnou četností, různé druhy mutací mají různou pravděpodobnost a také se budou tyto mutace různě fylogeneticky fixovat v populaci. Ve fylogenetice se ve spojitosti s distančními metodami používá řada evolučních modelů, ale především je potřeba nejprve sekvence navzájem zarovnat.

2.1.1 ZAROVNÁVÁNÍ SEKVENCÍ

Nukleotidové nebo aminokyselinové sekvence je nutné před konstrukcí fylogenetického stromu zarovnat. Zarovnání sekvencí zajistí, že budou porovnávány odpovídající si části sekvencí. Toto předzpracování sekvencí je zcela klíčové pro většinu bioinformatických analýz.

Existují dvě základní metody zarovnávání - globální a lokální. Oba druhy metod hledají v porovnávaném páru sekvencí podobnosti, stejné nebo maximálně podobné nukleotidy/aminokyseliny jsou zarovnány pod sebe, inzerce a delece jsou v zarovnání znázorněné mezerami v jedné či druhé sekvenci. Globální zarovnání hledá podobnosti po celé délce sekvencí, a je proto vhodnější pro sekvence podobných délek. Výsledkem jsou dvě zarovnané sekvence stejných délek. Lokální zarovnání vyhledá pouze nejpodobnější část sekvencí, tudíž ho lze využít tam, kde je jedna sekvence výrazně kratší. Výsledkem jsou dvě zarovnané sekvence ne nutně stejných délek. [53]

Pro zarovnávání je nutné definovat skórovací systém, který bude ohodnocovat jednotlivé párové shody či neshody mezi sekvencemi. K tomu slouží substituční matice, která vyjadřuje míru pravděpodobnosti změny nukleotidu nebo aminokyseliny na jiný nukleotid či aminokyselinu. V substituční matici je také vyjádřena párová shoda, většinou výrazně vyšší hodnotou než u neshodných párů. Každý znak (nukleotid nebo aminokyselina) může být během času na dané pozici v sekvenci nahrazen jiným. Pravděpodobnosti substituce nejsou pro všechny znaky stejné. Např. hydrofilní arginin může s větší pravděpodobností zmutovat na taktéž hydrofilní glutamin než na hydrofobní leucin. V případě aminokyselin, má na pravděpodobnost substituce vliv degenerativní vlastnost DNA kódu, který překládá podobné kodóny do podobných aminokyselin. [54]

U nukleotidů nejsou substituční matice nutností, postačuje ohodnocení shodného páru nukleotidů větší hodnotou než neshodného páru a inzerce/delece mají hodnotu nejmenší. Nejčastěji se používá matice NUC.4.4, která shodnému páru dává hodnotu 5 a neshodnému páru hodnotu -4. Naproti tomu u aminokyselinových sekvencí je volba substituční matice klíčová pro správný výsledek zarovnání. Pro aminokyseliny se většinou používají dva druhy substitučních matic: PAM a BLOSUM (viz dále).

Matice PAM (Point Accepted Mutation nebo Percent Accepted Mutation) byla odvozena Margaret Dayhoffovou v roce 1970. Základní matice PAM1 vychází z

předpokladu, že za daný časový úsek dojde v sekvenci ke změně 1 % aminokyselin. Tato matice byla vytvořena ze souboru globálně zarovnaných sekvencí s podobností >85 % a vychází z empirického stanovení frekvence jednotlivých substitucí. Další matice PAM vznikají násobením PAM1 se sebou. Nejvyšší je PAM250 = $PAM1^{250}$, která vyjadřuje, že za časový úsek dojde ke 250 mutacím v sekvenci o 100 aminokyselinách, tj. některá pozice podléhá vícenásobné substituci. Matice PAM s vysokým číslem se používají pro zarovnávání více rozdílných sekvencí. [55],[56]

Matice BLOSUM (BLOcks amino-acid SUbstitution Matrix) byla odvozena v roce 1992 z lokálně zarovnaných evolučně vzdálených sekvencí. Číslo matice udává podobnost bloků. BLOSUM62 byla vytvořena z bloků s 62% podobností. Čím vyšší číslo matice, tím podobnější sekvence pomocí matice zarovnáváme. [57],[58],[59]

Základním algoritmem pro globální zarovnání je Needlemanův-Wunschův dynamický algoritmus, který počítá matici ohodnocení každého páru nukleotidů/aminokyselin. Následné zpětné trasování v matici nalezne zarovnání s maximálním skóre. Algoritmus nalezne nejlepší možné zarovnání dle použité substituční matice a penalizace mezery (inzerce/delece). [60],[61]

Pro lokální zarovnání je základem Smithův-Watermanův algoritmus, který se od Needlemanova-Wunschova algoritmu liší jen v několika bodech. V sekvencích nalezne pouze úsek s maximálním skóre [62]. Princip lokálního zarovnání byl pak použit i pro základní techniky vícenásobného zarovnávání, tj. současného zarovnávání tří a více sekvencí. [63],[64],[65]

2.1.2 ALGORITMUS BLAST

V současnosti pravděpodobně nejpoužívanějším algoritmem pro zarovnávání je BLAST (Basic Local Alignment Search Tool), který patří do rodiny lokálně zarovnávacích algoritmů. Tento algoritmus a jeho různé vylepšené verze především slouží k vyhledávání homologních sekvencí v databázích na základě lokálních podobností a jejich statistických významností. [66],[67]

BLAST na rozdíl od Needlemanova-Wunchova a Smithova-Watermanova algoritmu využívá heuristického přístupu, na jehož základě nedochází k porovnávání každého znaku sekvence s každým, ale místo toho se nejprve porovnávají krátké subsekvence, tzv. slova, která pak tvoří jádra zarovnání. Pokud je ke slovu v databázi sekvencí nalezena shoda s hodnotou skóre nad zadaným prahem *T*, prodlužuje se zarovnání od jádra v obou směrech.

Prvním krokem algoritmu je rozdělení zadané sekvence na slova o zadané délce. Standardní délka slova jsou tři znaky. Slova jsou ze sekvence vytvořena v posuvném okně, např. první slovo je tvořeno prvním až třetím znakem, druhé slovo je tvořeno druhým až čtvrtým znakem atd. Pro každé slovo se následně hledá shoda v databázi sekvencí. Jako jádro pro prodlužování zarovnání se vezmou pouze ty shody, jejichž skóre je vyšší než zadaná hodnota *T*, v případě nukleotidů se používá úplná shoda slova.

Skórování je dáno volbou substituční matice. K jádru zarovnání jsou v obou směrech přidávány další znaky, dokud hodnota skóre nenabude lokální maximální hodnoty, či skóre nepoklesne o *X*. Pokud má prodloužené zarovnání skóre vyšší než je zadaná mezní hodnota *S*, pak je toto zarovnání součástí výsledku vyhledávání.

Níže je příklad principu algoritmu BLAST pro sekvenci nukleotidů, skórovací parametry: shoda 5, neshoda -4.

hledaná sekvence	ATTAGACAG	
knihovna slov $W = 3$	ATT, TTA, TAG, AGA, ACA, CAG	
příklad jádra	TTA	
porovnávaná sekvence	GCGATTAGCCT	
prodlužování zarovnání	ATTAG GCGATTAGCCT	skóre 25
	ATTAGA GCGATTAGCCT	skóre 21
výsledek s max. skóre	ATTAGAC GCGATTAGCCT	skóre 26

Pro každé nalezené zarovnání se skóre nad *S* je vyhodnocena *e-value*, což je hodnota vyjadřující počet možných náhodných výsledků se stejným skórem v databázi. Čím více se *e-value* blíží hodnotě 0, tím je šance náhodné shody menší a zarovnání je tudíž pravděpodobněji správné = kvalitnější.

Databáze GenBank pod správou NCBI (National Center for Biotechnology Information) používá BLAST jako standardní nástroj k vyhledávání nukleotidových i proteinových sekvencí. Výsledkem vyhledávání je list sekvencí s nejvyšším dosaženým skórem a nejnižší hodnotou *e-value*.

2.1.3 EVOLUČNÍ MODELY

Evoluční modely se používají ke korekci vícenásobných, zpětných, paralelních a konvergentních mutací v sekvencích DNA, které se následně podrobí fylogenetické analýze. V některých případech, kdy jsou si sekvence velmi podobné, není ani potřeba použít evoluční model a počítá se pouze se vzdálenostmi vypočítanými jako podíl počtu rozdílů mezi zarovnanými sekvencemi ku jejich délce (*p*-distance), což je hodnota:

$$p = \frac{n_d}{L},$$

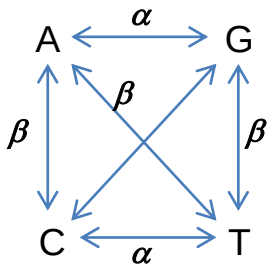
kde n_d je počet rozdílných znaků a L je délka zarovnaných sekvencí včetně zarovnávacích mezer.

Používané evoluční modely uvažují dva základní druhy mutací: transverze a tranzice. Transverzní mutace je substituce purinové báze za pyrimidinovou a naopak. Tranziční mutace je substituce purinové báze a purinovou nebo pyrimidinové báze za pyrimidinovou.

Nejjednodušším účelným evolučním modelem je Jukesův-Cantorův model (JC) [68]. Tento model předpokládá, že všechny transverzní i tranziční mutace jsou na všech pozicích v sekvenci stejně pravděpodobné. Mezi dvěma sekvencemi pozorujeme p rozdílů (substitucí, inzercí nebo delecí) a podle modelu můžeme vypočítat pravděpodobný celkový počet substitucí K, ke kterým došlo během vývoje sekvencí od jejich oddělení od společného předka:

$$K = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right).$$

Tento jednoduchý model dobře popisuje sekvence s malým počtem změn. Pro geneticky vzdálenější sekvence (s větším počtem mutací) je vhodnější použít některý sofistikovanější model, který předpokládá rozdílné pravděpodobnosti tranzicí a transverzí. Mezi takové modely patří dvouparametrický Kimurův model (K2P), který tranzicím přiřazuje pravděpodobnostní parametr α a transverzím parametr β. Jelikož je transverzních kombinací dvakrát více než tranzičních je celková hodnota změn rovna $\alpha+2\beta$ (viz **Obr. 2.2**) [69].



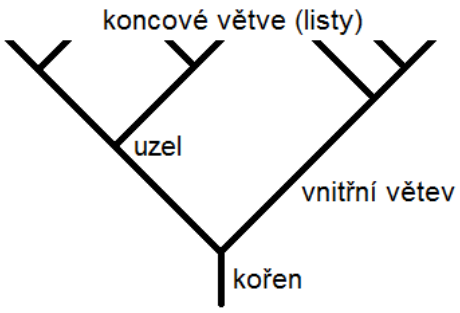
Obr. 2.2 Kimurův 2-parametrický model pro tranzice a transverze.

Další, propracovanější, modely (např. Tamurův-Neiův) přidávají parametry zohledňující rozdílné pravděpodobnosti různých druhů tranzicí a transverzí. Do dnešních dnů vzniklo mnoho různých modelů zohledňujících různé vlastnosti rodin sekvencí či časové variace evolučních rychlostí. [70]

2.1.4 DENDROGRAMY A FYLOGENETICKÉ STROMY

Dendrogram (zjednodušeně strom) je grafickým znázorněním kladogeneze, tj. způsobu rozvětvování vývojových linií organismů v průběhu evoluce. Strom je tvořen čarami (větvemi), které se postupně rozdělují v místech zvaných uzly až do koncových úseků zvaných listy reprezentujících jednotlivé druhy. Vnitřní větve reprezentují hypotetické či vymřelé předky druhů. Společného předka všech druhů zahrnutých do analýzy představuje kořen stromu, což je větev, ze které se všechny ostatní větve oddělují. Délky větví a úhly, které větve mezi sebou svírají, nemají žádnou informační hodnotu.

Příklad dendrogramu znázorňuje **Obr. 2.3**. Fylogenetický strom se od dendrogramu liší tím, že délka větví vyjadřuje dobu trvání druhů nebo množství evolučních změn.



Obr. 2.3 Schéma dendrogramu.

Nekořenový dendrogram zobrazuje pouze vzájemné podobnosti mezi druhy bez předpokladu společného předka. Tento typ dendrogramu lze vytvořit z každého zakořeněného stromu prostým odstraněním kořene, ale z nezakořeněného stromu není možné jednoduše vytvořit strom zakořeněný, neboť nevíme, mezi kterými uzly by se kořen měl nacházet.

2.1.5 METODY KONSTRUKCE FYLOGENETICKÝCH STROMŮ

Esenciální součástí molekulární fylogenetiky je tvorba fylogenetických stromů, což jsou diagramy zobrazující vzájemnou příbuznost mezi druhy, případně v časovém měřítku zobrazují i jejich evoluci. Fylogenetické stromy jsou většinou konstruovány na základě analýzy sekvencí DNA nebo proteinů. K tomuto by se měly výhradně používat sekvence z homologních sekvencí, tj. pocházejících od společného předka, i když jejich podobnost již může být malá.

Cílem konstrukce fylogenetického stromu je nalezení nejpravděpodobnějšího schématu vývoje analyzovaných znaků/genů/druhů. Můžeme postupovat v zásadě dvěma způsoby. Z dat lze sestrojit všechny možné stromy a na základě zvoleného kritéria optimality následně vybrat ten nejvíce vyhovující. Jiný přístup zahrnuje přímou konstrukci jediného stromu na základě zvoleného výpočetního algoritmu.

Prvně zmiňovaný přístup založený na konstrukci všech přípustných stromů a kritériu optimality je výpočetně velmi náročný i pro malý počet sekvencí (malý objem dat) a výsledkem nemusí být pouze jeden strom nejlépe vyhovující kritériu. V rámci tohoto přístupu existují dvě základní metody: metoda maximální úspornosti (maximum parsimony) a metoda maximální věrohodnosti (maximum likelihood).

Základní myšlenka metody maximální úspornosti spočívá v nalezení takového stromu, který bude popisovat vývoj sekvencí pomocí co nejmenšího možného počtu událostí. To zahrnuje nalezení nejmenšího počtu událostí vedoucích ke všem sestrojitelným stromům a výběr takového stromu, který má nejnižší počet událostí. Tato metoda pracuje se znakovými sekvencemi DNA nebo proteinů, které musí být

zarovnané. Existují heuristické přístupy, na základě kterých se některé stromy z analýzy vylučují, což zrychlí vyhledávání optimálního stromu. Při použití heuristického přístupu však neexistuje jistota, že výsledkem bude opravdu optimální strom. [70],[71],[72]

Metoda maximální věrohodnosti taktéž pracuje se zarovnanými znakovými sekvencemi. Tato metoda je založena na statistickém přístupu a výpočtu věrohodností všech možných stromů a jejich délek větví pro vybraný evoluční model. Věrohodnost vyjadřuje pravděpodobnost, že daná topologie stromu nejlépe vysvětluje vývoj předložených dat. Jako nejvhodnější se pak vybírá strom s největší hodnotou věrohodnosti. Výsledky jsou silně závislé na zvoleném evolučním modelu a metoda je extrémně výpočetně náročná. Výhodou je její robustnost. [70],[73],[74],[75]

Druhý přístup konstrukce jediného stromu nepracuje se sekvencemi jako takovými, ale s maticí jejich vzájemných vzdáleností. Každý pár sekvencí je zarovnán a je spočítána jejich vzdálenost (*p*-distance), na kterou je ještě možné aplikovat některý z evolučních modelů. Každá vzdálenost je vlastně odhadem délky větve nezakořeněného stromu pro dané dva druhy. Máme tedy soubor takovýchto dvoudruhových stromů, z kterých se některou z metod snažíme sestrojit jeden *n*-druhový strom. Avšak jednotlivé *p*-distance neodpovídají přesně délkám větví v konečném *n*-druhovém stromu. Konečné délky větví stromu nejsou funkcí času, ale odrážejí míru evoluce, tj. množství změn, ke kterým došlo během nespecifikovaného časového úseku od separace druhů od společného předka. [70],[76],[77],[78]

Ke konstrukci stromu z distančních dat se nejčastěji používají metody Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [79] a spojování sousedů (neighbor-joining) [80],[81],[82].

2.2 DNA BARCODING

2.2.1 ZÁKLADNÍ PRINCIP

DNA barcoding je přístup bez jednotné metodiky, který se pomocí krátké sekvence DNA organismu snaží identifikovat jeho příslušnost k jednotlivému živočišnému nebo rostlinnému druhu [83]. Identifikací druhů se rozumí přiřazení neznámého vzorku ke známému a popsanému druhu na základě shody znaků; v případě DNA barcodingu jde o shodu v rámci sekvence DNA. Často se kromě druhové identifikace přiřazuje k DNA barcodingu také klasifikace a molekulární taxonomie. Klasifikací se rozumí třídění souboru jedinců do skupin (druhů či vyšších taxonomických jednotek) podle sdílení shodných znaků. Barcode přeloženo do češtiny znamená čarový kód, který se používá k jednoznačné identifikaci zboží nebo objektů. DNA barcode sekvence by tedy měla obdobně jednoznačně vést k jedinému druhu organismu. Ačkoliv byl tento přístup znám už dříve, teprve práce Paula Heberta z roku 2003 [82] odstartovala praktické realizování tohoto přístupu s využitím mitochondriální DNA pro živočichy a chloroplastové DNA pro rostliny. Mitochondriální DNA živočichů má v tomto přístupu několik výhod a předpokládaných vlastností: je lehce extrahovatelná z buňky, kódující úseky nejsou

odděleny nekódujícími úseky, inzerce a delece jsou méně četné než substituce, substituční mutace mají převážně neutrální charakter, DNA se dědí po mateřské linii, většinou nedochází k rekombinaci, mutační rychlost mtDNA je několikanásobně větší než v jaderné DNA a měla by být dostatečná na odlišení i nedávno separovaných druhů (viz kapitola 1.5). Platnost některých vlastností, které určují mtDNA jako vhodný marker molekulární diverzity, je však omezená [30].

Pro živočichy byla v literatuře jako univerzální DNA barcode sekvence vybrána část mitochondriálního genu *coxI* (cytochrom *c* oxidáza podjednotka I, často také označovaná jako *coI*) o přibližné délce 650 bp. Sekvence *coxI* uložené ve veřejných databázích však mohou být o něco kratší či delší v závislosti na použitých primerech a úspěšnosti sekvenace. Tento úsek genu byl vybrán Hebertem jako „nejvhodnější a univerzální“ [82], avšak výběr byl zdůvodněn pouze analýzou omezeného souboru dat a hloubková analýza ostatních úseků mtDNA do dnešního dne chybí. Rozdílnost mitochondriálních genů v omezeném souboru druhů metazoí byla sice publikována, ale nebyla vyhodnocována vnitrodruhová variabilita [84],[85]. Existují však i práce založené na jiných částech mtDNA jako *nd2* či *cyt b* [87]. Nelze tedy obecně a jednoznačně říci, že úsek mtDNA *coxI* je optimální volbou. Pro některé skupiny organismů jako jsou ryby, ptáci, některé skupiny hmyzu, *cox1* skutečně postačuje, existují však skupiny organismů, jejichž vnitrodruhová variabilita na tomto úseku znemožňuje jednoznačnou identifikaci (např. obojživelníci [88]). Také u rostlin není mitochondriální gen *coxI* vhodný, neboť je naopak málo variabilní, používají se proto různé části plastidových genů jako *trnH-psbA* (především pro kvetoucí rostliny [89]) či *rbcL* a *matK* [90]. Existuje řada více či méně univerzálních postupů pro získání DNA barcodingových sekvencích pro různé skupiny organismů [91].

2.2.2 IDENTIFIKACE DRUHŮ

K identifikaci druhů pomocí DNA barcode sekvencí se převážně používají distanční metody, tak jak byly použity v první publikaci Paula Heberta [86] a dalších následujících pracích, na kterých Hebert spolupracoval (např.: [92],[93],[94]). Jeho způsob převzala celá řada dalších autorů a dalo by se říci, že v DNA barcodingu převažuje konzervativně tento směr. Jedná se o výpočet vzájemných *p*-distancí s korekcí některým evolučním modelem (nejčastěji Kimurův dvouparametrický model) a následné vytváření fylogenetického stromu nejčastěji metodou spojování sousedů. Drtivá většina studií využívající tohoto přístupu je založena na následujícím schématu: 1. sekvenace skupiny organismů v určité oblasti, kdy je každý jedinec morfologicky identifikován; 2. zpracování DNA barcodingových sekvencí distanční metodou na dendrogram; 3. určení vnitrodruhové a mezidruhové variability; 4. zveřejnění získaných sekvencí ve veřejné databázi. Jde tedy především o získávání množství dat.

Tento široce rozšířený přístup má však i svoji skupinu kritiků, kteří preferují znakově-orientované metody či úplnou revizi DNA barcodingu jako celku. Hlavní výtka k distančním metodám plyne z faktu, že je nutné určit práh hodnoty vnitrodruhové

variability, který bude určovat, kdy se ještě jedná o stejný druh. Prahová distanční hodnota musí reflektovat vnitrodruhovou variabilitu a současně delimitovat příbuzné druhy [95]. Bohužel je vnitrodruhová variabilita různá pro různé skupiny organismů a nelze mít tedy nastavenu jedinou hodnotu a takto orientovaná identifikace často dává nesprávné výsledky především z důvodu, že se vnitrodruhová a mezidruhová variabilita často překrývají [96],[97],[98],[99]. Odhlédneme-li od faktu, že se *cox1* barcode ukázal jako nepříliš univerzální a spolehlivý, pak mnozí autoři navrhují používat k identifikaci raději znakově orientované metody, které by měly být pro identifikaci druhů spolehlivější. [100],[101],[102],[103]

Distanční a některé znakové přístupy k DNA barcodingu jsou převážně založené na zkonstruování fylogenetického stromu či spíše dendrogramu. Díky své velké mutační rychlosti je mtDNA vhodná k analýze druhů příbuzných skupin organismů. U méně příbuzných skupin organismů dochází právě díky velké mutační rychlosti k saturaci bodových mutací a tím k homoplazii mezi mitochondriálními genomy. DNA barcodingové sekvence samy o sobě tedy nemohou sloužit k vytváření fylogenetických hypotéz mezi odlišnými skupinami organismů ani ke klasifikaci organismů do taxonomických skupin. I stromy vytvořené pro jednu skupinu organismů jsou málo robustní a často obsahují chyby [104]. Fylogenetická hypotéza založená na jednom úseku mtDNA není validní. V analogii to odpovídá určování příbuznosti organismů pouze na základě jediného morfologického znaku. DNA barcoding nemůže sloužit k určování fylogeneze druhů, může být pouze pomocným nástrojem. V DNA barcodingu tedy nejde o sestavování fylogenetických stromů v daném slova smyslu, ale jde obecný fenogram. [105],[106],[107]

Je nutné poznamenat, že z jistého pohledu je identifikace konkrétního druhu na základě analýzy genetických odlišností nemožná, neboť druh se stále vyvíjí a tedy teoreticky je nemožné statistickými parametry definovat dynamickou entitu [108]. Z tohoto hlediska selhává v identifikaci druhu každá metoda. Avšak samotná definice druhu je nejasná a existuje několik teoretických pojetí druhů. DNA barcoding používající distanční metody je víceméně fenetickým přístupem a druh je v tomto případě souborem jedinců charakterizovaný vnitrodruhovou variabilitou a od dalších souborů jedinců (druhů) je tento druh oddělen mezidruhovou variabilitou, která by měla být několikanásobně větší než vnitrodruhová variabilita. Takovéto skupiny jedinců identifikované molekulárními přístupy dostaly různá jména, např. MOTU (Molecular Operational Taxonomic Unit) [109]. Striktně vzato DNA barcoding neurčuje příslušnost vzorku ke druhu, ale jeho příslušnost k MOTU. MOTU a druh nejsou synonyma, dokud nebude existovat jednoznačné propojení mezi druhem a variabilitou používaného molekulárního znaku. [107],[110]

Kromě identifikace druhů Hebert navrhl, že by DNA barcodingové sekvence bylo možné využít k nalezení nových druhů, především těch morfologicky kryptických [82], [111]. Tato možná aplikace rozvířila značnou diskuzi. Je možné na základě cca 650 bp dlouhého úseku sekvence určit, jestli vzorek patří opravdu novému druhu? Někteří

autoři konstatují, že určování nového druhu by mělo zůstat v doméně klasické taxonomie založené na morfologických, behaviorálních a ekologických znacích zkombinovaných s molekulárně-biologickými znaky. Samotný krátký úsek DNA nemůže mít dostatečnou vypovídací hodnotu k určení nového druhu. DNA barcoding může alespoň odhalit druhy, které jsou potenciálně nové; samotné určení by pak bylo předmětem hlubší analýzy. [100],[112]

Kromě distančních metod, existuje ještě přístup znakový, kdy je každý druh, nebo spíše MOTU, určen specifickými znaky na určitých pozicích v sekvenci. Neporovnávají se potom celé sekvence ale pouze specifické pozice. Pro znakovou analýzu DNA barcode sekvencí byl například vyvinut software CAOS (Characteristic Attributes Organization System) [113]. Zjednodušeně řečeno, CAOS pracuje tak, že v souboru sekvencí ve formátu FASTA nalezne variabilní znaky (nukleotidy), které charakterizují jednotlivé skupiny jedinců (druhů). Takto získané charakteristické znaky jsou použity jako identifikátory pro určení příslušnosti dalších sekvencí. Tento přístup se ukázal být na některé, pro distanční metody obtížné, skupiny organismů efektivnější [103]. Podobně jako CAOS pracuje i metoda BLOG (Barcoding with LOGic), která je také znakově orientovaná [114]. Pro identifikaci druhů se vyzkoušely i další přístupy jako je strojové učení [115].

2.2.3 PRAKTICKÉ VYUŽITÍ DNA BARCODINGU

Aby byl DNA barcoding cenným nástrojem, je potřeba vytvořit databázi DNA barcode sekvencí všech organismů, které byly správně identifikovány klasickým způsobem zahrnujícím morfologické a další znaky. Od každého organismu by měl být nashromážděn soubor sekvencí tak, aby se pokryla vnitrodruhová diverzita v rámci různých populací. Současně musí být stanovena metoda, kterou se budou vzorky porovnávat s databázovými druhovými referencemi. Identifikační metoda musí být založena na diagnostických kritériích, pomocí kterých bude možné spolehlivě a jednoznačně druhy odlišit. [116]

Ve světovém měřítku byla vytvořena speciální databáze The Barcode of Life Data Systems (BOLD Systems), která DNA barcodingové sekvence uchovává i s popisnými daty (morfologie, fotografie, geografické údaje, atd.). V současnosti obsahuje databáze 4 318 601 záznamů DNA barcode sekvencí z 245 635 druhů bezobratlých, obratlovců, rostlin a hub [117] (údaj k 31. 7. 2015). Kromě uchovávání záznamů, databáze také poskytuje nástroj pro identifikaci „neznámé“ sekvence. Tímto nástrojem je klasický algoritmus BLAST, který v celé databázi vyhledá sekvence s nejnižší hodnotou podobnostního skóre. Avšak ani nejnižší hodnota skóre nemusí určovat nejbližší (nejpříbuznější) sekvenci [118], tj. přiřadit neznámou sekvenci k druhu. Problém také nastává v případě, kdy druh, ke které neznámá sekvence skutečně patří, ještě není v databázi zastoupen.

V BOLD databázi je většina záznamů shrnuta do jednotlivých projektů dle skupiny organismů a případně lokality sběru dat. Každý projekt má v databázi svůj jedinečný

písmenný kód (např. MNCN = Neotropical Birds). Kromě sekvencí obsahují jednotlivé projekty popisné informace o zastoupených jedincích. Sekvence z vybraného projektu lze z databáze souhrnně získat jako soubor ve FASTA formátu. Ze sekvencí lze také vytvořit fenogram metodou spojování sousedů z distanční matice, kdy máme na výběr z výpočtu samotných *p*-distancí či aplikací evolučního Jukesova-Cantorova modelu a dvouparametrického Kimurova modelu. Pro výpočet distancí je nutné vybrat algoritmus zarovnání, kterých nabízí databázový systém několik. Databáze v současnosti také nabízí (k 31. 7. 2015) pro zvolený projekt výpis charakteristických znaků pro zvolené taxonomické skupiny např. rody, třídy atd.

Od roku 2003 do poloviny roku 2015 bylo publikováno přes 590 časopiseckých článků obsahujících sousloví „DNA barcode“ v názvu a indexovaných v databázi Web of Knowledge. Lze tedy říci, že DNA barcoding je pro svoji jednoduchost častým nástrojem pro výzkum. Nutno podotknout, že nekritické přejímání Hebertovi metodiky analýzy, může vést k dezinterpretaci některých výsledků, především z důvodu, že zavádějící jednoduchost DNA barcodingu přilákala vědecké skupiny bez patřičné orientace v oblasti taxonomie [116].

Z praktického hlediska se DNA barcoding kromě obecného zkoumání diverzity organismů uplatnil v několika specifičtějších úlohách. Jelikož je DNA barcoding založen na molekulárních znacích, je možné ho použít k identifikaci organismů, se kterými si klasická taxonomie těžko poradí. Jedná se např. o identifikaci organismů majících rozdílné morfologie v různých životních stádiích jako bezobratlí [119], ryby [120] či obojživelníci [121] nebo rozdílné morfologie sociálního hmyzu [122]. DNA barcoding take může pomoci při odhalování blízce příbuzných a kryptických druhů (např.: [123],[124],[125]), i když v této oblasti nelze dělat konečné závěry bez další analýzy jiných molekulárních i nemolekulárních znaků.

DNA barcoding se v dnešní době využívá také ke kontrole potravin, jako jsou např. mořské plody a ryby [126],[127], nebo k ověření složení přírodních léčiv [128] a dalším kontrolním testům.

3 NUMERICKÉ REPREZENTACE

Genomické (nukleotidové) a proteomické (aminokyselinové) sekvence se standardně zapisují jako sled znaků $s = s[n]$, kde $1 \leq n \leq N$ a N je délka sekvence. Jde o symbolický zápis. V symbolickém zápise jsou sekvence uložené ve veřejných databázích. Pro sekvence DNA a RNA se používá abeceda *S{A, C, G, T/U}*, kde tyto znaky ve stejném pořadí zastupují výskyt bází adenin, cytosin, guanin a tymin, a pro RNA je místo tyminu uracil. Symbolické sekvence však nejsou vhodné pro analýzu metodami digitálního zpracování signálů. Z tohoto důvodu se symbolické sekvence převádí do numerického formátu, tzv. mapováním $S \rightarrow X$ za pomoci numerické mapy.

Ideální numerická mapa musí zajistit, že výsledný digitální genomický či proteomický signál ponese stejnou informaci jako sekvence v symbolickém zápisu. Tzn., že všechny biologicky významné charakteristiky sekvence jsou vyjádřené odpovídajícím matematickými vlastnostmi výsledného digitálního signálu. Dále by numerická mapa neměla zavádět další informaci, kterou původní sekvence nenese. Současně by měla být numerická mapa dostatečně jednoduchá, aby umožňovala rychlé a efektivní zpracování a byla čitelná i pro lidského operátora.

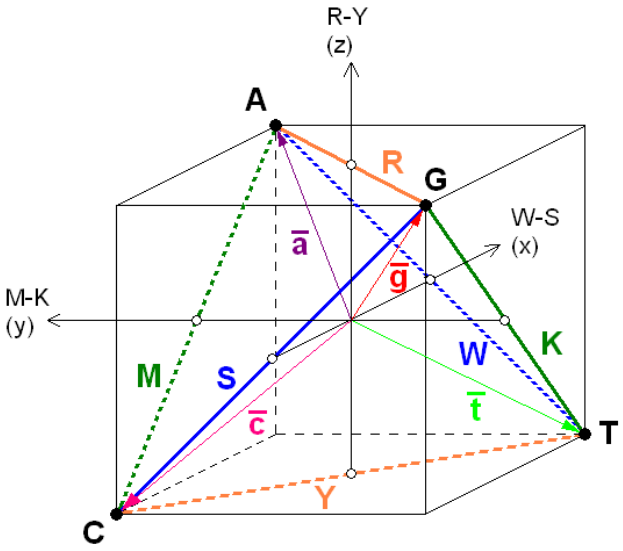
Ne všechny existující numerické mapy jsou stejně vhodné pro různé typy následného zpracování. Vhodně zvolenou numerickou mapou docílíme zvýraznění vlastností sekvencí, které následná analýza studuje. Nevhodně zvolená numerická mapa může poskytnout v dané analýze výsledky; jejich interpretace však bude obtížná, či přímo zavádějící. Pro některé druhy analýz lze zvolit numerickou mapu, která ponese jen redukované množství informace než původní symbolická sekvence.

Teoreticky lze odvodit či navrhnout nepřeberné množství numerických map. Jen některé však budou vhodné k následné analýze vybranou metodou. V následujících kapitolách budou popsány některé základní numerické mapy, které mají reálné využití. Numerických map a způsobů digitalizace genomických dat bylo publikováno mnoho, bohužel většina těchto přístupů reálné uplatnění nenašla. Vhodnost zvolené numerické mapy pro danou aplikaci je nutné důkladně otestovat a případně i srovnat výsledky se standardně používanými znakově orientovanými metodami, pokud taková metoda pro danou aplikaci existuje.

3.1 NUKLEOTIDOVÝ ČTYŘSTĚN

Mezi biologicky významné vlastnosti sekvencí patří biochemické vlastnosti nukleotidů, viz kapitola 1.1, dělené do tří skupin {R/Y}, {S/W} a {M/K}. Podle tohoto dělení můžeme vytvořit reprezentaci čtyřstěnem/tetrahedronem (viz **Obr. 3.1**) [129]. Pomocí tohoto nukleotidového čtyřstěnu je možné odvodit numerické mapy bez ztráty informace. Lze ho také rozšířit k použití pro aminokyseliny.

Každá hrana čtyřstěnu odpovídá jedné z biochemických vlastností nukleotidů a vrcholy na příslušné hraně odpovídají nukleotidům s danou vlastností. Hrany komplementárních vlastností (např. purin a pyrimidin) jsou naproti sobě kolmo orientované. Umístíme-li do středu čtyřstěnu počátek kartézského souřadného systému a osy orientujeme rovnoběžně s hranami čtyřstěnu (viz **Obr. 3.1**), pak jsou nukleotidy mapovány čtyřmi vektory symetricky umístěnými v prostoru a směřujícími do vrcholů.



Obr. 3.1 Nukleotidový čtyřstěn umístěný v pomocné krychli [129].

Zavedeme-li osy jako rozdíly mezi vlastnostmi:

$$x = W - S \quad y = M - K \quad z = R - Y,$$

a souřadnice (0,0,0) se bude nacházet ve středu čtyřstěnu, pak jsou vektory reprezentující čtyři nukleotidy vyjádřeny takto:

$$\vec{a} = \vec{i} + \vec{j} + \vec{k}$$
$$\vec{c} = -\vec{i} + \vec{j} - \vec{k}$$
$$\vec{g} = -\vec{i} - \vec{j} + \vec{k}$$
$$\vec{t} = \vec{i} - \vec{j} - \vec{k}\ .$$

Jednotlivým nukleotidům pak můžeme přiřadit tyto souřadnice:

$$A = (1,1,1)$$
$$C = (-1,1,-1)$$
$$G = (-1,-1,1)$$
$$T = (1,-1,-1),$$

což představuje třírozměrnou (3D) numerickou mapu.

V některých případech je nutné uvedený způsob numerické konverze doplnit ještě o vyjádření pomocí konvence IUPAC (International Union of Pure and Applied

Chemistry). Jde o zavedení dalších symbolů užitečných v případě, kdy v důsledku selhání sekvenace nelze jednoznačně určit typ nukleotidu. Mezi tyto symboly patří už zmíněné symboly pro skupiny (R, Y, M, K, S, W) a dále symboly pro trojice nukleotidů:

$$B = \{C, G, T\} = \sim A,$$

$$D = \{A, G, T\} = \sim C,$$

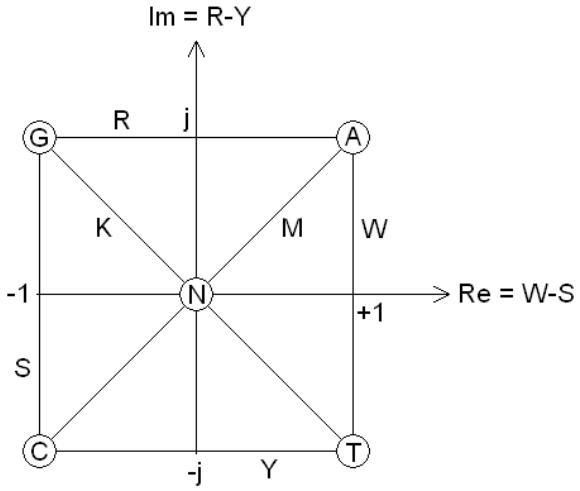
$$H = \{A, C, T\} = \sim G,$$

$$V = \{A, C, G\} = \sim T.$$

Tyto symboly pak ve čtyřstěnu zastupují následující vektory:

$$\vec{w} = \frac{\vec{a} + \vec{t}}{2} = \vec{i}$$
$$\vec{s} = \frac{\vec{c} + \vec{g}}{2} = -\vec{i}$$
$$\vec{m} = \frac{\vec{a} + \vec{c}}{2} = \vec{j}$$
$$\vec{k} = \frac{\vec{g} + \vec{t}}{2} = -\vec{j}$$
$$\vec{r} = \frac{\vec{a} + \vec{g}}{2} = \vec{k}$$
$$\vec{y} = \frac{\vec{c} + \vec{t}}{2} = -\vec{k}$$
$$\vec{b} = \frac{\vec{c} + \vec{g} + \vec{t}}{3} = -\frac{\vec{a}}{3}$$
$$\vec{d} = \frac{\vec{g} + \vec{t} + \vec{a}}{3} = -\frac{\vec{c}}{3}$$
$$\vec{h} = \frac{\vec{t} + \vec{a} + \vec{c}}{3} = -\frac{\vec{g}}{3}$$
$$\vec{u} = \frac{\vec{a} + \vec{c} + \vec{g}}{3} = -\frac{\vec{t}}{3}.$$

Tato 3D numerická mapa může být zredukována na dvourozměrnou (2D), i když se ztrátou informačního obsahu. Některé metody analýzy si vystačí i se sníženým informačním obsahem. Redukované numerické mapy získáme projekcí nukleotidového čtyřstěnu obvykle do jedné ze 3 rovin rovnoběžných se stěnami čtyřstěnu. Výběr projekční roviny závisí na volbě, kterou informaci o biochemických vlastnostech můžeme zanedbat (resp. která informace je pro nás zajímavá). Analýzou širokého rozsahu vlastností DNA sekvencí bylo zjištěno, že skupina amino/keto je méně důležitá než silná/slabá vazba a purin/pyrimidin. Skupina amino/keto nemá především vliv na komplementaritu bází, zatímco silná/slabá vazba a skupina purin/pyrimidin přímo komplementaritu určují. Výběru zachování informačního obsahu o skupinách R/Y a W/S odpovídá projekce čtyřstěnu do roviny komplexní *x*-*y* (viz **Obr. 3.2**). [130]



Obr. 3.2 Reprezentace nukleotidů v komplexní rovině *x-y*.

V této 2D reprezentaci leží na kladné části reálné osy nukleotidy tvořící slabou vazbu, na záporné části reálné osy leží nukleotidy tvořící silnou vazbu. V kladné části imaginární osy jsou pak umístěny puriny a v záporné části pyrimidiny. Jednotlivé nukleotidy jsou analyticky vyjádřené:

$$a = 1 + j$$
$$c = -1 - j$$
$$g = -1 + j$$
$$t = 1 - j \quad .$$

Symboly IUPAC jsou v reprezentaci v rovině *x-y* vyjádřeny:

$$w = 1$$
$$y = -j$$
$$s = -1$$
$$r = j$$
$$k = m = n = 0$$
$$d = \frac{1}{3}(1 + j)$$
$$h = \frac{1}{3}(1 - j)$$
$$b = \frac{1}{3}(-1 - j)$$
$$v = \frac{1}{3}(-1 + j) \quad .$$

Existuje mnoho dalších numerických reprezentací, které však nevycházejí z modelu nukleotidového čtyřstěnu. Závisí na metodě zpracování sekvencí DNA, jakou numerickou reprezentaci zvolíme. Špatně zvolená numerická reprezentace může vyústit v naprosto špatné výsledky analýzy. Nejjednodušší z numerických reprezentací jsou

jednorozměrné, používají se však i reprezentace čtyřrozměrné. Dále jsou některé z nich popsány.

3.2 DALŠÍ NUMERICKÉ MAPY

Reprezentace reálnými čísly

Jedná se o jednu z nejjednodušších metod numerické reprezentace DNA sekvence. Každému nukleotidu je přiřazeno jedno reálné číslo, např.:

A = 1 C = 2 G = 3 T = 4.

Tato reprezentace nenese žádné informace o biochemických vlastnostech sekvence, naopak zavádí vlastnost, kterou DNA sekvence nemají a to, že A < C < G < T (případně jinou relaci podle hodnoty přiřazených čísel), což nemá z biochemického hlediska žádný smysl. Tuto reprezentaci je však možné použít pro některé jednoduché úkoly spíše pomocného charakteru.

Binární reprezentace

Binární numerická reprezentace využívá k mapování nukleotidů pouze číslic 0 a 1. Existují dvě možnosti binární reprezentace a to buď přiřazením binárního kódu jednotlivým nukleotidům, nebo binární reprezentace vztažená na celou sekvenci.

Pro první možnost můžeme binární numerickou mapu zkonstruovat dle vybraných biochemických vlastností. Např.:

{R/Y} = 0/1 {S/W} = 0/1 {M/K} = 0/1,

pak jsou jednotlivé nukleotidy binárně vyjádřeny:

A = 010 C = 100 G = 001 T = 111.

V redukované formě pro {R/Y} = 0/1 a {S/W} = 0/1 platí:

A = 01 C = 10 G = 00 T = 11.

4D binární reprezentace

Jako druhou možnost uveďme 4D binární reprezentaci neboli také reprezentaci indikačními vektory [131]. Tato reprezentace nezachovává žádnou informaci o biochemických vlastnostech DNA sekvence, ale zachovává informaci o periodicitě výskytu nukleotidů v sekvenci. Používá se např. při zpracování sekvencí diskrétní Fourierovou transformací pro vyhledávání kódujících úseků [132],[133].

Tato metoda vytváří čtyři indikační vektory $u_A[n]$, $u_C[n]$, $u_G[n]$ a $u_T[n]$, které indikují přítomnost nebo nepřítomnost daného nukleotidu na pozici n:

$u_X[n] = 1$ jestliže $s[n] = X$,

kde *s[n]* pro *n = 1, 2, ..., N* je symbolická sekvence o délce *N*.

Příklad pro sekvenci `AGCGATGCA`, pak jsou indikační vektory rovny:

u_A = [100010001]

u_C = [001000010]

u_G = [010100100]

u_T = [000001000]

3D numerická reprezentace vzniklá redukcí 4D binární reprezentace

4D binární reprezentaci sekvencí lze zredukovat na třírozměrnou bez ztráty informace. Tato reprezentace se využívá například pro kódování sekvencí v RGB barevném prostoru. Každému z nukleotidů je přiřazen jednotkový 3D vektor směřující ze středu kartézského souřadného systému do jednoho ze čtyř vrcholů pravidelného čtyřstěnu, např.: [132]

$$A = (a_r, a_g, a_b) = (0,0,1)$$

$$C = (c_r, c_g, c_b) = \left(-\frac{\sqrt{2}}{3}, \frac{\sqrt{6}}{3}, -\frac{1}{3}\right)$$

$$G = (g_r, g_g, g_b) = \left(-\frac{\sqrt{2}}{3}, -\frac{\sqrt{6}}{3}, -\frac{1}{3}\right)$$

$$T = (t_r, t_g, t_b) = \left(\frac{2\sqrt{2}}{3}, 0, -\frac{1}{3}\right)$$

DNA sekvence je pak reprezentována třemi numerickými sekvencemi $x_r[n]$, $x_g[n]$ a $x_b[n]$:

$$x_r[n] = \frac{\sqrt{2}}{3}(2u_T[n] - u_C[n] - u_G[n])$$

$$x_g[n] = \frac{\sqrt{6}}{3}(u_C[n] - u_G[n])$$

$$x_b[n] = \frac{1}{3}(3u_A[n] - u_T[n] - u_C[n] - u_G[n])$$

Pro vykreslení takto kódované sekvence v RGB je vhodné ještě provést normalizaci hodnot do rozsahu hodnot 0 až 1.

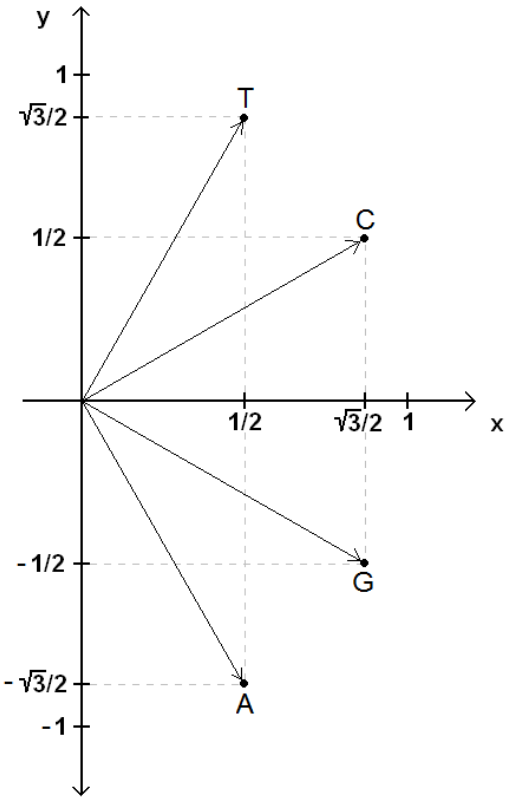
2D numerická reprezentace v 1. a 4. kvadrantu

Tato metoda zobrazuje DNA sekvenci v 2D prostoru pouze v 1. a 4. kvadrantu kartézského souřadného systému [134]. Každé z bází A, C, G a T jsou přiřazeny jedinečné souřadnice na osách *x* a *y*. Pyrimidinovým bázím (T a C) odpovídá 1.

kvadrant, purinovým (A a G) pak 4. kvadrant. Jednotkové vektory reprezentující jednotlivé báze jsou následující:

$$A = \left(\frac{1}{2}, -\frac{\sqrt{3}}{2} \right)$$
$$C = \left(\frac{\sqrt{3}}{2}, \frac{1}{2} \right)$$
$$G = \left(\frac{\sqrt{3}}{2}, -\frac{1}{2} \right)$$
$$T = \left(\frac{1}{2}, \frac{\sqrt{3}}{2} \right)$$

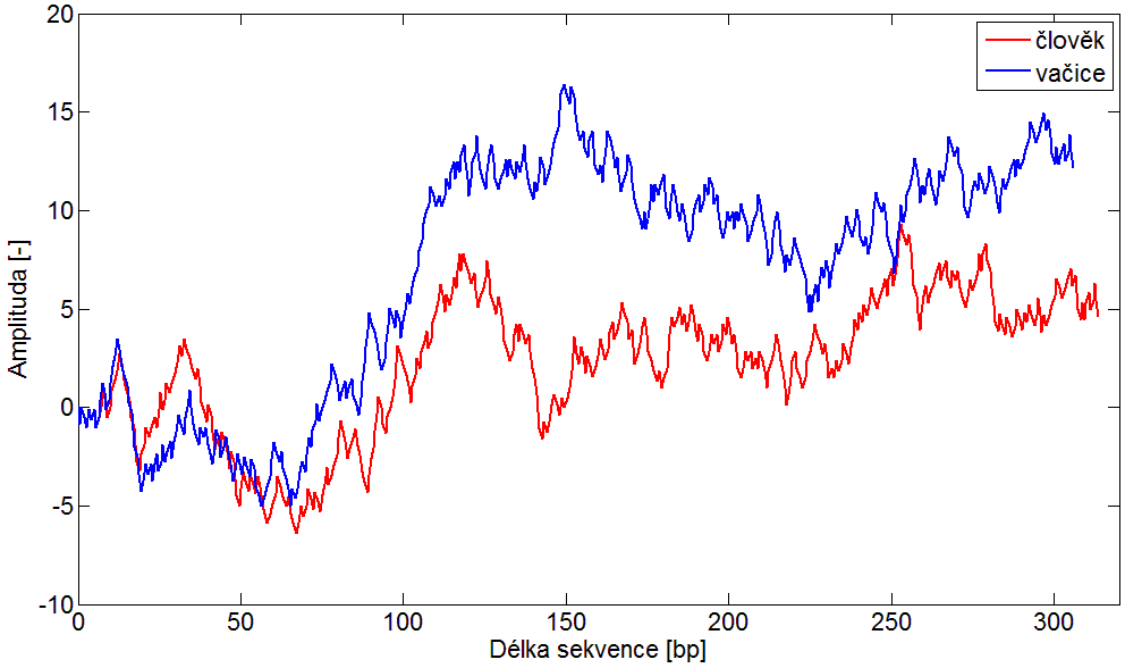
Na **Obr. 3.3** je znázorněno uspořádání jednotkových vektorů.



Obr. 3.3 Reprezentace nukleotidů v 1. a 4. kvadrantu kartézského souřadného systému.

S výhodou se tato reprezentace používá pro grafické znázornění DNA sekvence. Pokud budeme vyjádření jednotlivých nukleotidů přičítat k předchozí hodnotě, dostaneme zobrazení sekvence jako graf s počátkem v počátku souřadného systému a celá křivka sekvence se bude nacházet v kladné části osy *x*. Nedochází tím k vytváření smyček, jak by se tvořily např. při reprezentaci ve všech čtyřech kvadrantech nebo v 1. a 2. kvadrantu. Vytváření smyček způsobuje degeneraci zobrazované numericky

vyjádřené sekvence a z takového grafického zobrazení sekvence nelze zpětně zrekonstruovat původní sekvenci. Současně s tím odpovídá každé sekvenci pouze jedna grafická reprezentace [134],[135]. Příklad takové reprezentace viz **Obr. 3.4**.



Obr. 3.4 Příklad kumulované reprezentace v 1. a 4. kvadrantu pro 1. exon genu hemogloninu β pro člověka a vačici.

4D vektorová reprezentace

Každému z nukleotidů je přidělen čtyř-prvkový vektor [136]:

$$A=(1,0,0,0) \quad C=(0,1,0,0) \quad G=(0,0,1,0) \quad T=(0,0,0,1)$$

Toto přidělení je volitelné, jelikož jsou všechny směry v 4D prostoru rovnocenné, a nemá žádný vliv na následnou analýzu sekvence. Sekvence nukleotidů se pak přepisuje tak, že se vektor na pozici *n* přičte k vektoru na pozici *n-1*.

Různé numerické charakterizace sekvencí lze vypočítat z další 4D vektorové reprezentace, které lze využít k podobnostní analýze [137], nebo z jednodušší 2D reprezentace [138].

Grafické reprezentace

Grafická reprezentace sekvencí se od numerické reprezentace liší pouze tím, že je zobrazitelná např. v kartézském souřadném systému a dává dodatečnou vizuální informaci o sekvenci, která není v prostém sledu číslic (nebo čísel) patrná [139],[140],[141]. Grafickou reprezentaci můžeme například využít pro informativní porovnání sekvencí [138],[142],[143]. Metody grafické reprezentace poskytují jednoduchý nástroj k prohlížení, třídění a srovnávání různých genomických i proteomických sekvencí. Grafické

reprezentace sekvencí existují v různých dimenzích (2D, 3D i vícerozměrné). Některé grafické reprezentace mohou způsobovat degeneraci sekvence [144], tj. v reprezentaci vznikají smyčky, z kterých není možné zpětně získat původní symbolickou sekvenci. Nové metody degeneraci nezpůsobují, např. kumulovaná nebo rozbalená fáze viz níže [134],[135].

V mnohých publikacích jsou metody numerické konverze demonstrovány na porovnávání krátké sekvence 1. exonu β-globinu u různých druhů obratlovců. Hlubší analýza vlastností metod a především smysluplný návrh jejich praktického využití s demonstrací na širším okruhu reálných dat chybí. Existují samozřejmě i metody s přínosným využitím jako je např. kumulovaná a rozbalená fáze, které mohou sloužit k porovnávání celý genomů nebo chromozomů.

Kumulovaná a rozbalená fáze

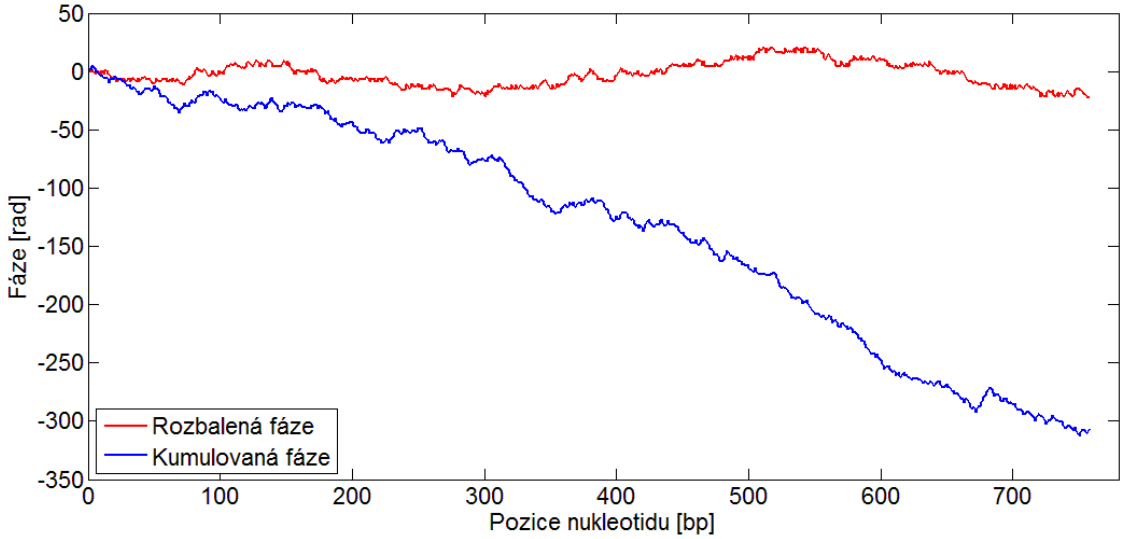
Nukleotidové sekvence mapované komplexními čísly mají fázovou charakteristiku. Jsou-li nukleotidy vyjádřené dle obrázku 3.2, pak může její fáze nabývat diskrétních hodnot úhlu v radiánech $\{-3\pi/4, -\pi/4, \pi/4, 3\pi/4\}$.

Kumulovaná fáze představuje sumu fází komplexně vyjádřených nukleotidů od prvního do aktuálního nukleotidu dle [130]:

$$\theta_M = \frac{\pi}{4}\left[3\left(n_G - n_C\right) + \left(n_A - n_T\right)\right],$$

kde n_A, n_C, n_G a n_T jsou počty příslušných nukleotidů od začátku sekvence do pozice M.

Rozbalená fáze se vypočítá ze sekvence vyjádřené komplexními čísly, kde absolutní hodnota rozdílu mezi fázemi sousedních prvků je udržována menší než π tím, že se přičítá či odečítá příslušný násobek 2π od fáze aktuálního prvku. Na **Obr. 3.5** je znázorněn příklad kumulované a rozbalené fáze sekvence mitochondriálního genu *coxI* ptačího druhu *Branta hutchinsii* (berneška aljašská). Kumulovaná fáze souvisí se statistikou jednotlivých nukleotidů, kdežto rozbalená fáze souvisí se statistikou dvojic nukleotidů.



Obr. 3.5 Rozbalená a kumulované fáze sekvence části mitochondriálního genu *coxI* druhu *Branta hutchinsii*.

EIIP hodnoty

Sekvence nukleotidů či aminokyselin lze převést do jednoho číselného vektoru dle hodnot průměrné energie všech stavů valenčních elektronů, tzv. electron-ion interaction potential (EIIP). Hodnoty pro jednotlivé nukleotidy a aminokyseliny shrnuje **Tab. 3.1**. Toto kódování lze s výhodou použít místo 4D binárního vektoru pro výpočet frekvenčního spektra sekvence pro vyhledávání kódujícího úseku DNA nebo pro analýzu společných funkcí proteinů. [145]

Tab. 3.1 Hodnoty EIIP pro nukleotidy a aminokyseliny [145].

Nukleotidy	EIIP [Ry]	Aminokyseliny	EIIP [Ry]	Aminokyseliny	EIIP [Ry]
A	0,1260	Leu	0,0000	Tyr	0,0516
C	0,1340	Ile	0,0000	Trp	0,0548
G	0,0806	Asn	0,0036	Gln	0,0761
T	0,1335	Gly	0,0050	Met	0,0823
		Val	0,0057	Ser	0,0829
		Glu	0,0058	Cys	0,0829
		Pro	0,0198	Thr	0,0941
		His	0,0242	Phe	0,0946
		Lys	0,0371	Arg	0,0959
		Ala	0,0373	Asp	0,1263

3.3 PRAKTICKÉ VYUŽITÍ NUMERICKÝCH MAP

Jak již bylo řečeno v kapitole o numerických reprezentacích, numerických map lze definovat mnoho, ale málokterá nalezla skutečně hodnotné praktické využití. V následující kapitole je diskutováno několik oblastí, ve kterých bylo aplikací numerických reprezentací a deterministických metod analýzy dosaženo potencionálně hodnotných výsledků.

3.3.1 PREDIKCE KÓDUJÍCÍCH ÚSEKŮ

Kódující úseky DNA (tzv. exony) vykazují periodicitu opakování nukleotidů rovnou 3. Tato periodicita vyjadřuje korelace pozicí nukleotidů podél kódujícího úseku a je způsobena nestejnoměrným zastoupením jednotlivých nukleotidů v rámci tří pozic v kodonech. Během evoluce se mezi skupinami organismů ustálily různé kodonové vzory exonů preferující odlišné typy nukleotidů na třech pozicích v kodonech. Každá skupina organismů preferuje i různé kodony pro stejnou aminokyselinu (tzv. codon bias). I přes to je periodicita v kódující oblasti víceméně univerzální. V nekódujících oblastech (tzv. introny) není charakteristický vzor zastoupení nukleotidů, a tudíž introny nevykazují stejnou periodicitu jako exony. [146],[147],[148],[149]

Pro predikci kódujících úseků DNA sekvencí existuje mnoho přístupů, např. se využívají HMM (skryté Markovovy modely) či metody založené na homologii sekvencí [150]. Principiálně jednoduchou signálově orientovanou metodou je vyhledání oblastí sekvence s periodicitou 3 pomocí diskrétní Fourierovy transformace (DFT). Tato periodicita se vyskytuje převážně v kódujících oblastech eukaryotických organismů. U některých prokaryotických organismů se tato periodicita vyskytuje a u některých ne.

Symbolická sekvence DNA musí být převedena do numerického formátu, který musí splňovat podmínku: do sekvence nesmí zavádět umělou periodu. Nejjednodušším a také nejčastěji používaným mapováním je 4D binární reprezentace, kdy se sekvence převede do 4 indikačních vektorů u_A, u_C, u_G a u_T. Následně se na všechny indikační vektory aplikuje diskrétní Fourierova transformace:

$$U_X[k] = \sum_{n=1}^{N} u_X[n] e^{-i\frac{2\pi}{N}k(n-1)}, \; X = A,C,G,T$$

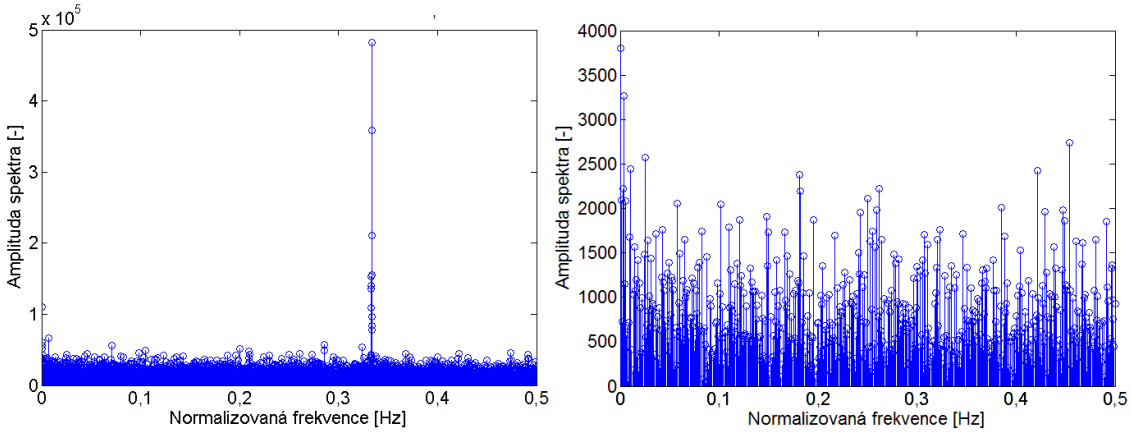
kde n je pořadí vzorku indikačního vektoru, $i = \sqrt{-1}$ a N je délka sekvence [151],[152].

Následně se vypočítá „výkonové“ spektrum (poznámka: termín výkonové spektrum je ustálený výraz vztahující se k signálům jakožto záznamům změny nějaké fyzikální veličiny, např. elektrického napětí; v případě genomického signálu však výraz výkon nemá žádný význam):

$$S[k] = |U_A[k]|^2 + |U_C[k]|^2 + |U_G[k]|^2 + |U_T[k]|^2$$

Jelikož nejvyšší frekvence opakování nukleotidů může být 1 (např. sekvence AAA…AAA) a nejde o jednotku časovou, nýbrž o vyjádření frekvence výskytu znaku v posloupnosti, budeme mluvit o „normalizované frekvenci“.

Na **Obr. 3.6** vlevo je příklad „výkonového“ spektra celého mitochondriálního genomu člověka o délce 16 569 bp. MtDNA živočichů neobsahuje introny, ale jen kódující sekvenci (vyjma nekódujícího krátkého úseku D-loop). Z uvedeného důvodu je pak v průběhu spektra zřetelný výrazný vrchol na normalizované frekvenci 1/3 Hz, který odpovídá charakteristice exonu. Pravá část **Obr. 3.6** pak zobrazuje „výkonové“ spektrum pouze části D-loop (o délce 1 121 bp), která podle očekávání nevykazuje žádnou výraznou periodicitu. Při podrobnější analýze zjistíme, že pouze části mtDNA kódující 13 proteinů mají opravdu výraznou periodicitu 3; ostatní části kódující rRNA a tRNA stejně jako D-loop periodicitu nevykazují.



Obr. 3.6 Výkonové spektrum celé lidské mtDNA (vlevo) a pouze části D-loop (vpravo).

K vyhledávání pozic exonů v extrémně dlouhých sekvencích je potřeba počítat DFT v definovaném okně s postupným posouváním okna po celé délce sekvence. Délka okna a překryv oken bude určovat poziční rozlišovací schopnost metody. Tato metoda není schopna určit přesnou pozici začátku a konce kódujícího úseku, pouze vymezí oblast, kde se tento úsek může nacházet. Při postupném procházení sekvence stačí počítat pouze frekvenční koeficient $k = W/3$, kde W je délka posuvného okna, a analyzovat amplitudu tohoto koeficientu v závislosti na pozici posuvného okna v sekvenci.

DFT lze aplikovat i na jiné numerické reprezentace než na 4D binární vektory. Stejné výsledky získáme použitím 3D RGB mapy redukované ze 4D binární, kde ve výsledku je odlišná pouze poměrově amplituda frekvenčních koeficientů. Stejně tak lze použít jednovektorovou reprezentaci EIIP hodnotami nukleotidů [153]. Tato metoda šetří výpočetní čas, neboť frekvenční koeficienty se počítají z jednoho číselného vektoru oproti 4, respektive 3 vektorům ze 4D binární či 3D RGB reprezentace.

3.3.2 PREDIKCE REPETITIVNÍCH ÚSEKŮ

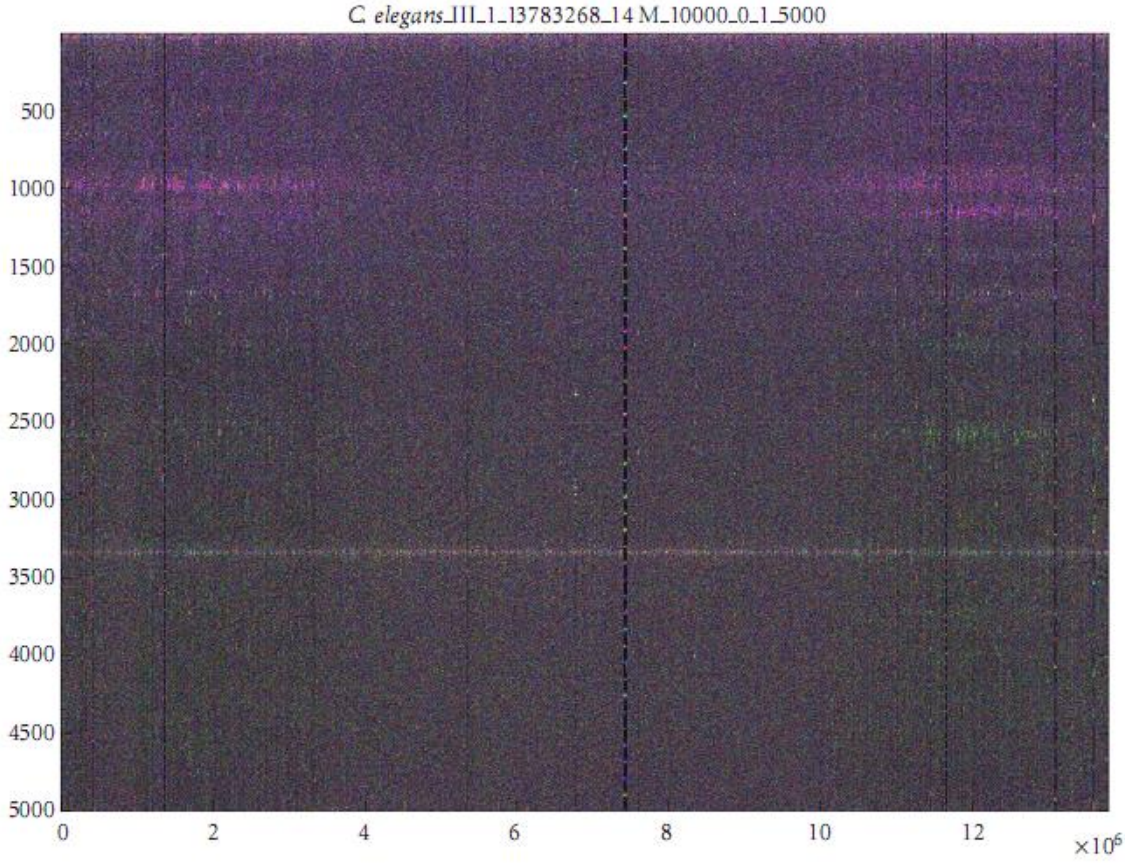
V sekvencích DNA se vyskytují oblasti, ve kterých se opakovaně vyskytuje stejný úsek (případně s mutacemi). Repetice se mohou týkat celých genů a částí chromozomů. Variace v počtu kopírovaných genů často souvisí s výskytem onemocnění a různými fenotypovými projevy, což souvisí s jejich vlivem na expresi genů. Variace repetitivních oblastí je obecně označována jako CNVs (Copy Number Variations) a dle definice to jsou oblasti o velikosti minimálně 1 kb (kilobází) [154]. CNVs tvoří podstatnou součást genomové diverzity mezi organismy [155]. I mezi lidmi jsou CNVs nedílnou součástí variability stejně jako jednonukleotidový polymorfismus [156].

Tandemové repetice jsou naproti tomu posloupnosti krátkých posloupností několika nukleotidů opakujících se za sebou i stokrát či více. Tyto úseky se často označují zkratkou VNTRs (Variable Number Tandem Repeats) nebo jako minisatelity. I v rámci jedné populace organismů může být počet opakování tandemového úseku v jednom genu různý. Vyvstává tedy otázka, jakou mají tandemové repetice pro organismus význam? Tandemové repetice jsou během replikace náchylné ke specifickému typu mutace, která vede k inzerci další repetice do sekvence, případně její deleci. Pravděpodobnost tohoto jevu je podstatně vyšší, než jiné mutace způsobené replikačním aparátem. Fondon a Garner zjistili významnou korelaci mezi tandemovými repeticemi a morfologickými znaky různých plemen psů [157]. Má se tedy za to, že VNTRs mají velký význam v evoluci druhů [158]. Nadměrné zkopírování třínukleotidové repetice v některých místech lidské DNA způsobuje vážná onemocnění jako je Huntingtonova choroba, syndrom fragilního X a mnoho dalších. Počet opakování těchto repetic koreluje se závažností choroby [1].

Tandemové repetice lze nalézt mnohými pravděpodobnostními a statistickými metodami a metodami původně určenými ke zpracování textu. Většina těchto metod potřebuje apriorně znát vzor, tj. nukleotidové složení tandemové repetice nebo alespoň délku vzoru. Nalezení repetic především znesnadňují bodové mutace vzoru. Algoritmy založené na výpočtu zarovnávacích matic mají nevýhodu v dlouhém výpočetním čase a nejsou vhodné pro sekvence delší než maximálně několik tisíc bází [159]. Principiálně jednoduchá je metoda založená na výpočtu editační vzdálenosti v rámci sekvence [160]. Metoda nevyžadující vzor ani jeho délku využívající výpočet *k*-tic shod zarovnání a statistická kritéria je popsána v [161].

Stejně jako v případě predikce kódujících úseků v DNA lze k vyhledávání repetitivních oblastí využít spektrální analýzu diskrétní Fourierovou transformací, konkrétně spektrogramy tvořené posloupností DFT v krátkém úseku. Spektrogram lze vytvořit z jednotlivě počítaných DFT pro krátké výpočetní okno (desítky až stovky bp) s překryvem či bez překryvu oken. Dobrých výsledků lze dosáhnout použitím numerické mapy 3D RGB, pomocí které zavedeme do spektrogramu i pseudobarvení, které usnadní následnou analýzu. Před vykreslením barevného spektrogramu je potřeba jednotlivé barevné složky RGB reprezentace normalizovat, aby dosahovaly pouze hodnot 0 až 1. [162],[163],[164]

Na **Obr. 3.7** je znázorň spektrogram celého chromozomu III háďátka obecného (*Caenorhabditis elegans*). Na ose *x* je vyneseno pořadové číslo nukleotidu, osa *y* má charakter „frekvence“. Světlá horizontální linie přes celou délku sekvence na pozici y rovnající se zhruba 3300 charakterizuje periodicitu 3 pro kódující oblasti. Linie je přerušovaná z důvodu výskytu intronů. Výrazná vertikální linie zhruba v oblasti x = ~7,4 Mbp představuje minisatelit o velikosti 95 bp, který se opakuje v úseku o celkové délce 30 kbp [164].



Obr. 3.7 Spektrogram chromozomu III háďátka obecného (*C. elegans*), převzato z [164]. Světlá horizontální linie představuje periodicitu 3 kódujících oblastí, vertikální linie představují minisatelity.

4 CÍLE DIZERTAČNÍ PRÁCE

Cílem disertační práce je návrh nové deterministické metody pro zpracování genomických dat, realizace této metody v programovém prostředí Matlab a její ověření na reálných datech z veřejně přístupných databází. Cíle lze rozdělit na jednotlivé složky takto:

1. Studium, testování a zhodnocení moderních přístupů k reprezentaci a analýze DNA sekvencí:
   a. studium tradičních přístupů k reprezentaci a analýze DNA sekvencí,
   b. využití numerické konverze pro analýzu DNA,
   c. metody podobnostní analýzy numericky vyjádřených genomických dat.
2. Návrh a realizace nových algoritmů numerické reprezentace a analýzy genomických dat:
   a. numerická konverze DNA sekvencí vhodná pro vzájemné porovnávání sekvencí,
   b. podobnostní analýza genomických dat ve specifickém numerickém formátu.
3. Aplikace vytvořených postupů na reálné DNA sekvence a vyhodnocení:
   a. ověření navržené numerické reprezentace a metod pro podobnostní analýzu na souboru reálných DNA sekvencí z veřejně přístupných databází,
   b. vyhodnocení výsledků podobnostní analýzy využívající novou numerickou reprezentaci.

Nová numerická reprezentace DNA sekvencí bude navrhována s ohledem na její využití pro DNA barcoding, tj. porovnávání krátkých úseků sekvencí. Metoda pro podobností analýzu využívající nového numerického formátu DNA sekvencí by měla být vhodná pro porovnání velkých souborů dat. Reálné sekvence budou získány z veřejné databáze DNA barcodingových sekvencí BOLD. U sekvencí, které nebudou navrženou metodikou správně určeny, bude provedena analýza vyhledáním homologií algoritmem BLAST v databázi GenBank.

Body 2a. a 2b. cílů dizertační práce jsou zpracovány v kapitole 5. Body 3a. a 3b. cílů jsou zpracovány v kapitolách 6, 7 a 8.

5 NUKLEOTIDOVÉ DENZITNÍ VEKTORY

Metoda denzity nukleotidů, resp. nukleotidových denzitních vektorů, představuje jednoduchý a efektivní způsob numerické reprezentace symbolické sekvence DNA či RNA. V principu vyjadřuje průměrné zastoupení jednotlivých nukleotidů v definované části sekvence. Využití této reprezentace k moderní analýze biologických sekvencí nebylo dosud publikováno v relevantní vědecké literatuře. Popis distribuce nukleotidů v sekvencích se objevuje pouze v publikacích vydaných před rokem 1980, kdy byla bioinformatika teprve na začátku svého vývoje a orientovala se pouze na kvantitativní charakterizaci sekvencí.

5.1 VÝPOČET NUKLEOTIDOVÝCH DENZITNÍCH VEKTORŮ

Pro výpočet nukleotidových denzitních vektorů (ND vektorů) se nejprve ze symbolické DNA sekvence vytvoří binární indikační vektory $b_A[n]$, $b_C[n]$, $b_G[n]$, a $b_T[n]$, které obsahují na pozici n hodnotu 1, pokud se daný nukleotid na pozici v sekvenci vyskytuje, nebo hodnotu 0 v případě, že se nukleotid nevyskytuje. V sekvencích se kromě znaků A, C, G a T pro jednotlivé nukleotidy mohou vyskytovat ještě další speciální znaky podle konvence IUPAC. Znak N, který reprezentuje neznámý nukleotid, se projeví ve všech indikačních vektorech hodnotou 0,25. Tato hodnota reprezentuje 25% možnost výskytu některého ze čtyř nukleotidů. Další používané znaky reprezentují nukleotidy o určité biochemické vlastnosti (viz kapitola 1.1). V indikačních vektorech příslušných nukleotidů se pak vyskytuje hodnota 0,5. Příklad: bude-li v sekvenci znak R (purinový nukleotid), pak v indikačních vektorech pro adenin a guanin bude hodnota 0,5, která představuje 50% možnost výskytu purinu, tedy buď adeninu A nebo guaninu G. **Tab. 5.1** shrnuje hodnoty binárních indikačních vektorů pro všechny IUPAC povolené znaky pro nukleotidy.

Nukleotidové denzitní vektory se určují z binárních indikačních vektorů pomocí průměrování v posuvném okně o definované velikosti, které se posouvá vždy o jednu pozici (jeden nukleotid). Průměrováním v okně však vzniká problém, že výsledné denzitní vektory jsou kratší než původní sekvence. Např. pro sekvenci o délce M=100 bp a velikost výpočetního posuvného okna W=5 nukleotidů získáme denzitní vektory o M-W+1=96 hodnotách. Výpočetní okno o velikosti W tak zkrátí denzitní vektory celkem o $W–1$ hodnot. Toto zkrácení není pro malé velikosti výpočetní okna významné, je však možné toto eliminovat tak, že se na začátky a konce binárních indikačních vektorů doplní přídatnými hodnotami v počtu rovnajícím se polovině velikosti posuvného okna zaokrouhleno na nejbližší nižší celé číslo. Nejlogičtější se jeví přídatná hodnota 0,25 vyjadřující stejně jako v případě znaků N 25% možnost výskytu jednoho ze čtyř možných nukleotidů před začátkem či koncem sekvence, se kterou pracujeme. Analýza vlivu hodnoty ošetření okrajů binárních indikačních vektorů je uvedena dále.

Tab. 5.1 Hodnoty binárních indikačních vektorů pro nukleotidové znaky.

Znak	*Indikační vektory*				Znak	*Indikační vektory*			
	b_A	b_C	b_G	b_T		b_A	b_C	b_G	b_T
A	1	0	0	0	**Y**	0	0,5	0	0,5
C	0	1	0	0	**K**	0	0	0,5	0,5
G	0	0	1	0	**V**	0,33	0,33	0,33	0
T	0	0	0	1	**H**	0,33	0,33	0	0,33
U	0	0	0	1	**D**	0,33	0	0,33	0,33
M	0,5	0,5	0	0	**B**	0	0,33	0,33	0,33
R	0,5	0	0,5	0	**N**	0,25	0,25	0,25	0,25
W	0,5	0	0	0,5	**-**	0,25	0,25	0,25	0,25
S	0	0,5	0,5	0					

Po úpravě (rozšíření) binárních indikačních vektorů se počítají denzitní vektory aplikováním posuvného okna o velikosti *W*:

$$d_X[n] = \frac{\sum_{i=n-W/2}^{n+W/2} u_X[i]}{W}, n = 1...M \tag{5.1}$$

kde *M* je délka sekvence a *X* je typ nukleotidu. Posuvné výpočetní okno se pohybuje po celé délce indikačních vektorů a výsledkem výpočtu je průměr z hodnot v okně. Takto získáme sadu čtyř denzitních vektorů. Velikost posuvného okna je volitelná. Volba velikosti okna je závislá na délce sekvence a na požadované rozlišovací schopnosti výsledného signálu. Při využití nukleotidové denzity jako vstupního signálu do různých metod analýzy sekvencí je potřeba velikost posuvného okna pro každou metodu optimalizovat, tj. vyzkoušet na zkušebním souboru dat různá nastavení velikosti výpočetního okna a pro další analýzy stejnou metodou použít to nastavení, které vedlo k nejlepším výsledkům.

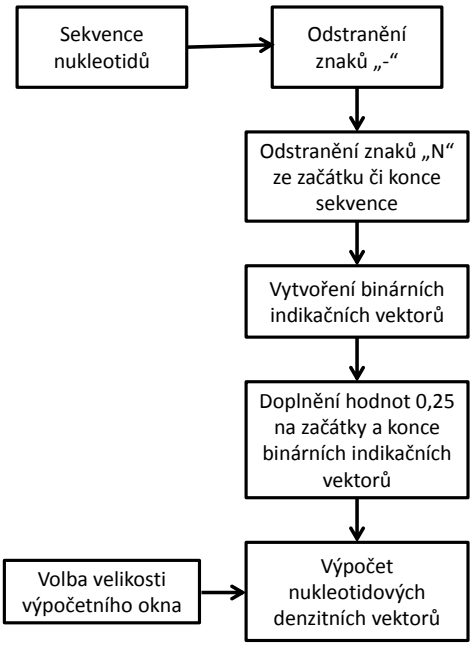
Následující **Tab. 5.2** zobrazuje příklad tvorby binárních indikačních vektorů a nukleotidových denzitních vektorů pro velikost výpočetního okna *W*=5 nukleotidů s použitím doplnění binárních indikačních vektorů o hodnoty 0,25 na začátky a konce vektorů. **Obr. 5.1** pak znázorňuje blokové schéma tvorby ND vektorů.

Jednoduchou analýzou lze zjistit, která z možností ošetření okrajů binárních indikačních vektorů je nejlepší, tj. způsobuje nejmenší zkreslení. Bylo provedeno testování ošetření okrajů pomocí přidávání hodnot 0 až 1 s krokem 0,05 pro velikosti výpočetního okna *W* =3 až *W*=29 nukleotidů. Analýza byla provedena na reálné sekvenci DNA s délkou 500 bp. Tato sekvence byla převedena na ND vektory pro různé hodnoty ošetření okrajů binárních vektorů a délky výpočetního okna. Následně byly

denzitní vektory zkráceny o *W* prvků na začátku i konci, čímž se odstranily zkreslené konce, a tyto denzitní vektory sloužily jako referenční správné. Pro vyhodnocení zkreslení byla stejná DNA sekvence zkrácena o prvních a posledních *W* nukleotidů ještě před výpočtem ND vektorů. Po převodu do denzitních vektorů pak měla stejnou délku jako referenční denzita a obsahovala zkreslené okraje. Mezi referenčními a zkreslenými denzitami byly spočítány Euklidovské vzdálenosti viz vzorec (5.1).

Tab. 5.2 Příklad tvorby nukleotidových denzitních vektorů pro *W*=5.

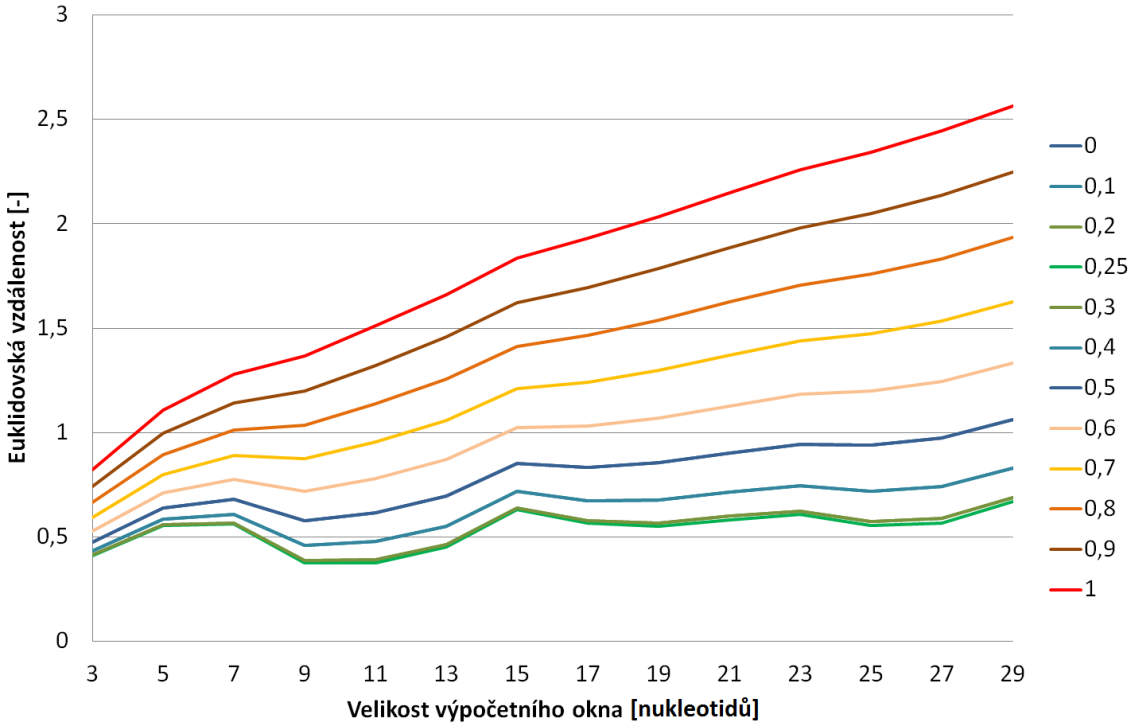
	Sekvence:		**A**	**G**	**T**	**T**	**N**	**A**	**C**	**R**	**A**	**C**		
Binární indikační vektory	b_A	.25	.25	1	0	0	0	.25	1	0	.5	1	0	.25	.25
	b_C	.25	.25	0	0	0	0	.25	0	1	0	0	1	.25	.25
	b_G	.25	.25	0	1	0	0	.25	0	0	.5	0	0	.25	.25
	b_T	.25	.25	0	0	1	1	.25	0	0	0	0	0	.25	.25
Nukleotidové denzitní vektory	d_A			.30	.25	.25	.25	.25	.25	.45	.40	.25	.30		
	d_C			.10	.05	.05	.05	.25	.25	.25	.40	.45	.30		
	d_G			.30	.25	.25	.25	.05	.05	.05	0	.05	.10		
	d_T			.30	.45	.45	.45	.45	.25	.05	0	.05	.10		



Obr. 5.1 Blokové schéma tvorby nukleotidových denzitních vektorů.

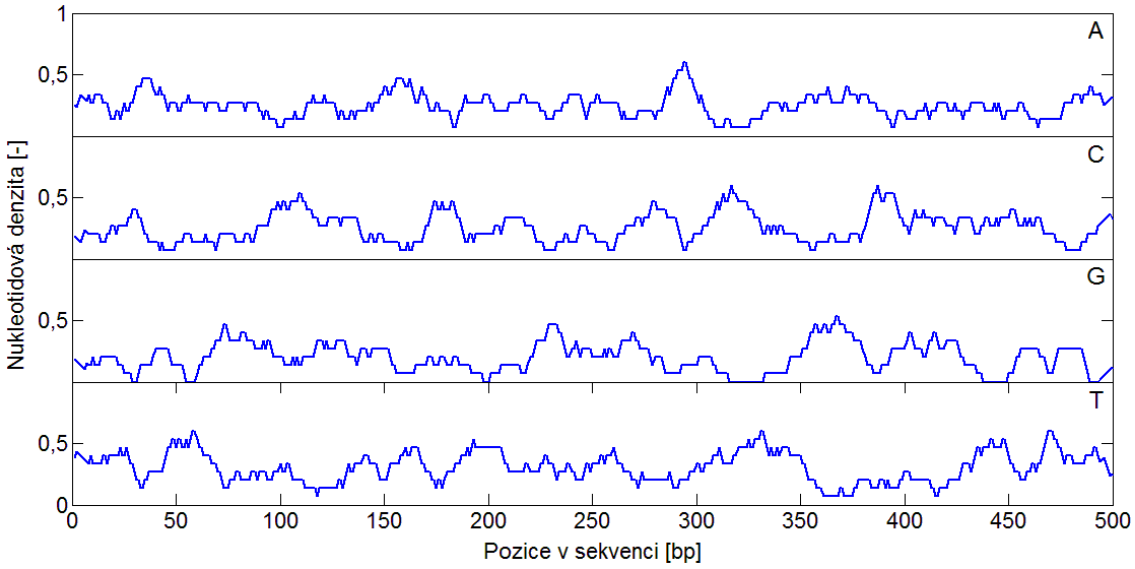
S délkou výpočetního okna Euklidovské vzdálenosti mezi denzitními vektory bez zkreslení a se zkreslením mírně rostly. Nejmenších rozdílů dosáhlo ošetření okrajů přidáváním hodnot 0,25. Ošetřování okrajů hodnotami 0 a 0,5 dávalo stejné chyby. Chybovost byla v intervalu 0 až 0,5 symetrická s minimem v hodnotě 0,25. Nad hodnoty 0,5 chybovost výrazně rostla, viz **Obr. 5.2** (pro přehlednost jsou vykresleny průběhy jen pro některé z testovaných hodnot ošetřování okrajů). Touto analýzou bylo

potvrzeno, že ošetřování okrajů binárních vektorů hodnotami 0,25 je nejlepší, tj. způsobuje nejmenší zkreslení a toto zkreslení roste s velikostí výpočetního okna jen málo.

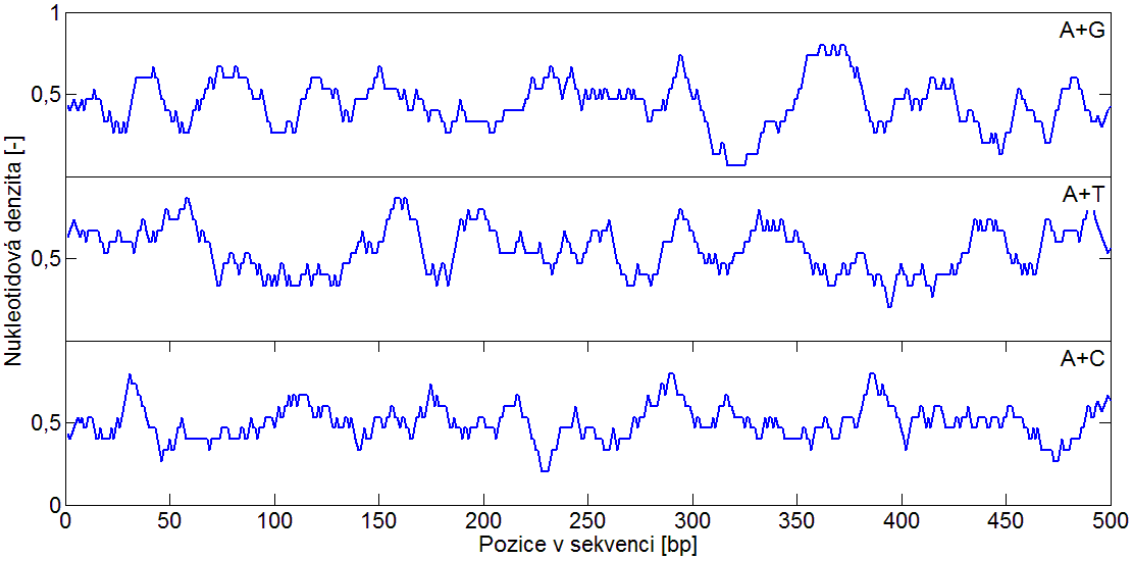


Obr. 5.2 Vliv volby ošetření okrajů binárních vektorů na zkreslení distančních vektorů v závislosti na velikosti výpočetního okna *W*. Překrývají se průběhy ošetření hodnotami 0 s 0,5, 0,1 s 0,4 a 0,2 s 0,3.

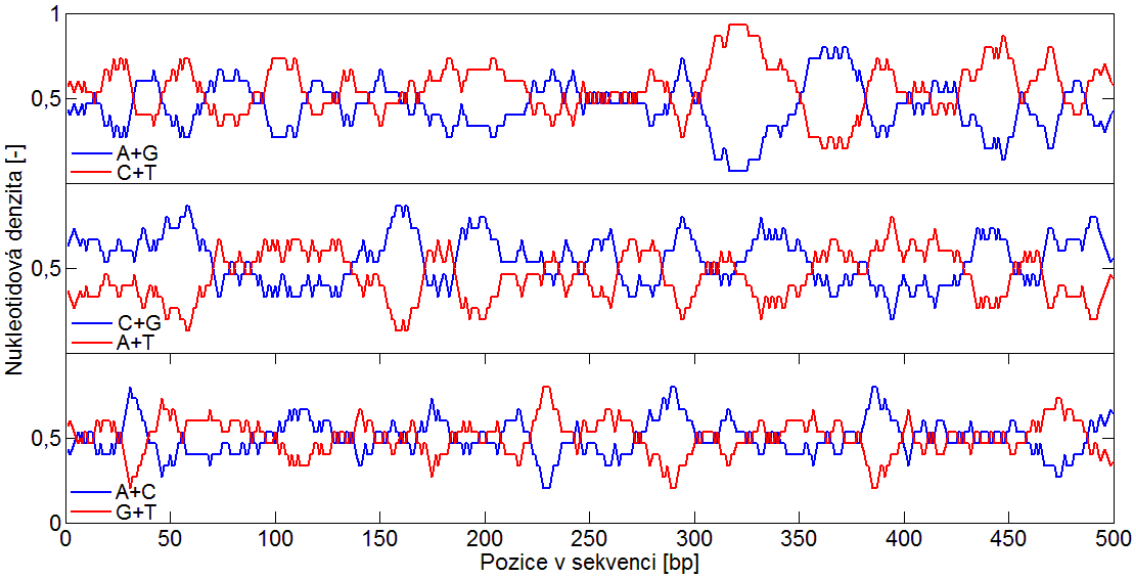
Grafická reprezentace nukleotidových denzit je možná několika způsoby. Nejjednodušší je samostatné vykreslení jednotlivých nukleotidových denzitních vektorů, viz příklad na **Obr. 5.3**. Tato vizualizace poskytuje informaci o výskytu jednotlivých nukleotidů v rámci sekvence a můžeme například rychle identifikovat části sekvence s převažujícím výskytem určitého nukleotidu. Druhý způsob vizualizace tkví v zobrazení sum nukleotidových denzitních vektorů pro nukleotidy podobných biochemických vlastností, viz **Obr. 5.4** kde jsou zobrazeny 3 možnosti sumace z 6. Sumace vektorů se řídí biochemickými vlastnosti nukleotidů dle dělení na purinové/pyrimidinové nukleotidy, nukleotidy tvořící silnou/slabou vodíkovou vazbu a nukleotidy obsahující amino/keto skupinu. Toto dělení je detailněji rozebráno v kapitole 1.1. Všech 6 možností sumace vektorů podle biochemických vlastností lze vykreslit také po dvojicích dle dané kategorie vlastností nukleotidů (např. puriny a pyrimidiny) do jednoho obrázku, viz **Obr. 5.5**. V tomto typu vykreslení je na první pohled dobře patrná komplementarita jednotlivých kategorií vlastností.



Obr. 5.3 Nukleotidové denzitní vektory části mitochondriálního genu *coxI* organismu *Canis lupus campestris*, *W*=15.

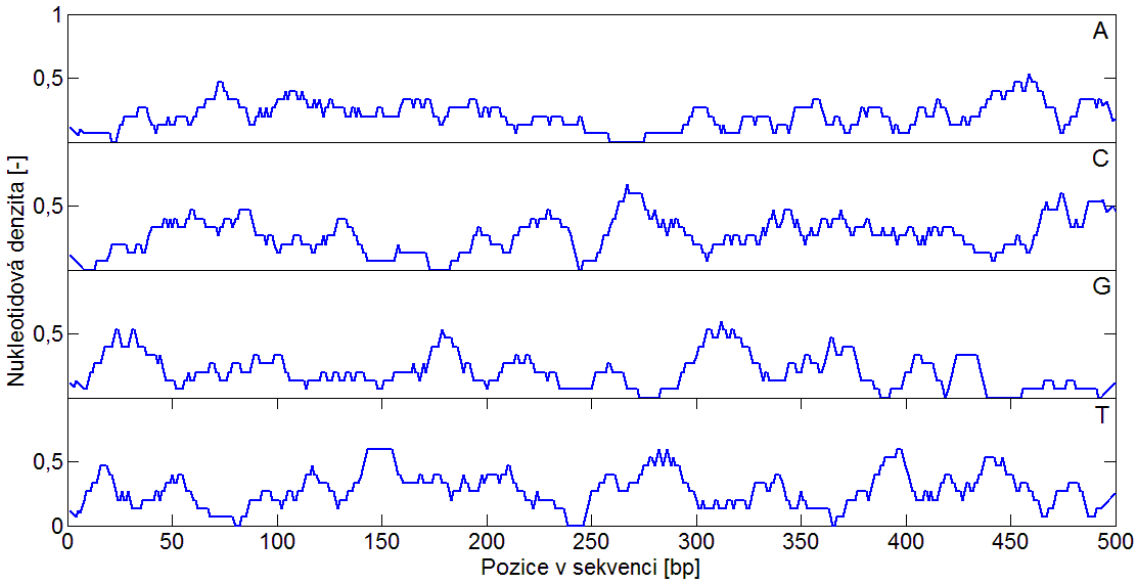


Obr. 5.4 Sumy nukleotidových denzitních vektorů pro purinové, slabou vazbu tvořící a amino nukleotidy části mitochondriálního genu *coxI* organismu *Canis lupus campestris*, *W*=15.

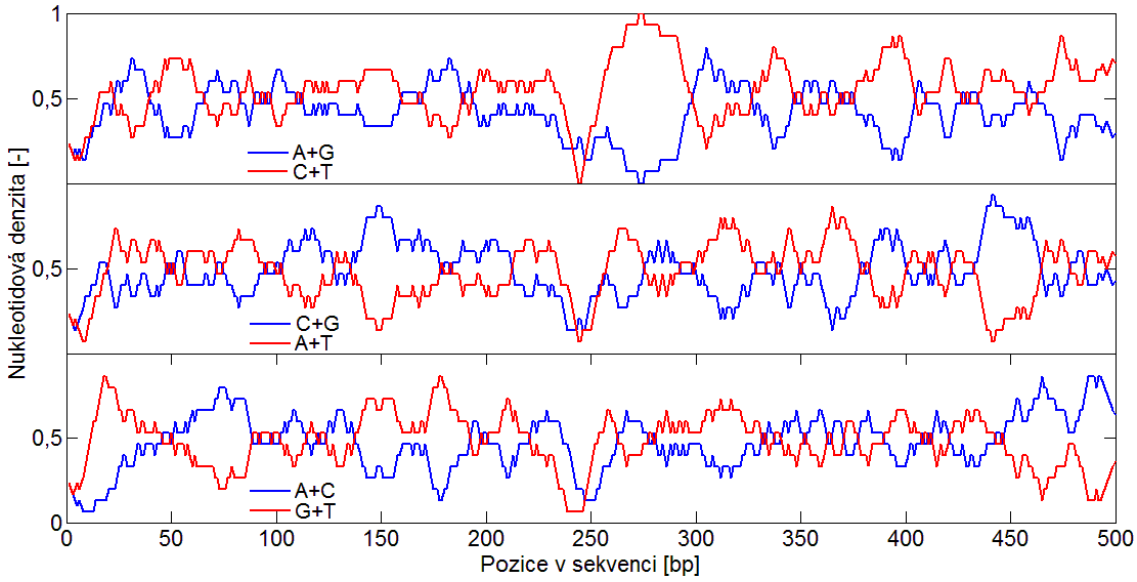


Obr. 5.5 Sumy nukleotidových denzitních vektorů části mitochondriálního genu *coxI* organismu *Canis lupus campestris* dle biochemických vlastností nukleotidů, *W*=15.

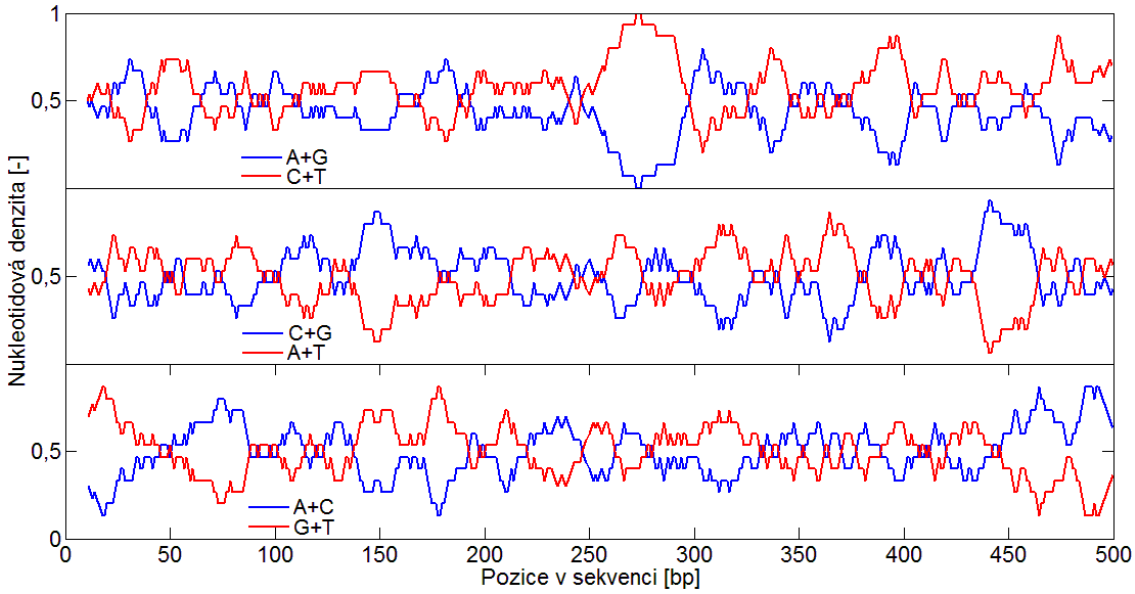
V sekvencích v symbolickém zápisu se kromě znaků A, C, G a T reprezentujících jednotlivé nukleotidy občas vyskytují i znaky N, které značí neznámý nukleotid, nebo znaky -, které pocházejí ze zarovnávání sekvencí a reprezentují inzerci nebo deleci nukleotidu. Výskyt těchto znaků je potřeba řešit již při tvorbě binárních indikačních vektorů. Pokud se znaky N vyskytují na začátku či konci sekvence lze je odstranit, neboť nemají žádný význam a jejich ponechání pouze způsobí zkreslení okrajů denzitních vektorů. Výskyt ojedinělých znaků N uvnitř sekvence lze řešit hodnotou 0,25 ve všech binárních indikačních vektorech. Zarovnávací mezery (znak -) je vhodné ze sekvence odstranit v jakékoli pozici výskytu. Pokud však budeme chtít zobrazit denzity vícenásobně zarovnaných sekvencí, pak musí být zarovnávací mezery řešeny nahrazením hodnotami 0,25 v binárních indikačních vektorech. **Obr. 5.6** ukazuje jednotlivé denzitní vektory sekvence, která obsahuje na prvních 10 pozicích znaky N a pak jsou tyto znaky přítomny na 240 až 250 pozici. Pouhou vizuální inspekcí obrázku nejsme schopni určit výskyt znaků N; znalému může přijít jen podezřelé nízké hodnoty na začátku vektorů. Záleží také na velikosti výpočetního okna. Zřetězený výskyt znaků N je lépe odhalitelný vizualizací sum denzit viz **Obr. 5.7**. **Obr. 5.8** pak ukazuje denzity vzniklé po odstranění počátečních N a nahrazením vnitřních N hodnotami 0,25 v binárních vektorech.



Obr. 5.6 Nukleotidové denzitní vektory sekvence s výskytem znaků N na pozici 1 – 10 a 240 – 250 bp, *W*=15.



Obr. 5.7 Sumy nukleotidových denzitních vektorů sekvence s výskytem znaků N na pozici 1 – 10 a 240 – 250 bp, *W*=15.



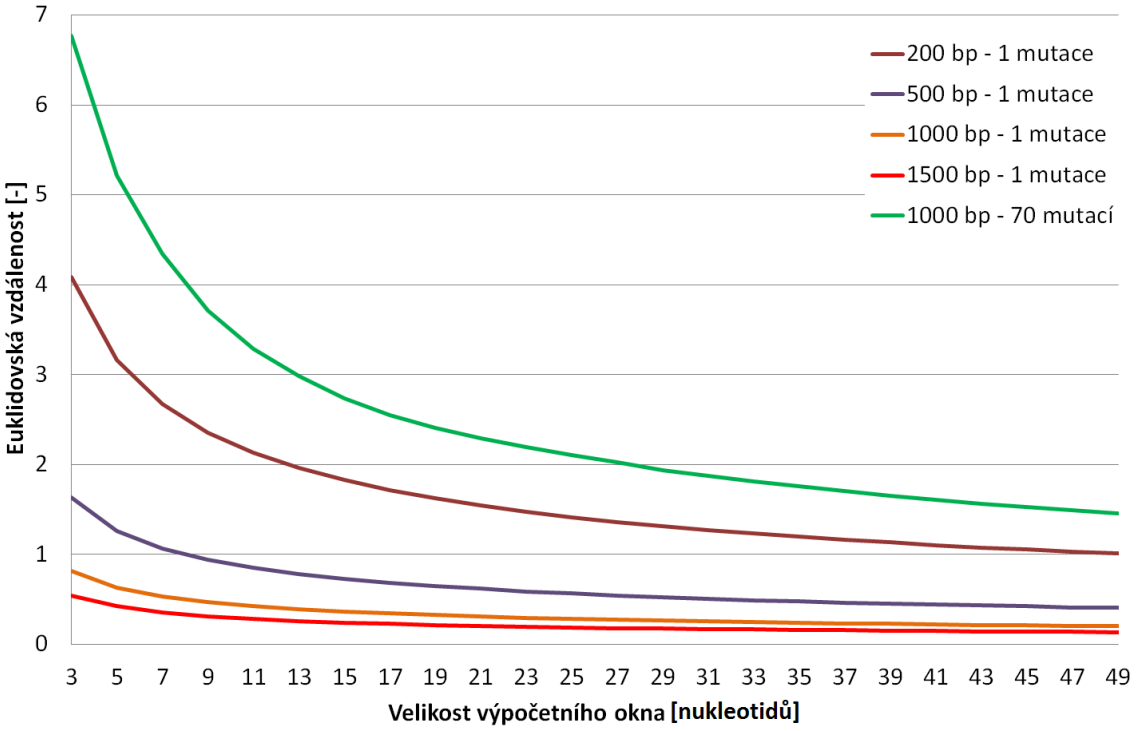
Obr. 5.8 Sumy nukleotidových denzitních vektorů sekvence s výskytem znaků N na pozici 1 – 10 a 240 – 250 bp, s korekcí odstraněním počátečních N a nahrazením vnitřních N hodnotami 0,25 v binárních indikačních vektorech, *W*=15.

5.2 VLASTNOSTI ND VEKTORŮ A JEJICH PRAKTICKÉ VYUŽITÍ

Teoretická pravděpodobnost, že bychom pro dvě rozdílné nukleotidové sekvence získali stejné denzitní vektory, je velmi malá avšak ne nulová. Nukleotidové denzitní vektory jsou závislé na pozicích jednotlivých nukleotidů a na velikosti posuvného výpočetního okna. Byla provedena analýza vlivu změny nukleotidu a velikosti výpočetního okna na Euklidovskou vzdálenost, viz vzorec (5.2), mezi dvěma sekvencemi. Jako referenční sekvence, se kterou byly porovnávány sekvence se změněnými nukleotidy, byla použita část sekvence lidského mitochondriálního genu 16S rRNA. Při změně jednoho nukleotidu na libovolný jiný nukleotid je vzdálenost mezi sekvencemi stejná, tj. nezávislá na typu změny nukleotidu. Euklidovská vzdálenost s velikostí výpočetního okna klesá s mocninou okolo ½ i pro reálné sekvence s množstvím mutací, viz **Obr. 5.9**. Obdobně klesá vzdálenost s délkou sekvence s mocninou -1.

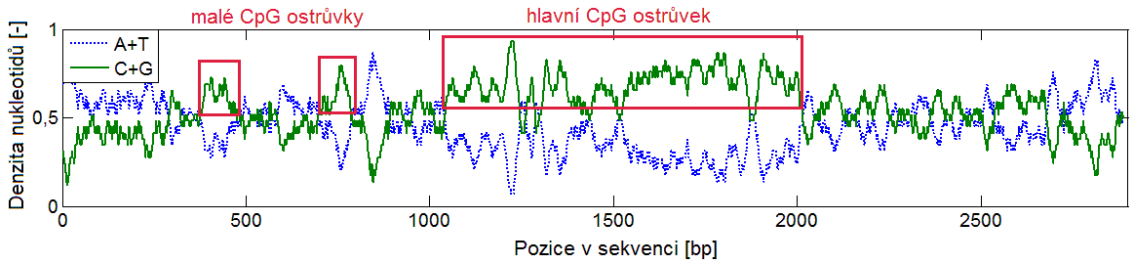
Zpětná rekonstrukce symbolické sekvence z denzity nukleotidů nemusí být jednoznačná. Především na začátku a konci sekvence není možné přesně určit polohu některých nukleotidů. Tento typ numerické reprezentace ztrácí informační obsah o přesné pozici některých nukleotidů.

Denzity náhodně vygenerované sekvence DNA nelze na první pohled odlišit od skutečné DNA sekvence a stejně tak nemají denzity sekvencí intronů od denzit sekvencí exonů charakteristické či jinak významné odlišnosti. Z vizualizace denzitních vektorů nelze říci, o jaký typ sekvence se jedná.



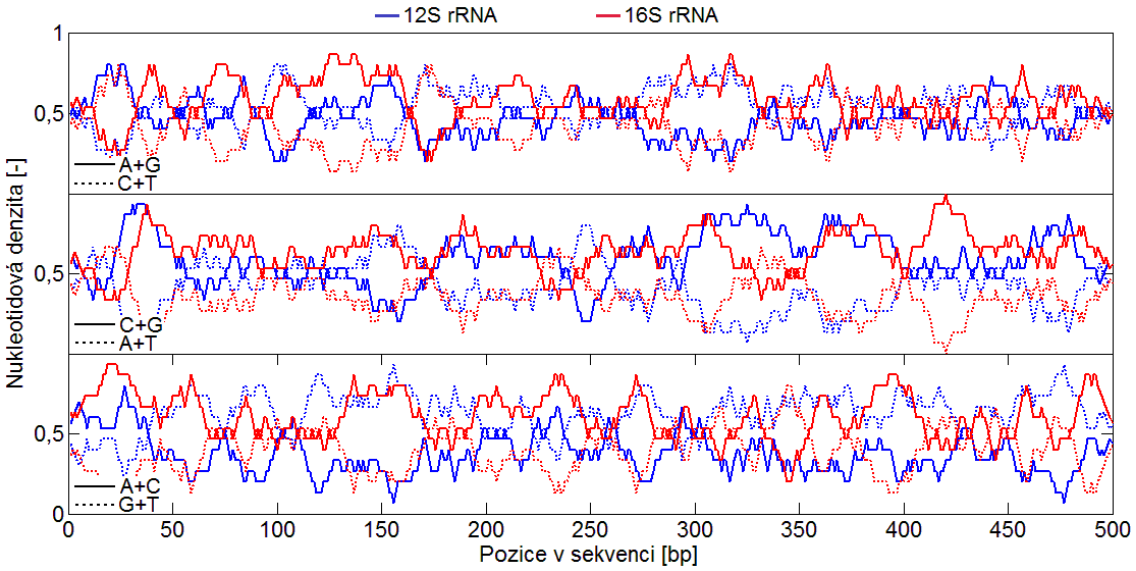
Obr. 5.9 Vliv velikosti výpočetního okna na Euklidovskou vzdálenost mezi denzitami dvou sekvencí různých délek při změně jednoho nukleotidu a distance dvou sekvencí se 70 mutacemi.

Nukleotidové denzitní vektory mohou najít uplatnění v mnohých oblastech analýzy nukleotidových sekvencí. Jelikož vektory vytváří jakýsi charakteristický vzor pro danou sekvenci, jejich využití ke komparativní genomice se samo nabízí. Nukleotidové denzitní vektory by také šlo využít k rychlému vyhledávání známého motivu v sekvenci nebo homologních sekvencí s použitím metod, které se využívají ve zpracování digitálních signálů. Lokální charakter zastoupení nukleotidů také může pomoci při vyhledávání CpG ostrůvků, tj. oblastí s vysokým výskytem cytosinu a guaninu (viz příklad na **Obr. 5.10**). Konzervovanost některých vzorů v ND vektorech může mít též spojitost s aktivními místy kódovaného proteinu. Dále je to vhodná numerická reprezentace sekvencí pro signálové zarovnávání, např. s využitím metody dynamického borcení času (dynamic time warping) [129]. Níže jsou na konkrétních příkladech ukázány některé možnosti využití denzit.

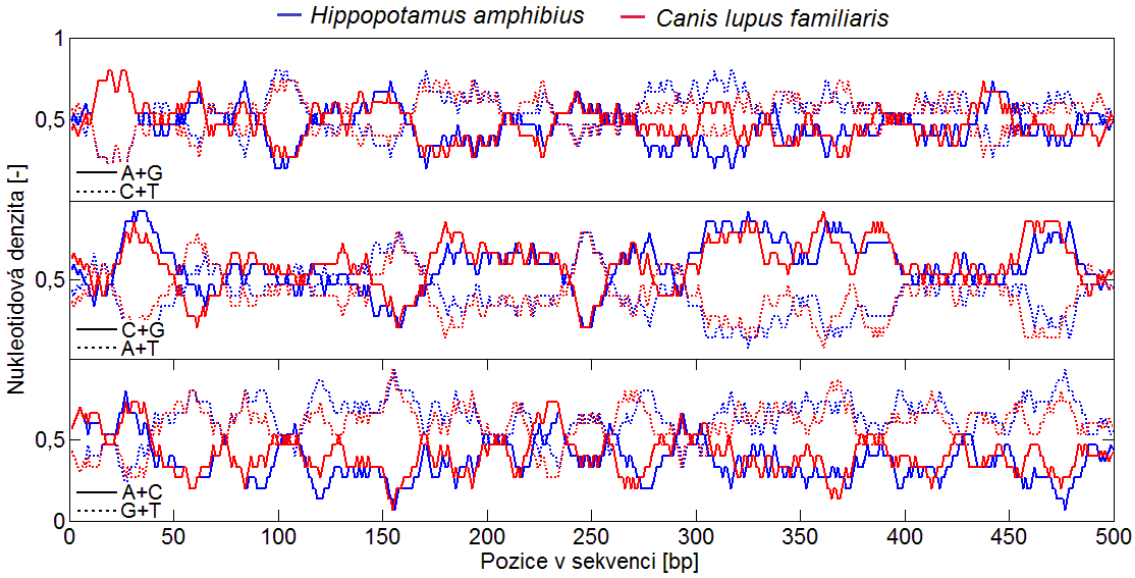


Obr. 5.10 Sumy denzit purinové/pyrimidinové nukleotidy s lokalizací CpG ostrůvků v lidské sekvenci X07398.1, *W*=49.

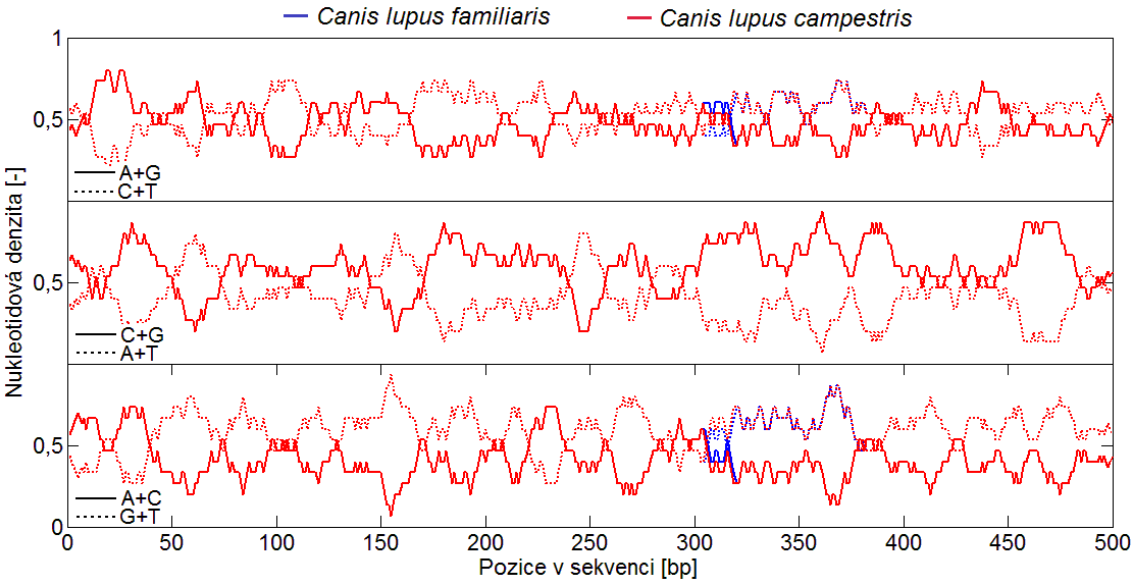
Grafické zobrazení nukleotidových denzitních vektorů pro delší výpočetní okna (od W=7 nukleotidů) dává dobrou vizuální informaci o pozici a míře odlišností mezi sekvencemi. Sekvence různých genů mají pochopitelně naprosto odlišné ND vektory, viz příklad na **Obr. 5.11** zobrazující prvních 500 nukleotidů mitochondriálních genů kódujících 12S a 16S rRNA organismu *Hippopotamus amphibius* (hroch obojživelný). Globální zarovnání těchto dvou sekvencí vykazuje 47% shodu, což je shoda náhodná. Naopak sekvence jednoho genu ale jiných organismů mají značně podobný charakter. Čím příbuznější organismy, tím podobnější ND vektory, viz **Obr. 5.12** pro dva savce *Hippopotamus amphibius* (hroch obojživelný) a *Canis lupus familiaris* (pes domácí). Na **Obr. 5.13** jsou ND vektory prvních 500 nukleotidů sekvence 12S rRNA pro blízce příbuzné organismy *Canis lupus familiaris* (pes domácí) a *Canis lupus campestris* (vlk euroasijský). Rozdíl mezi sekvencemi je pouze v jednom nukleotidu na pozici 312, kde u *Canis l. familiaris* je adenin a u *Canis l. campestris* je tymin. Proto v prostřední části obrázku, kde jsou zobrazeny sumy ND vektorů slabá/silná vazba, jsou průběhy zcela totožné.



Obr. 5.11 Sumy nukleotidových denzitních vektorů dvou různých mitochondriálních sekvencí pro rRNA o délce 500 bp organismu *Hippopotamus amphibius*, W=15.



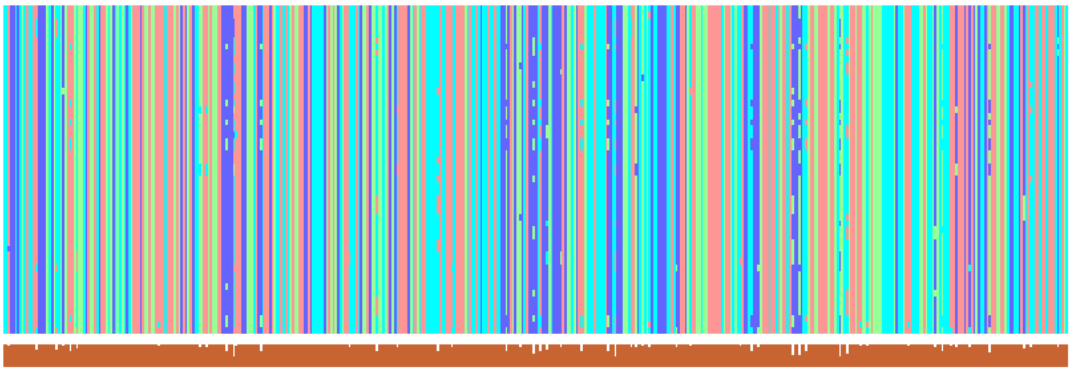
Obr. 5.12 Sumy nukleotidových denzitních vektorů dvou mitochondriálních sekvencí pro 12S rRNA o délce 500 bp organismů *Hippopotamus amphibius* a *Canis lupus familiaris*, *W* = 15.



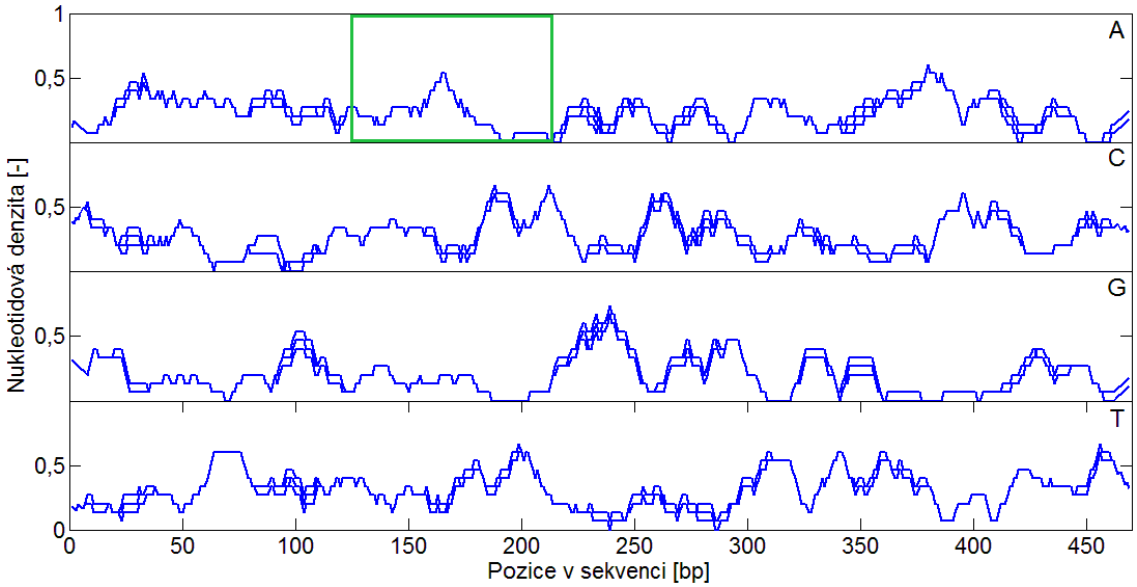
Obr. 5.13 Sumy nukleotidových denzitních vektorů dvou mitochondriálních sekvencí pro 12S rRNA o délce 500 bp organismů *Canis lupus familiaris* a *Canis lupus campestris*, *W*=15.

Analýzou ND vektorů pro jeden gen, ale pro různé organismy, lze jednoduše odhalit, v kterých částech sekvence a jaké typy nukleotidů jsou konzervované. **Obr. 5.14** znázorňuje barevně kódované vícenásobné zarovnání 52 sekvencí různých jedinců jednoho druhu *Etheostoma proeliare* 470 bp dlouhé části mitochondriálního genu *coxI*. Hnědý hřebínek pod zarovnáním zobrazuje míru konzervovanosti nukleotidů. Plná plocha v daném sloupci znamená, že na dané pozici se vyskytuje pouze jeden druh nukleotidu. Hloubka bílého místa udává míru výskytu dalších druhů nukleotidů. Čím méně je obdélník „zubatý“, tím více jsou nukleotidy v sekvencích konzervované. Z tohoto obrázku však nemáme bezprostřední informaci, které typy nukleotidů jsou

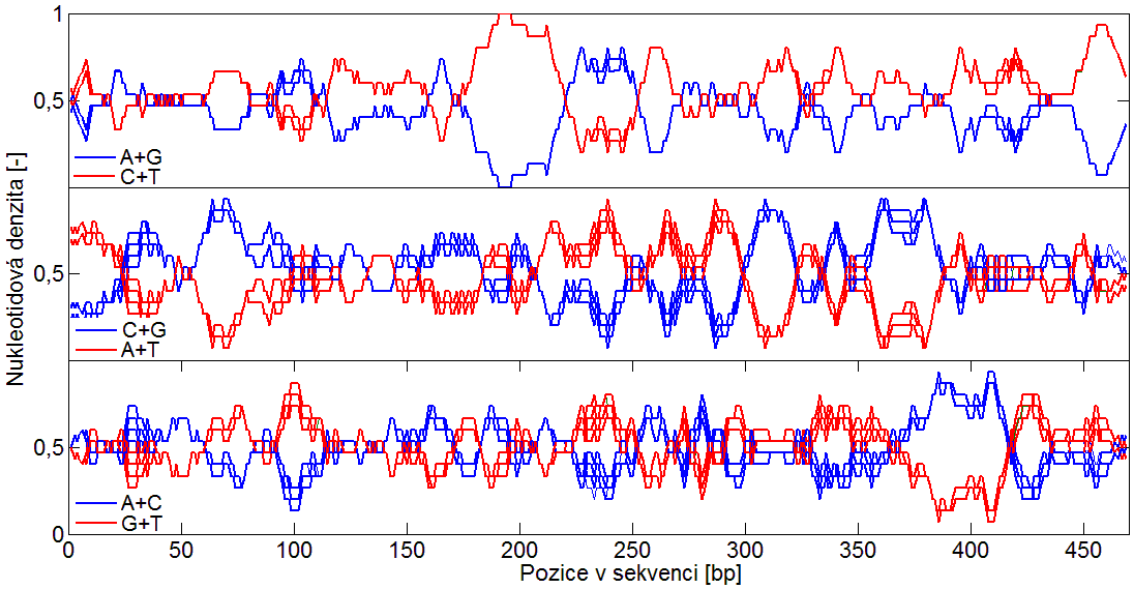
v určitých částech sekvence konzervovanější. Tuto informaci můžeme dostat buď další výpočetní analýzou zarovnaných sekvencí, nebo vykreslením ND vektorů, viz **Obr. 5.15**, kde jsou ND vektory jednotlivých nukleotidů pro všechny sekvence vykresleny do jednoho obrázku. Tam, kde se všechny průběhy překrývají, je daný nukleotid v sekvenci pro všechny jedince konzervovaný. Např. adenin je v přibližné oblasti 120 až 220 konzervovaný. Při vykreslení sum ND vektorů dle biochemických vlastností dostaneme další informaci, které typy nukleotidů jsou více konzervované než jiné. Z **Obr. 5.16** je patrné, že obsah purinových/pyrimidinových nukleotidů je podstatně více konzervován než obsah nukleotidů tvořících slabou/silnou vazbu či amino/keto nukleotidy. Větší konzervovanost purinových/pyrimidinových nukleotidů v souboru podobných sekvencí souvisí s rozdílnou mírou tranzicí a transverzí v rámci evoluce (viz kapitola 2.1.3).



Obr. 5.14 Vícenásobné zarovnání části mitochondriálního genu *coxI* pro 52 různých jedinců druhu *Etheostoma proeliare*. Řádky odpovídají sekvencím a sloupce znázorňují barevně kódované nukleotidy.

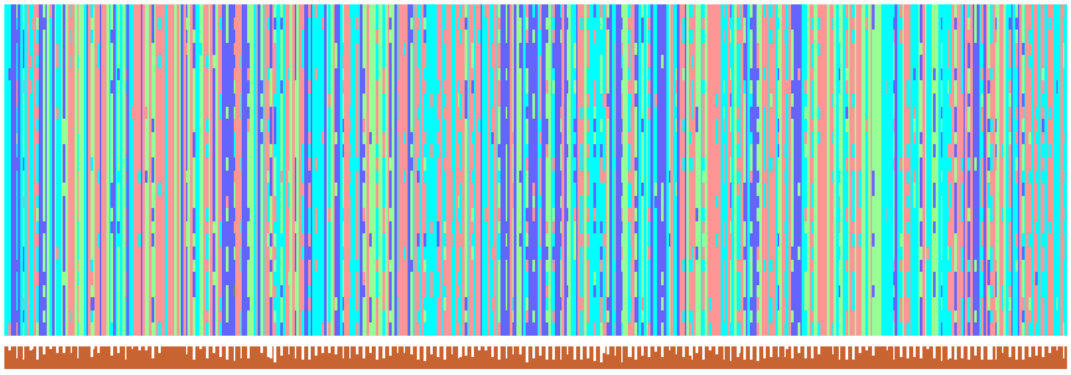


Obr. 5.15 Nukleotidové denzitní vektory části mitochondriálního genu *coxI* pro 52 různých jedinců druhu *Etheostoma proeliare*, zeleně je ohraničena oblast konzervovaného obsahu adeninu, W=15.

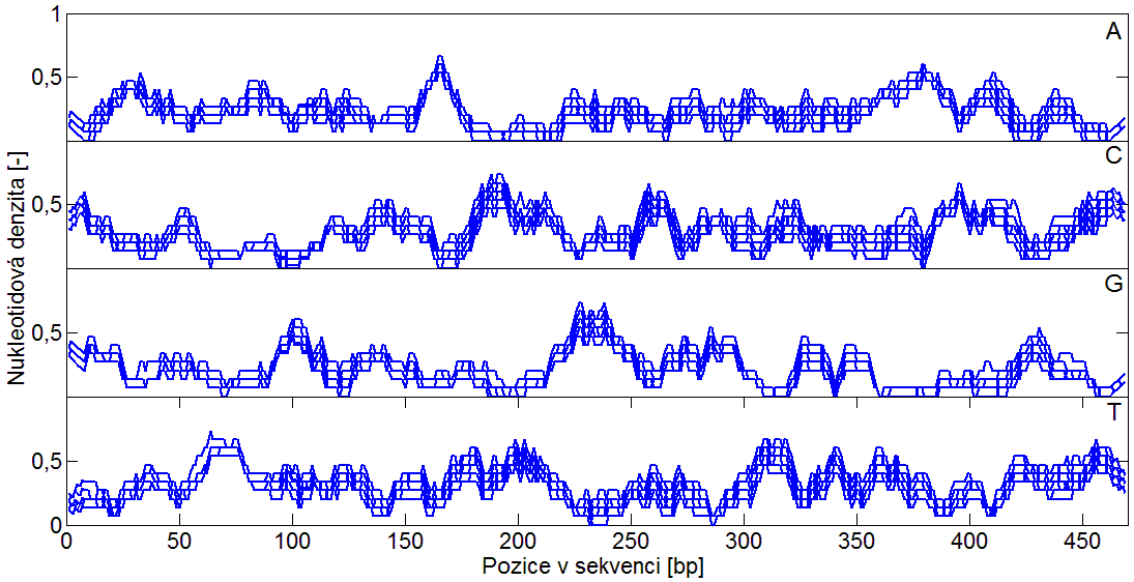


Obr. 5.16 Sumy nukleotidových denzitních vektorů části mitochondriálního genu *coxI* pro 52 různých jedinců druhu *Etheostoma proeliare*, *W*=15.

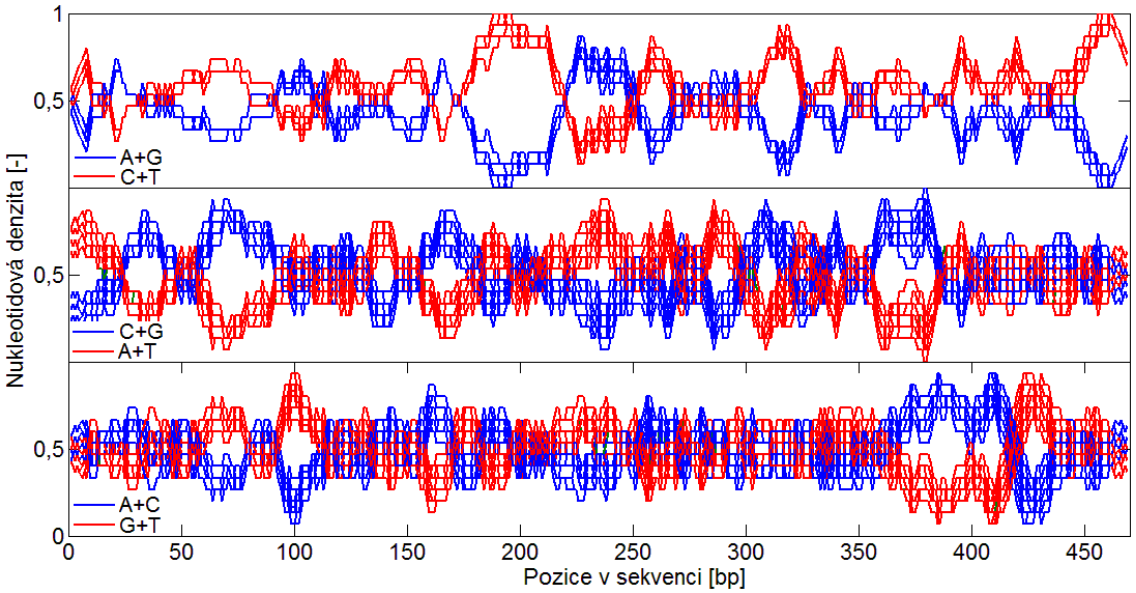
Obdobný případ pro 26 sekvencí různých druhů rodu *Etheostoma* části genu *coxI* představují **Obr. 5.17**, **Obr. 5.18** a **Obr. 5.19**. Míra konzervovanosti sekvencí je podstatně menší, ale ND vektory mají zachovaný trend.



Obr. 5.17 Vícenásobné zarovnání části mitochondriálního genu *coxI* pro 26 různých druhů rodu *Etheostoma*. Řádky odpovídají sekvencím a sloupce znázorňují barevně kódované nukleotidy.

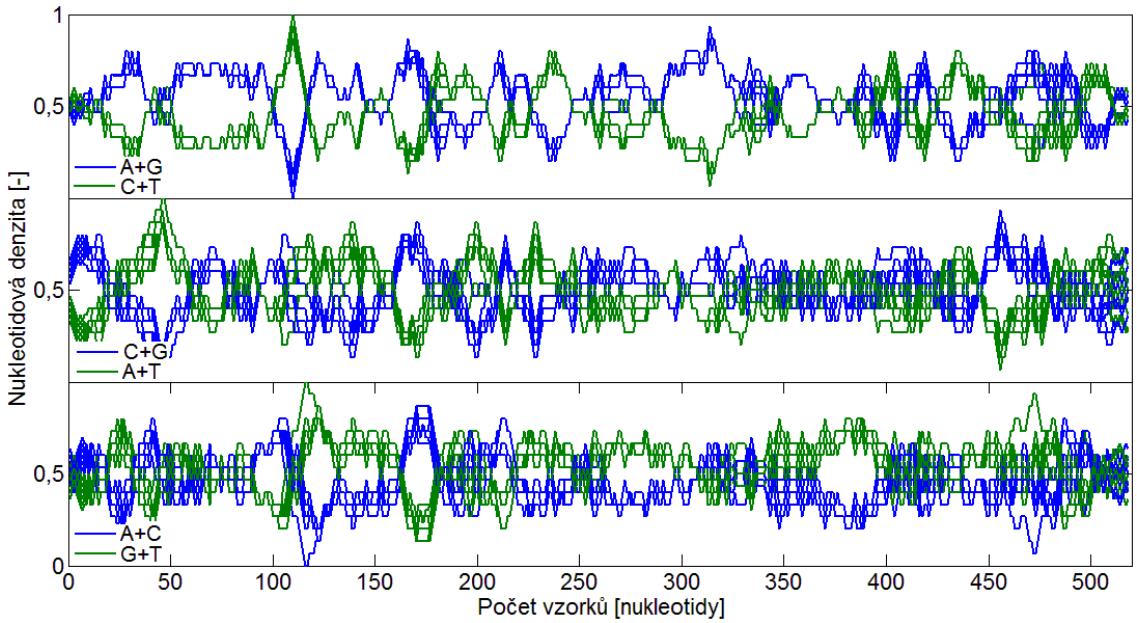


Obr. 5.18 Nukleotidové denzitní vektory části mitochondriálního genu *coxI* pro 26 různých druhů rodu *Etheostoma*, *W*=15.

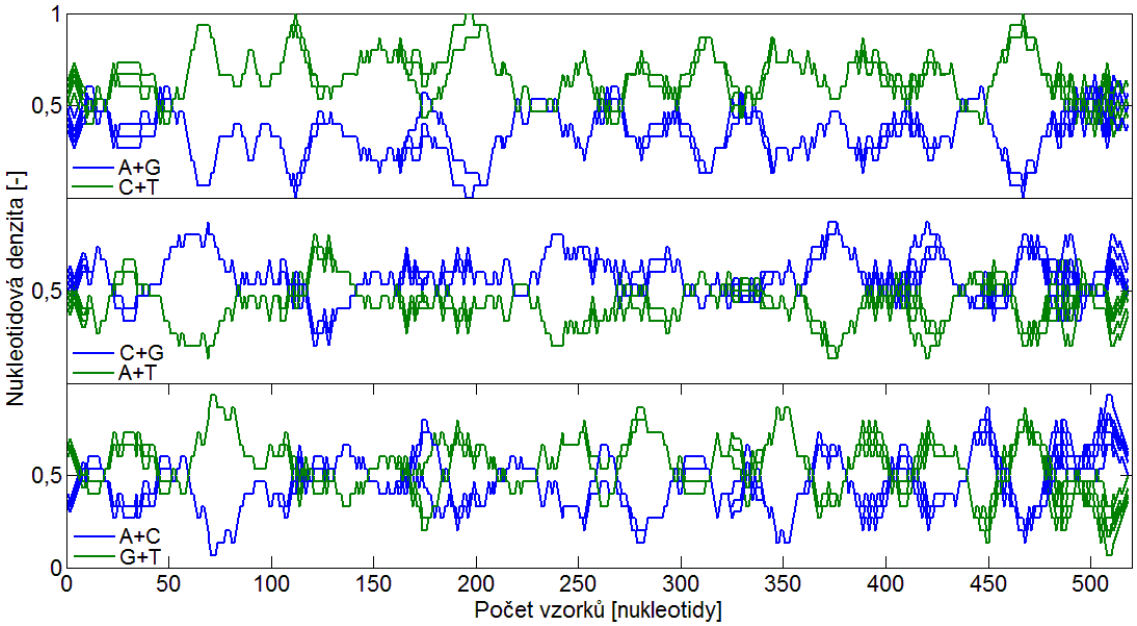


Obr. 5.19 Sumy nukleotidových denzitních vektorů části mitochondriálního genu *coxI* pro 26 různých druhů rodu *Etheostoma*, *W*=15.

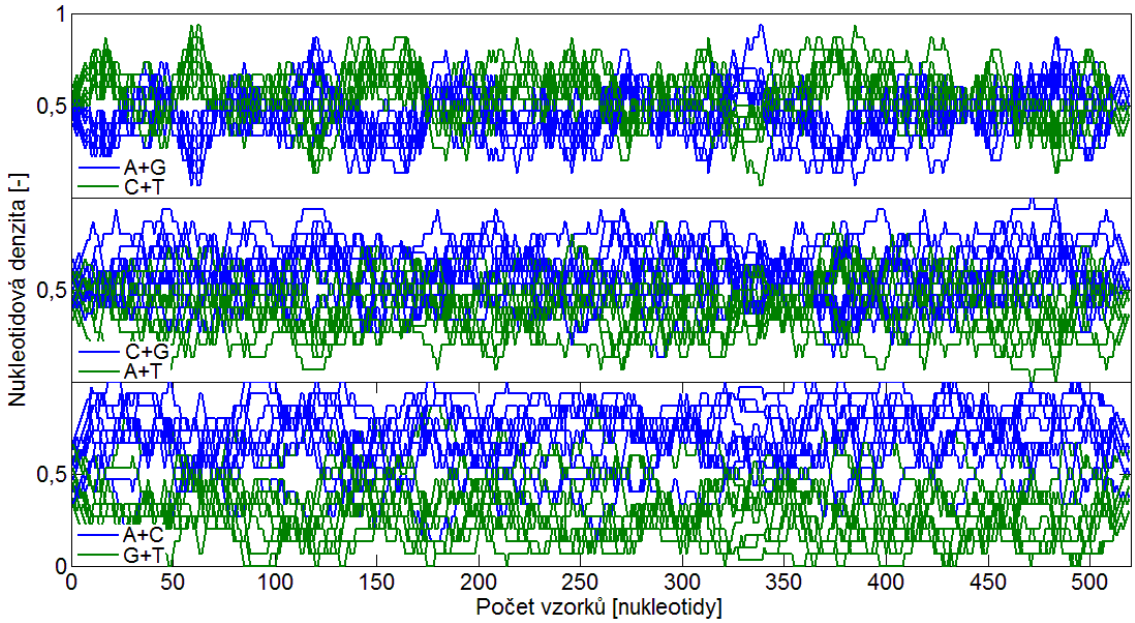
V ND vektorech je také dobře patrná konzervativnost nukleotidů v rámci kodonů (trojice nukleotidů kódujících aminokyselinu). Následující **Obr. 5.20**, **Obr. 5.21** a **Obr. 5.22** zobrazují ND vektory pouze z prvních, druhých a třetích nukleotidů kodonů ze sekvencí celého mitochondriálního genu *coxI* 10 různých obratlovců. Nejvíce mutací se projevuje v třetím nukleotidu kodonu, čemuž odpovídá i vysoká variabilita ND vektorů, jak je zřetelně vidět na **Obr. 5.22**. V případě genu *coxI* je nejméně variabilní druhý nukleotid v kodonu, viz **Obr. 5.21**.



Obr. 5.20 Sumy nukleotidových denzitních vektorů prvních nukleotidů v kodonech 10 sekvencí obratlovců celého mitochondriálního genu *coxI*, *W*=15.

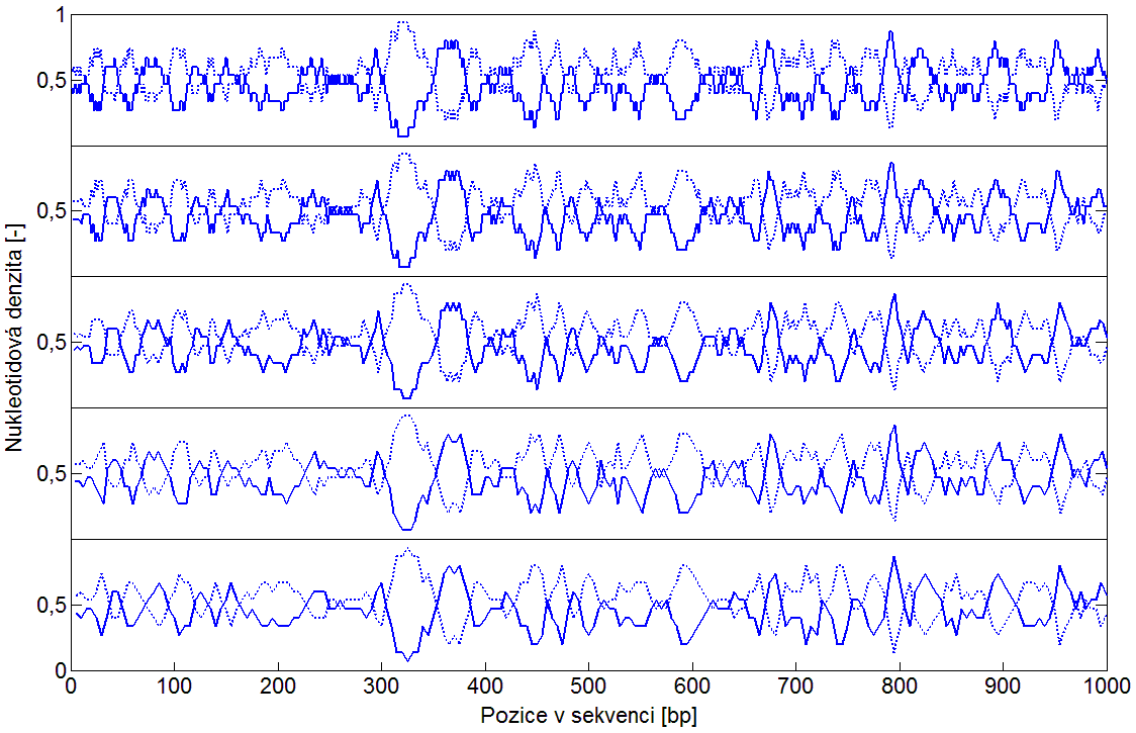


Obr. 5.21 Sumy nukleotidových denzitních vektorů druhých nukleotidů v kodonech 10 sekvencí obratlovců celého mitochondriálního genu *coxI*, *W*=15.

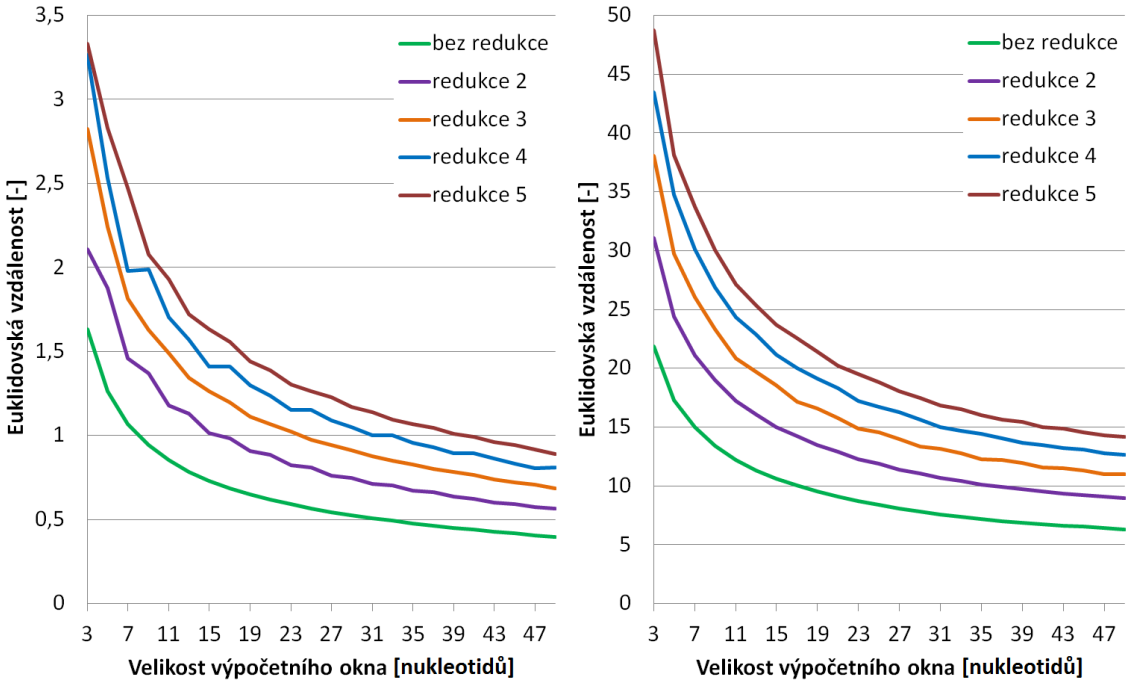


Obr. 5.22 Sumy nukleotidových denzitních vektorů třetích nukleotidů v kodonech 10 sekvencí obratlovců celého mitochondriálního genu *coxI*, *W*=15.

Také byla provedena analýza vlivu redukce počtu vzorků ND vektorů na Euklidovskou vzdálenost mezi ND vektory dvou sekvencí. Snížení počtu vzorků je výhodné při porovnávání velmi dlouhých sekvencí, např. celých mitochondrií. Redukce byla provedena pouhým vynecháním vzorků, např. redukce 2 znamená vynechání každého druhého vzorku, redukce 3 znamená zachování každého třetího vzorku apod. K testování bylo použito prvních 1000 nukleotidů mitochondriálního genu *coxI* u organismů *Canis l. familiaris* a *Canis l. campestris*, které se liší jen o 4 nukleotidy. A dále pak prvních 1000 nukleotidů z genů *coxI* a *nd2* organismu *Canis l. campestris*, přičemž tyto sekvence jsou naprosto odlišné. **Obr. 5.23** zobrazuje sumy ND vektorů jen purinových/pyrimidinových nukleotidů sekvence *coxI* organismu *Canis l. campestris*, při různé volbě redukce vzorků. Je patrná postupná ztráta podrobností se zvyšující se hodnotou redukce, ale základní charakter ND vektorů zůstává zachován. **Obr. 5.24** pak zobrazuje vliv redukce počtu vzorků a volba velikosti výpočetního okna na Euklidovské vzdálenosti mezi ND vektory dvou velmi podobných a dvou naprosto odlišných sekvencí. Trend poklesu vzdálenosti mezi sekvencemi s rostoucím oknem zůstal zachován, ale v rámci stejného výpočetního okna distance roste s nárůstem hodnoty redukce vzorků zhruba lineárně. Trend distancí je stejný pro podobné i odlišné sekvence, ale hodnoty distancí pro odlišné sekvence jsou zhruba o jeden řád vyšší.



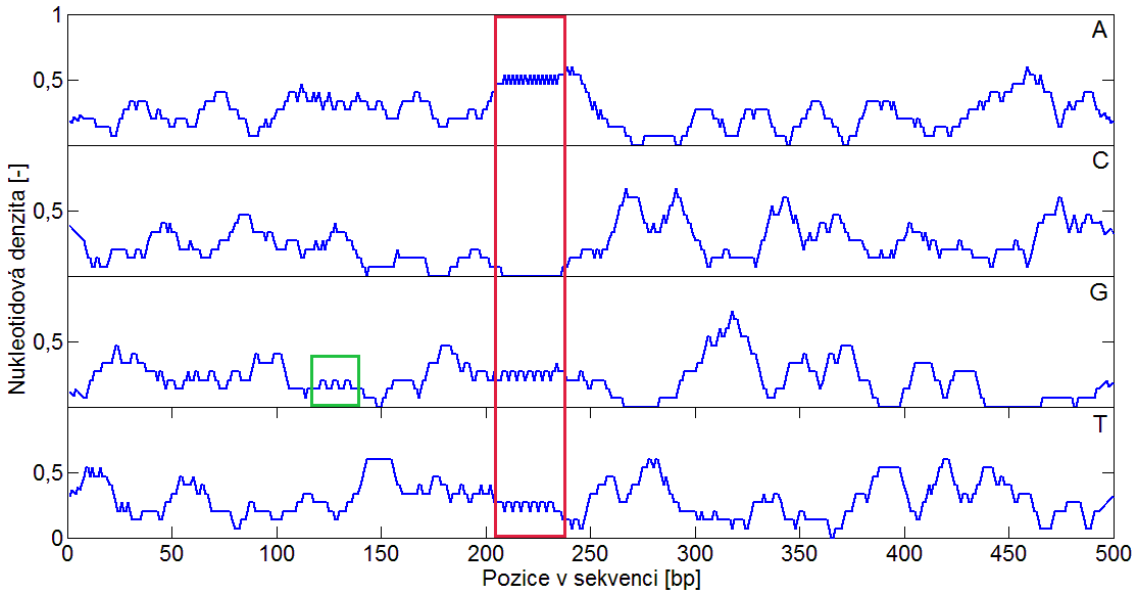
Obr. 5.23 Srovnání nukleotidových denzitních vektorů puriny/pyrimidiny při redukci vzorků pro prvních 1000 bází mitochondriálního genu *coxI* organismu *Canis lupus campestris*. Shora: originál s 1000 vzorky, 500 vzorků (redukce 2), 332 vzorků (redukce 3), 250 vzorků (redukce 4) a 200 vzorků (redukce 5).



Obr. 5.24 Vliv redukce vzorků na Euklidovskou vzdálenost mezi sekvencemi. Vlevo pro sekvence genu *coxI* dvou příbuzných organismů *Canis l . familiaris* a *Canis l. campestris*; vpravo sekvence dvou různých genů *coxI* a *nd2* organismu *Canis l. campestris*.

Dále byla otestována další možnost redukce vzorků pomocí průměrování zvoleného počtu po sobě jdoucích hodnot. Distance mezi takto upravenými ND vektory se jen málo liší od hodnot distancí mezi denzitami redukovanými vynecháváním vzorků.

V nukleotidových denzitních vektorech se také projevují tandemové repetice, což jsou zřetězené stejné úseky DNA. Na **Obr. 5.25** je příklad tandemové repetice TAGA opakující se 10 krát za sebou. Ačkoliv byly ND vektory sekvence s repeticí vytvořeny pro výpočetní okno *W*=15 nukleotidů a repetice má délku jen 4 nukleotidy, je repetice v ND vektorech dobře patrná jako „zoubkovaný“ průběh. Z průběhu vektorů můžeme částečně i odhadnout složení repetice. Průběh pro cytosin je ve vyznačené oblasti nulový, takže repetice cytosin neobsahuje. Guanin a tymin mají stejnou šířku „zoubků“, což značí, že obsah těchto nukleotidů je stejný a střídají se. Adenin má „zoubky“ dvakrát hustší než guanin a tymin, tudíž jeho obsah je v repetici dvojnásobný. Délku motivu repetice nelze z pouhé vizuální inspekce ND vektorů odhadnout; k přesnému určení repetice je nutná signálová analýza průběhů. V obrázku je dále patrná zeleně vyznačená oblast pro guanin, kde se tento nukleotid vyskytuje opakovaně čtyři krát za sebou s periodou opakování 6.



Obr. 5.25 Nukleotidové denzity sekvence s uměle vytvořenou repeticí 10*TAGA na pozici 200 – 240 bp (červeně), jednotlivé G opakující se s periodou 6 čtyřikrát (zeleně), *W* = 15.

5.3 IDENTIFIKACE S VYUŽITÍM ND VEKTORŮ

V této dizertační práci jsou nukleotidové denzitní vektory využity ke komparativní genomice v rámci identifikace do taxonomických skupin na základě DNA barcodingových sekvencí. Základní schéma identifikace porovnáváním nukleotidových denzitních vektorů je následující:

1) Shromáždí se dostatečný počet sekvencí jednoho druhu tak, aby byla podchycena vnitrodruhová variabilita druhu, jak z hlediska pohlaví jedinců, tak věku jedinců a geografie populací.

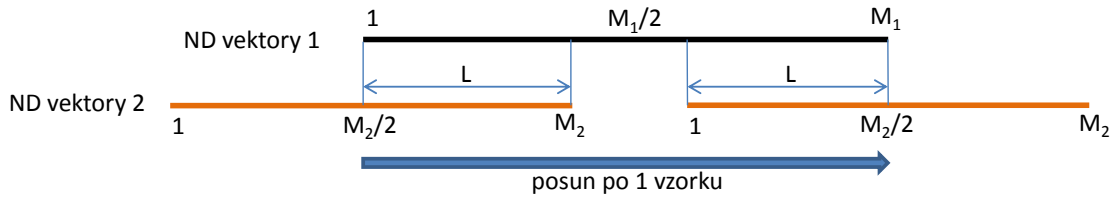
2) Vytvoří se referenční nukleotidové denzity pro jednotlivé druhy.

3) Analyzovaná sekvence se přiřadí k tomu druhu, se kterým má nejpodobnější (nejbližší) nukleotidové denzitní vektory.

Shromáždění dostatečného počtu vhodně vybraných sekvencí je úkol laboratorního charakteru, kdy je nutné získat organické vzorky a provést sekvenaci DNA. Předložená práce se však nezabývá přípravou vzorků; proto byly k analýzám použity sekvence z databáze BOLD.

Metoda k vytvoření referenčních nukleotidových denzitních vektorů konkrétního druhu musí vhodným způsobem zohledňovat vnitrodruhovou variabilitu, tzn. jednotlivé bodové mutace v sekvencích jedinců a jejich frekvenci výskytu v celé populaci. Byly navrženy tři metody založené na nukleotidových denzitních vektorech, které byly otestovány, a byla vyhodnocena jejich vhodnost na velkém souboru reálných sekvencí: metoda mediánu ND vektorů, metoda průměru ND vektorů a metoda průměru ND vektorů jen purinových nukleotidů.

Při výpočtu referenčních ND vektorů je potřeba ND vektory různých sekvencí signálově zarovnat. Pro signálové zarovnání ND vektorů dvou sekvencí se postupuje následovně. ND vektory delší sekvence označme jako ND_1 a mají délku M_1 a ND vektory kratší sekvence označme jako ND_2 a mají délku M_2. V případě stejných délek na pořadí sekvencí nezáleží. **Obr. 5.26** zobrazuje schematicky počáteční a koncové zarovnání obou ND vektorů.



Obr. 5.26 Schéma signálového zarovnávání ND vektorů dvou sekvencí. Zobrazeno je počáteční a koncové vzájemné umístění vektorů.

ND_2 se porovnává vůči ND_1. První okrajová pozice ND_2 vůči ND_1 je mezi vzorky 1 až $M_2/2$ z ND_1 a vzorky $M_2/2$ až M_2 z ND_2 a těchto vzorků je $L=M_2/2$. Pro každou pozici se spočítá Euklidovská vzdálenost mezi částmi ND vektorů dle vzorce (5.2). Následně se ND_2 posune vůči ND_1 o jeden vzorek, takže se porovnávají vzorky 1 až $M_2/2+1$ z ND_1 a vzorky $M_2/2-1$ až M_2 z ND_2 a těchto vzorků je $L=M_2/2+1$. Při každém posunu se zvětšuje porovnávaná oblast ND_2. Tento posun o jeden vzorek s výpočtem END vzdálenosti pokračuje do pozice porovnávání vzorků 1 až M_2 z ND_1 a vzorky 1 až M_2 z ND_2 a těchto vzorků je $L=M_2$. Nyní se již porovnává celý úsek ND_2 s ND_1. Pokračuje se v posunu do pozice M_1-M_2+1 až M_1 z ND_1 a vzorky 1 až M_2 z ND_2 a těchto vzorků je

$L=M_2$. Toto je poslední pozice, kdy se porovnává celé ND_2, následně se opět oblast porovnávání zmenšuje až k pozici vzorků $M_1–M_2/2+1$ až M_1 z ND_1 a vzorky 1 až $M_2/2$ z ND_2 a porovnávaných vzorků je $L=M_2/2$. Nejlepší vzájemná pozice ND vektorů je ta s minimální hodnotou Euklidovské vzdálenosti.

5.3.1 METODA MEDIÁNU ND VEKTORŮ

Pro každou sekvenci jedince daného druhu jsou vypočítány nukleotidové denzitní vektory. Nukleotidové denzitní vektory všech jedinců druhu (nebo vyšší taxonomické skupiny) jsou signálově zarovnány tak, že se nalezne jejich vzájemná pozice, která minimalizuje Euklidovskou vzdálenost mezi ND vektory, viz rovnice (5.2). Pro signálové zarovnání lze použít i jiných přístupů, např. pomocí korelace. Výsledné referenční ND vektory jsou tvořeny mediánovými hodnotami zarovnaných ND vektorů. **Tab. 5.3** zobrazuje příklad tvorby mediánové referenčního ND vektoru d_A. Kompletní reference je tvořena čtyřmi mediánovými ND vektory d_A, d_C, d_G a d_T.

Tab. 5.3 Příklad tvorby mediánového referenčního ND vektoru d_A.

Zarovnané vektory											
1. d_A	0,3	0,2	0,3	0,4	0,3	0,2	0	0,1	0,3		
2. d_A			0,3	0,3	0,3	0,2	0,1	0,2	0,3	0,1	
3. d_A		0,2	0,3	0,2	0,3	0,2	0	0,1			
4. d_A			0,4	0,4	0,3	0,2	0,1	0,2	0,3	0,1	0,2
5. d_A	0,3	0,2	0,3	0,4	0,3	0,1	0,1				
Mediánový referenční d_A	0,3	0,2	0,3	0,4	0,3	0,2	0,1	0,1	0,3	0,1	0,2

5.3.2 METODA PRŮMĚRU ND VEKTORŮ

Pro každou sekvenci jedince daného druhu jsou vypočítány nukleotidové denzitní vektory. Nukleotidové denzitní vektory všech jedinců jsou zarovnány tak, že se nalezne jejich vzájemná pozice, která minimalizuje Euklidovskou vzdálenost dle rovnice (5.2) mezi ND vektory. Výsledné referenční ND vektory jsou tvořeny průměrnými hodnotami zarovnaných ND vektorů. **Tab. 5.4** zobrazuje příklad tvorby průměrného referenčního ND vektoru d_A. Kompletní reference je tvořena čtyřmi průměrnými ND vektory d_A, d_C, d_G a d_T.

Tab. 5.4 Příklad tvorby průměrného referenčního ND vektoru d_A.

Zarovnané vektory											
1. d_A	0,3	0,2	0,3	0,4	0,3	0,2	0	0,1	0,3		
2. d_A			0,3	0,3	0,3	0,2	0,1	0,2	0,3	0,1	
3. d_A		0,2	0,3	0,2	0,3	0,2	0	0,1			
4. d_A			0,4	0,4	0,3	0,2	0,1	0,2	0,3	0,1	0,2
5. d_A	0,3	0,2	0,3	0,4	0,3	0,1	0,1				
Průměrný referenční d_A	0,3	0,2	0,32	0,34	0,3	0,18	0,06	0,15	0,3	0,1	0,2

5.3.3 METODA PRŮMĚRU PURINOVÝCH ND VEKTORŮ

Referenční ND vektory pro druh se tvoří obdobně jako v předchozí metodě průměru nukleotidových denzitních vektorů. Po výpočtu čtyř průměrných vektorů pro každý druh nukleotidu se však jako reference bere pouze součet průměrných vektorů pro adenin a guanin. Při identifikaci se pak využívá výpočet Euklidovské vzdálenosti dle vzorce (5.3).

5.3.4 METODA IDENTIFIKACE

Pokud je již vytvořená databáze referenčních nukleotidových denzitních vektorů, můžeme neznámé sekvence identifikovat, tj. přiřadit k některému z referenčních druhů. Byla navržena metoda založená na porovnávání nukleotidových denzitních vektorů neznámé sekvence s referenčními ND vektory. Ke kvantifikaci podobnosti/nepodobnosti dvou setů ND vektorů, tj. ND vektorů dvou sekvencí či ND vektorů neznámé sekvence s referenčními ND vektory, byl vybrán výpočet Euklidovské vzdálenosti mezi ND vektory. Tato Euklidovská vzdálenost slouží také k signálovému zarovnání ND vektorů, tj. při hledání vzájemné pozice vektorů s minimální Euklidovskou vzdáleností.

Pro identifikaci neznámé sekvence se postupuje následovně. Pro neznámou sekvenci se vypočítají nukleotidové denzitní vektory a následně se tyto vektory porovnávají se všemi referenčními ND vektory pro druhy či jiné taxonomické skupiny. Porovnávání funguje tak, že se ND vektory neznámé sekvence signálově zarovnají se všemi referenčními ND vektory. Signálové zarovnání znamená, že se nalezne vzájemná pozice ND vektorů taková, která dává nejmenší hodnotu délkově normalizované Euklidovské distance na 1000 nukleotidů, zde označovaná jako END vzdálenost (Euklidovská vzdálenost ND vektorů):

$$END_{i,j} = \frac{1}{L}\sqrt{\sum_{n=1}^{L}\left(\sum_{x=d_A,d_C,d_G,d_T}\left(x_{i,n}-x_{j,n}\right)^2\right)} * 10^3 \qquad (5.2)$$

kde i je index ND vektorů neznámé sekvence; j je index referenčních ND vektorů; d_A, d_C, d_G a d_T jsou ND vektory příslušných nukleotidů a L je délka porovnávaného úseku ND vektorů. Neznámá sekvence je přiřazena k tomu druhu, s jehož referencí má nejmenší hodnotu END vzdálenosti.

Z předběžných analýz ND vektorů pro sekvence ptáků vyšlo najevo, že obsah a pozice guaninu jsou značně konzervovány a s menší mírou to platí i pro adenin [129]. Adenin a guanin jsou purinové nukleotidy. Konzervovanost purinových nukleotidů může být klíčová pro přiřazování analyzovaných sekvencí k vyšším taxonomickým jednotkám než je druh. Proto byla vedle identifikace na základě Euklidovských vzdáleností mezi ND vektory všech nukleotidů navrhnuta ještě identifikace porovnáváním ND vektorů jen purinových nukleotidů. ND vektory pro adenin a guanin jsou sečteny a následně se počítá Euklidovská vzdálenosti těchto sumovaných vektorů

pro neznámou sekvenci a referenční ND vektory. Tuto vzdálenost označujme jako PEND (Purinová Euklidovská vzdálenost ND vektorů) a počítá se dle vzorce:

$$PEND_{i,j} = \frac{1}{L}\sqrt{\sum_{n=1}^{L}\left(\left(a_{i,n} + g_{i,n}\right) - \left(a_{j,n} + g_{j,n}\right)\right)^2} * 10^3 \qquad (5.3)$$

kde i je index ND vektorů neznámé sekvence; j je index referenčních ND vektorů; a se rovná d_A a g je d_G; L je délka porovnávaného úseku ND vektorů.

6 DRUHOVÁ IDENTIFIKACE

Databáze BOLD poskytuje DNA barcodingové sekvence velkého množství různých typů organismů. Sekvence z databáze nelze cíleně vybírat a získávat podle konkrétního druhu. Sekvence lze získat pouze ve formě souhrnného FASTA souboru zvoleného barcodingového projektu. Projekty jsou většinou zaměřené na jednu skupinu organismů nebo na geografickou lokalitu. V jednom souboru se pak vyskytují sekvence většího množství druhů a jejich jedinců.

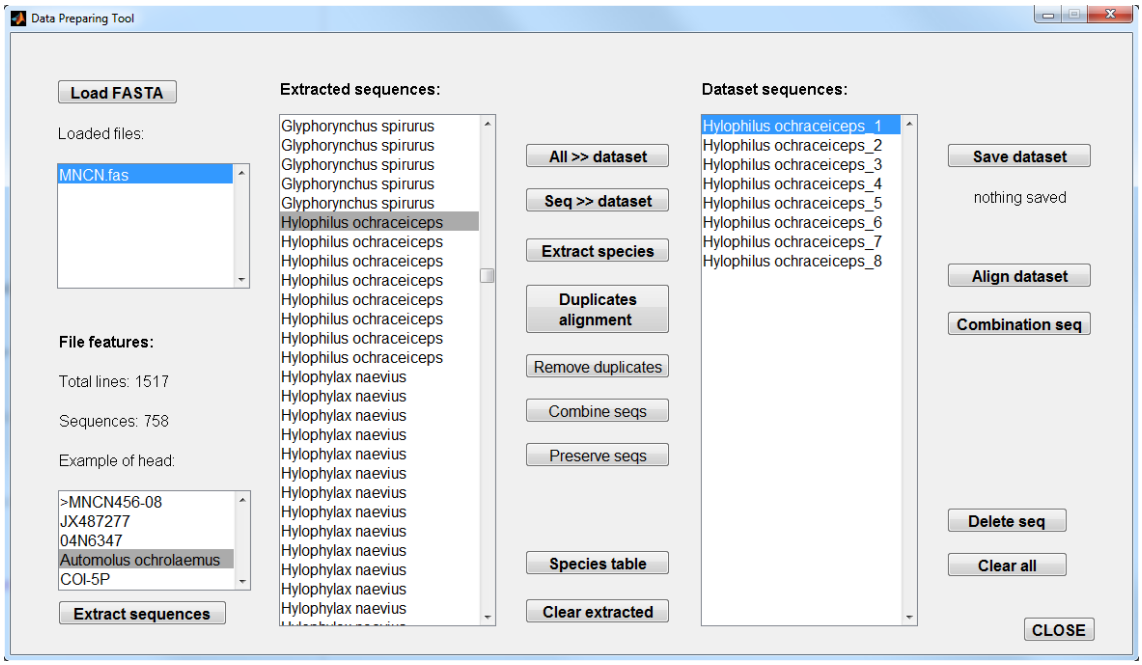
Z databáze GenBank (pod správou organizace NCBI) také není možné jednoduchým automatizovaným způsobem sekvence získat. Každý druh, který má být zahrnut do analýzy, by se v databázi musel vyhledávat samostatně, v nalezených sekvencích by se musely ručně vybrat úseky odpovídající *coxI*, a teprve poté je možné vybrané sekvence získat v jednom FASTA souboru. To by znamenalo nezvládnutelné množství stereotypní práce, proto byla dána přednost získávání sekvencí z BOLD databáze ve formě projektů. V poslední době se v BOLD databázi objevují i projekty obsahující extrahovaná data z GenBank databáze.

FASTA soubor DNA barcodingového projektu obsahuje sekvence řazené za sebou a oddělené popisnými hlavičkami jednotlivých sekvencí. Příklad takového FASTA souboru:

```
>BROMB378-06|JN801479|A.aburriB6|Aburria aburri|COI-5P
GTGACCTTCATCAATCGATGATTATTCTCAACTAACCATAAAGACATCGGCACCCTCTACTTAATTT…
>LGEMA345-08|JN801480|LGEMA-363|Accipiter bicolor|COI-5P
GTGACATTTATTAATCGATGAATATTCTCAACCAACCACAAAGACATTGGTACTCTTTACCTAATCT…
>BROMB692-07|JN801481|DOT 12780|Agamia agami|COI-5P
GTGACCTTCATTACCCGATGATTATTCTCAACCAACCACAAAGATATCGGCACCCTATACCTAATCT…
```

Název druhu, kterému sekvence přísluší, je uveden v hlavičce. Bohužel se ne vždy dodržuje pořadí jednotlivých částí hlavičky oddělených znakem „|“ ani v rámci jednoho souboru. Tento fakt komplikuje automatickou extrakci sekvencí ze souboru podle příslušnosti k druhu.

Pro získání sekvencí z projektových souborů z databáze BOLD byla použita k tomuto účelu vytvořená vlastní uživatelská aplikace DPT (Data Preparing Tool) v prostředí Matlab. Obrázek uživatelské aplikace je na **Obr. 6.1**. Tato aplikace načítá FASTA soubory se sekvencemi. Podle zobrazené hlavičky první sekvence se vybírá pozice názvu druhu v hlavičce a následně dochází k extrakci sekvencí a znázornění jmen druhů pro každou sekvenci v souboru. Pokud není dodržena stejná pozice názvu druhu v hlavičce u všech sekvencí, pak je nutná ruční korekce výběru názvu pro danou sekvenci. Po extrakci sekvencí ze souboru je možné sekvence vybírat ručně a ukládat je do samostatného souboru, nebo se sekvence automaticky roztřídí podle druhů a uloží se do separátních souborů pro každý druh. Separátní soubory pro druhy pak slouží k automatické tvorbě druhových referencí. Aplikace také umožňuje vícenásobné zarovnávání sekvencí a export názvů druhů do souboru v `xls` formátu.



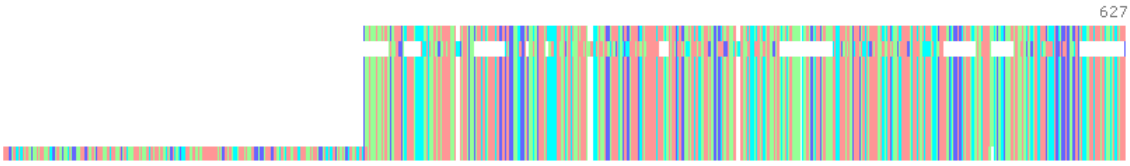
Obr. 6.1 Panel uživatelské aplikace DPT v Matlabu pro extrakci sekvencí z projektových FASTA souborů z databáze BOLD.

6.1 REFERENČNÍ DRUHY

Metody tvorby referenčních nukleotidových denzitních vektorů a metody identifikace byly otestovány na sekvencích živočišných druhů různých skupin organismů. Pro každou skupinu organismů existuje v databázi BOLD množství dat, jen obojživelníci a savci jsou zastoupeni méně, proto byl výběr dat z této oblasti značně omezený. Projektové soubory sekvencí byly vybrány takové, aby obsahovaly alespoň desítku rozdílných druhů. U některých skupin organismů bylo možné mít jeden samostatný soubor pro vytváření referencí a jiný soubor, který obsahoval sekvence stejných druhů ale jiných jedinců, mít pouze k identifikaci. Ve většině případů však neexistovaly dva podobné soubory, bylo proto nutné najít takové soubory, kde bylo ke každému druhu dostatek jedinců, a tyto jedince pak rozdělit na poloviny, kde první část sloužila k vytvoření referencí a druhá k nezávislé identifikaci.

Stažené FASTA soubory sekvencí bylo nejprve nutné zpracovat aplikací DPT. Automaticky byly ze souborů extrahovány sekvence jednotlivých druhů. Pro každý druh se vytvořil samostatný soubor ve formátu `mat`, který obsahoval sekvence všech jedinců daného druhu. Pro vytváření referencí se však následně vybraly pouze ty druhy, které měly alespoň 3 sekvence jedinců. V případě jednoho souboru společného pro vytváření referencí i identifikaci byly použity pouze ty druhy, které měly v souboru alespoň 4 sekvence jedinců, kdy 3 byly určeny k tvorbě reference a 1 pro identifikaci. V případě většího počtu sekvencí jednoho druhu, pak polovina sloužila k vytvoření referencí a druhá polovina k identifikaci.

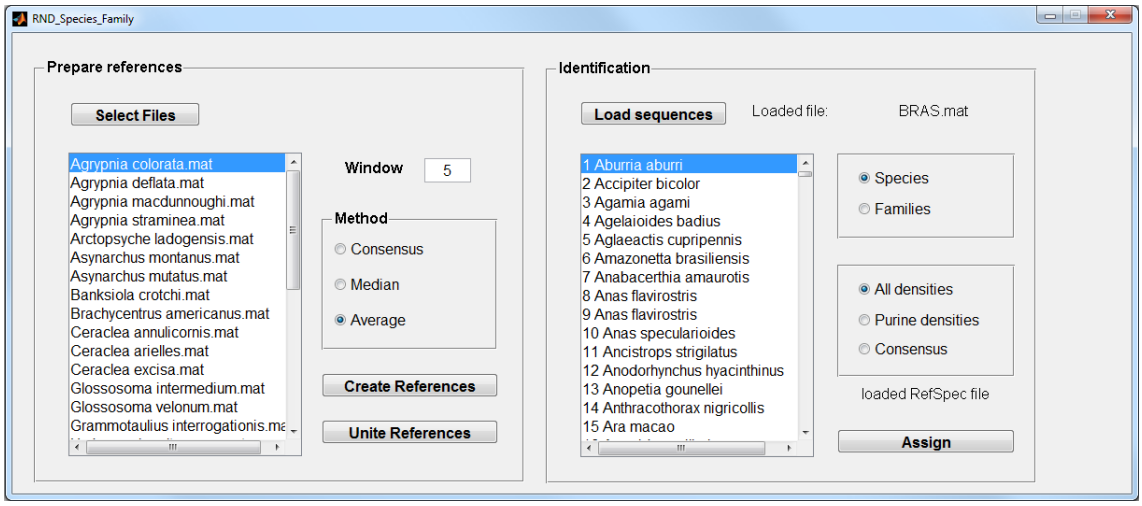
Ačkoliv je celá extrakce sekvencí druhů automatická, je vhodné sekvence každého druhu zkontrolovat zobrazením jejich vícenásobného zarovnání, neboť se občas vyskytly případy, kdy byly do databáze vloženy sekvence jiného úseku než *coxI*, či sekvence velmi krátké nebo patřící pravděpodobně jinému druhu, protože množství odlišných bází překračovalo obvyklou mez vnitrodruhové variability (> 2 %). Takováto sekvence by pak negativně ovlivnila charakter vytvořených referenčních ND vektorů. Kontrolu vícenásobným zarovnání lze provést aplikací DPT. Přítomnost o něco kratších či delších sekvencí v souboru sekvencí jedinců druhu není na závadu. Vytvořené algoritmy pro tvorbu referencí toto akceptují a druhová reference pak má délku jako nejdelší sekvence mezi jedinci. **Obr. 6.2** znázorňuje příklad, kdy v souboru sekvencí pro jeden druh je 1 sekvence naprosto rozdílná, tudíž je potřeba ji ze souboru odstranit, a 1 sekvence je pouze delší ale jinak shodná s ostatními, a proto je ponechána.



Obr. 6.2 Grafické znázornění vícenásobného zarovnání sekvencí jednoho druhu, kde je 1 sekvence naprosto odlišná (musí se odstranit) a 1 je pouze delší (ponechá se).

Větší část vybraných druhů organismů v projektových souborech byla zastoupena malým počtem jedinců, tudíž nebylo možné dobře specifikovat jejich vnitrodruhovou variabilitu. V drtivé většině zastoupených druhů byly vnitrodruhové rozdíly v sekvencích i v případě většího počtu jedinců pouze v několika málo nukleotidech. To by vedlo k žádným nebo velmi malým rozdílům mezi referencemi vytvořenými metodou mediánu ND vektorů a metodou průměru ND vektorů. Z tohoto důvodu byla identifikace druhů provedena pouze metodou průměru ND vektorů a průměru ND vektorů jen purinových nukleotidů. Rozdíl mezi všemi metodami je patrný až na výsledcích pro identifikaci do čeledí, viz kapitola 7.

Pro vytváření referencí druhů a následnou identifikaci byla v prostředí Matlab vytvořena uživatelská aplikace RND (Reference Nucleotide Density), viz obrázek **Obr. 6.3**. K vytvoření druhových referencí je nutné nejprve načíst samostatné druhové soubory dříve vytvořené v aplikaci DPT. Následně se zadává velikost výpočetního okna pro výpočet nukleotidových denzitních vektorů a případně se volí metoda tvorby reference (konsenzus, medián nebo průměr). Pro vytvoření druhových referencí je implicitně nastavena metoda průměrováním. Vypočtené druhové reference pro všechny načtené druhové soubory se uloží do jednoho souboru ve formátu `mat`, v jehož názvu je obsažena velikost nastaveného výpočetního okna. Dříve vytvořené soubory s referencemi lze spojovat dohromady a tím je umožněno rozšiřování referenčního souboru o nové druhy.



Obr. 6.3 Panel uživatelské aplikace RND v Matlabu pro vytváření druhových referencí a identifikaci sekvencí.

K tvorbě druhových referencí byly vybrány soubory několika skupin organismů: hmyz, ryby, obojživelníci, ptáci a savci. Soubory sekvencí byly vybírány s ohledem na omezené možnosti databáze a také tak, aby se na nich vyzkoušely některé specifické problémy identifikace jako např. identifikace blízce příbuzných druhů. **Tab. 6.1** obsahuje stručnou specifikaci souborů vybraných pro tvorbu referencí a následnou nezávislou identifikaci. Reference byly vytvořeny pro velikosti výpočetního okna W=3, 5, 7 až 25 nukleotidů. Některé z těchto souborů byly použity pouze k tvorbě referencí (reference byly vytvořeny pro všechny jedince druhu) a některé soubory sloužily současně i k identifikaci (reference byly tvořeny z poloviny jedinců druhu a ostatní byly použity k identifikaci).

Tab. 6.1 Soubory z databáze BOLD použité k tvorbě referencí druhů a následné identifikaci.

Kód projektu	Název projektu	Jedinců	Druhů	Ref. sek.	Ref. druhů	Rodové sek.	Typ organismu	Použito jako
CUCAD	Trichoptera of Churchill 2007	719	55	692	32	16	chrostíci	reference
DSTRI	Trichoptera of Churchill 2006	540	23	245	-	240	chrostíci	identifikace
GBANO	GenBank *Arthropoda-Diptera-Culicidae- Anophelinae*	2093	111	1007+1002	58	84	komáři	reference + identifikace
FFNA	Freshwater Fishes of North America	4313	673	2108+1733	437	343	sladkovodní ryby	reference + identifikace
FCFP	Fishes of Portugal – West Coast 1	188	48	175	38	8	mořské ryby	reference
FCFPS	Fishes of Portugal – South Coast	200	55	124	-	25	mořské ryby	identifikace
FCFPW	Fishes of Portugal – West Coast 2	207	49	130	-	24	mořské ryby	identifikace
GBAP	Genbank - *Amphibia*	1515	477	490+428	74	188	obojživelníci	reference + identifikace
MNCN	Neotropical Birds	758	43	756	42	0	ptáci	reference
BRAS	Neotropical	637	432	85	-	84	ptáci	identifikace
BCBNC	ROM – Bats of Guyana 1	840	96	807	70	22	netopýři	reference
BCDR	ROM – Bats of Guyana 2	4134	68	4051	-	54	netopýři	identifikace

Ref. sek. – počet sekvencí použitých k vytvoření druhových referencí
Rodové sek. – počet sekvencí stejného rodu ale jiného druhu než jsou druhové reference

Z bezobratlých živočichů obsahuje databáze BOLD sekvence především parazitů, chrostíků a měkkýšů, v menší míře mravenců, včel, žížal a dalších. Jelikož je přesné taxonomické zařazení sekvencí parazitů sporné (většinou je v souborech uveden pouze rod), byly vybrány soubory hmyzu, kde jsou sekvence ve větší míře přiřazeny k jednotlivým druhům. Databáze BOLD obsahuje množství sekvencí chrostíků, tudíž byl vybrán soubor se zkratkou CUCAD, který byl celý použit k tvorbě referencí. Existuje další soubor s podobnými sekvencemi, který bylo možné použít k identifikaci, se zkratkou DSTRI. Ze skupiny hmyzu byl dále vybrán soubor GBANO, který obsahuje sekvence komárů jednoho rodu *Anophelinae*. Jelikož k souboru neexistuje žádný podobný, který by se použil k identifikaci, sloužila k tvorbě referencí jen polovina sekvencí jedinců daného druhu, a druhá polovina pak sloužila k identifikaci. Tento soubor je specifický právě v množství druhů a jejich variant patřících do jednoho rodu, tj. obsahuje sekvence blízce příbuzných druhů.

Z ryb byl vybrán k vytvoření referencí jeden soubor pro sladkovodní ryby a jeden pro mořské ryby. Soubor FFNA pro sladkovodní ryby severní Ameriky je největším souborem ryb v databázi. K tvorbě referencí byla použita polovina jedinců druhů a druhá polovina pro identifikaci. Soubor portugalských mořských ryb FCFP byl použit k tvorbě referencí celý, neboť existují ještě dva další soubory FCFPS a FCFPW se sekvencemi stejných druhů, ale jiných jedinců, které se použily pouze k identifikaci.

Databáze BOLD obsahuje pouze okolo 2000 sekvencí obojživelníků, což je v porovnání s ostatními skupinami živočichů velmi málo. Důvodem je uváděná nevhodnost genu *coxI* k identifikaci této skupiny organismů kvůli jejich vysoké vnitrodruhové variabilitě [88]. Většina sekvencí *coxI* obojživelníků v databázi BOLD byla získána extrakcí dat z celých mitochondriálních sekvencí z databáze GenBank. K tvorbě referencí byl tedy vybrán soubor GBAP, který obsahuje právě sekvence extrahované z GenBank databáze. Jelikož soubor obsahuje velké množství druhů a jen malé množství jedinců pro jednotlivé druhy, bylo vytvořeno jen 74 referencí ze 477 druhů v souboru se vyskytujících. Tvorba referencí a identifikace sekvencí z toho souboru měla za cíl ukázat, s jakou mírou úspěšnosti si zde navržená metoda poradí i s údajně problémovými sekvencemi.

DNA barcoding ptáků poskytuje několik tisíc sekvencí ze všech kontinentů. Ač je počet dostupných sekvencí velký, u většiny druhů bylo sekvenováno málo jedinců. Soubory se v obsahu druhů také málo překrývají. K tvorbě referencí se jako vhodný ukázal soubor neotropických jihoamerických ptáků MNCN, k němuž existuje další soubor a větším počtem stejných druhů BRAS. Další soubory ptáků z různých kontinentů pak byly vybrány pro analýzu schopnosti navržených metod přiřazovat sekvence do vyšších taxonomických skupin (čeledě a řády) viz kapitola 7.

Pro savce neobsahuje databáze BOLD mnoho dat. Důvodem nezájmu barcodingových skupin je, že savci jsou již taxonomicky dobře popsáni. Jediná skupina savců, která se nějakému zájmu těší, jsou netopýři. Pro tvorbu referencí byl vybrán

soubor guyanských netopýrů BCBNC. Použiti byli všichni jedinci, jejichž sekvence se v databázi nacházely. Soubor BCDR byl použit k identifikaci.

Pro všechny referenční druhy byly vypočítány průměrné Euklidovské vzdálenosti mezi jejich ND vektory a směrodatné odchylky vzdáleností pro výpočetní okno W=5 nukleotidů. Průměrná vzdálenost mezi ND vektory pro druh byla získána tak, že každá sekvence daného druhu byla převedena na ND vektory a pro každou dvojici ND vektorů sekvencí byla spočítána normalizovaná Euklidovská vzdálenosti na 1000 nukleotidů a tyto vzdálenosti byly zprůměrovány (E_V). Dále byly spočítány vzdálenosti mezi druhy tak, že se porovnávaly všechny dvojice referenčních ND vektorů druhů (průměrné ND vektory pro druh). Vzdálenosti byly zprůměrovány v rámci druhů patřících do stejných rodů (E_R) a pak všechny ostatní, tj. mezidruhové vzdálenosti (E_M). Tyto hodnoty shrnuje **Tab. 6.2**. Průměrné vzdálenosti mezi ND vektory jednoho druhu (vnitrodruhová vzdálenost) jsou několikanásobně nižší než vzdálenosti mezi referenčními druhovými ND vektory (mezidruhová vzdálenost), což znamená, že identifikace porovnáváním ND vektorů by měla mít vysokou úspěšnost. Nejvyšší hodnoty vnitrodruhové vzdálenosti vykazují druhy obojživelníků ze souboru GBAP, a proto lze očekávat problematickou identifikaci. Odstup průměrných vzdáleností pro druhy stejného rodu a ostatní druhy je s přihlédnutím k jejich směrodatným odchylkám podstatně menší než odstup průměrných vzdáleností uvnitř druhu a vzdáleností mezi druhy, ale i tak lze předpokládat, že sekvence druhů, které nebudou mít druhovou referenci, budou při identifikaci přiřazovány s vyšší úspěšností k referenčním druhům stejného rodu, pokud takový bude v databázi.

Tab. 6.2 Průměrné Euklidovské vzdálenosti a jejich směrodatné odchylky pro druhy z referenčních souborů.

Projekt	E_V	σ_V	E_R	σ_R	E_M	σ_M
CUCAD	1,4128	1,1334	8,1728	1,7116	11,3525	1,0311
GBANO	2,5125	1,0929	10,1482	3,9916	-	-
FCFP	1,1221	0,7318	6,7378	2,4042	11,6123	0,7807
FFNA	1,7519	1,5693	8,4390	1,4229	10,9859	1,1394
GBAP	3,5974	3,0632	8,6413	3,0030	13,3193	3,7825
MNCN	2,6232	1,1006	5,8688	1,6662	8,8621	0,7178
BCBNC	1,8561	1,0531	7,0139	1,8550	10,8034	0,6932

E_V – průměrné vzdálenosti mezi ND vektory jednoho druhu,
E_R – průměrné vzdálenosti mezi referenčními ND vektory druhů stejného rodu,
E_M – průměrné vzdálenosti mezi referenčními ND vektory všech druhů mimo stejný rod.

Dva druhy (*Cheumatopsyche campyla* a *Cheumatopsyche ela*) ze souboru CUCAD vykazují nízkou mezidruhovou vzdálenost, která odpovídá spíš vnitrodruhové vzdálenosti, u těchto druhů můžeme předpokládat problematickou identifikaci.

V souboru GBANO pro komáry rodu *Anopheles* se vyskytuje množství druhů, jejichž mezidruhové vzdálenosti jsou ve stejném rozsahu jako vnitrodruhové vzdálenosti. U těchto druhů není možná jednoznačná identifikace žádnou metodou založenou pouze na porovnání *coxI* barcode sekvencí. Výčet těchto druhů viz kapitola 6.2.1. Taktéž v souboru FFNA pro severoamerické sladkovodní ryby se nachází množství druhů, jejichž mezidruhové vzdálenosti se překrývají s hodnotami vnitrodruhových vzdáleností, či se těmto hodnotám blíží. Tyto druhy nebyla identifikační analýza schopna správně určit, viz kapitola 6.2.2. V souborech FCFP, MNCN a BCBNC není žádný druh, jehož vzdálenost s jiným druhem by se překrývala s vnitrodruhovou vzdáleností.

6.2 VÝSLEDKY IDENTIFIKAČNÍ ANALÝZY

Automatická identifikace sekvencí se provádí přes aplikaci RND, která slouží i k tvorbě referencí. Všechny sekvence načteného souboru, které se mají identifikovat, se převedou na nukleotidové denzitní vektory, přičemž se automaticky použije stejná velikost výpočetního okna, jakou má načtený soubor referenčních druhů. Následně se každá sekvence porovnává se všemi referencemi a je přiřazena k tomu referenčnímu druhu, se kterým má nejmenší Euklidovskou vzdálenost. Výsledkem automatické identifikace je soubor ve formátu `xls`, který v prvním sloupci obsahuje názvy druhů jednotlivých identifikovaných sekvencí, v druhém sloupci jsou názvy druhů, ke kterým byly sekvence přiřazeny, a třetí sloupec obsahuje hodnoty nejmenší Euklidovské vzdálenosti, na základě které došlo k přiřazení k druhům. Tento výstupní soubor se dále používá k vyhodnocení úspěšnosti identifikace.

Sekvence, které byly použity k tvorbě druhových referencí, byly také identifikovány. Úspěšnost identifikace těchto sekvencí by měla být teoreticky vyšší než úspěšnost identifikace sekvencí, ze kterých reference tvořeny nebyly. Úspěšnost identifikace silně závisí na vnitrodruhové variabilitě druhů. Dále v textu se budou vyskytovat tyto výrazy:

- *referenční sekvence* – sekvence použitá k vytvoření druhové reference,
- *druhově referenční sekvence* – sekvence náleží druhu mající druhovou referenci,
- *rodově referenční sekvence* – sekvence, která není druhově referenční, ale její druh je stejného rodu jako nějaký referenční druh.

Kromě úspěšnosti identifikace referenčních a druhově referenčních sekvencí byla vyhodnocována i úspěšnosti identifikace rodově referenčních sekvencí. Jako úspěšná identifikace těchto sekvencí se bralo jejich přiřazení k libovolnému referenčnímu druhu, avšak stejného rodu jako je identifikovaná sekvence.

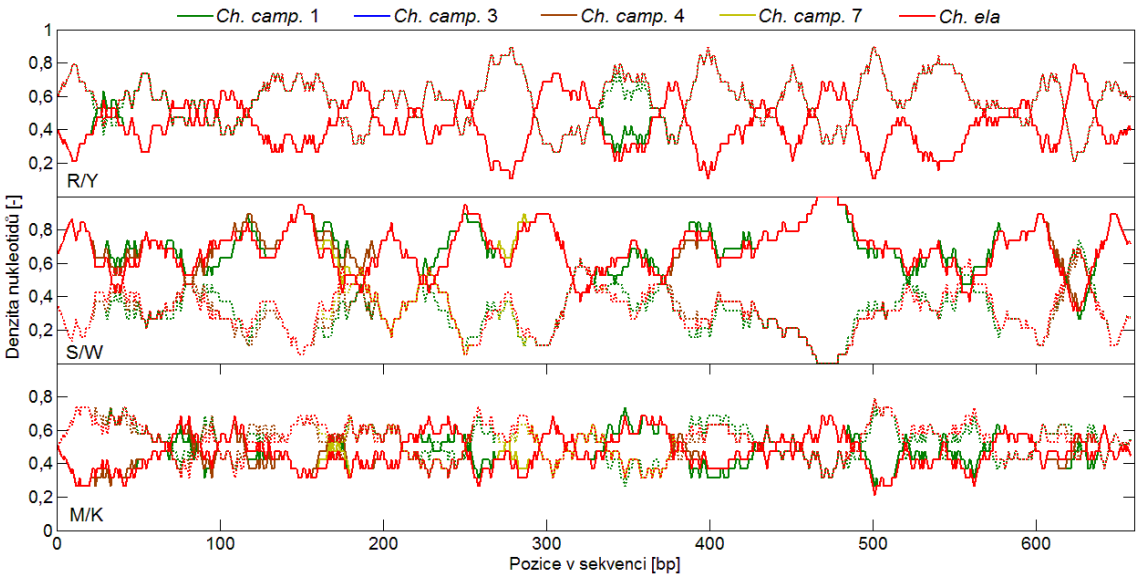
6.2.1 HMYZ

Soubor CUCAD

Referenční soubor CUCAD obsahuje celkem 716 sekvencí, přičemž z 692 sekvencí byly vytvořeny druhové reference. Zbylých 16 sekvencí patří do stejných rodů jako referenční druhy (jsou rodově referenční) a dalších 8 sekvencí patří jiným rodům či čeledím. Při verifikační identifikaci (identifikovány byly referenční sekvence) bylo pro celé spektrum délek výpočetního okna dosaženo 100% úspěšnosti identifikace referenčních sekvencí, kromě dvou problematických sekvencí viz níže.

Nesprávně byly vždy přiřazeny 2 ze 7 sekvencí druhu *Cheumatopsyche campyla*. Zařazeny byly do druhu *Cheumatopsyche ela*. Morfologická taxonomie druhů řádu chrostíků vyskytujících se v arktické oblasti Kanady je vzhledem k velkému počtu druhů a jejich podobnosti a rozdílné morfologii jednotlivých vývojových stádií a pohlaví značně obtížná. Dle literatury publikující sběr DNA barcode sekvencí obsažených v souboru CUCAD, jsou právě problémové sekvence druhu *Ch. campyla* zmíněny, neboť vykazovaly nestandardně vysokou vnitrodruhovou variabilitu [166]. Autoři usoudili, že tyto sekvence by mohly poukazovat na morfologicky kryptický druh. **Obr. 6.4** představuje sumy ND vektorů podle biochemických vlastností pro 1., 3., 4. a 7. sekvenci druhu *Ch. campyla* a jedné sekvence druhu *Ch. ela* pro výpočetní okno délky W=21 nukleotidů (toto okno bylo zvoleno z důvodu příznivé formy obrazové informace). První sekvence *Ch. campyla* byla určena jako příslušník daného druhu. Co se týče vysoce konzervativního obsahu purinových bází, je z obrázku patrná výrazná odlišnost této sekvence od sekvence druhu *Ch. ela* v přibližné oblasti 330 – 370 bp, v ostatních oblastech jsou ND vektory téměř shodné. Z vícenásobného zarovnání určené problémové sekvence druhu *Ch. campyla* číslo 3, 4 a 7 naprosto kopírují průběh R/Y ND vektorů druhu *Ch. ela*. Variabilnější sumy ND vektorů nukleotidů typu S/W ukazují více odlišností pro všechny sekvence, ale v mnohých oblastech je patrná větší shoda mezi problémovými sekvencemi se sekvencí *Ch. ela* než s *Ch. campyla*. Stejně tak to platí pro M/K sumy ND vektorů. Jen 4. sekvence *Ch. campyla* má mírně odlišný charakter než všechny ostatní. Z charakteru ND vektorů tedy usuzuji, že došlo k nesprávnému určení sekvencí 3 a 7 k druhu *Ch. campyla*, nýbrž by měly patřit druhu *Ch. ela*; určení 4. sekvence je nejisté, ale identifikační analýza ji konzistentně přiřazovala k *Ch. campyla*.

Algoritmus BLAST v BOLD databázi při hledání homologních sekvencí k těmto sporným sekvencím určil jako 100% podobné samotné sporné sekvence a dále sekvence druhu *Ch. ela*. Podobné výsledky dalo i hledání v databázi GenBank. Z toho vyplývá, že pravděpodobnější vysvětlení než kryptický druh, jak uvádí literatura, je nesprávné morfologické určení sbíraných vzorků chrostíků. Pokud tedy budeme brát, že 3. a 7. sekvence označené jako *Ch. campyla* jsou ve skutečnosti *Ch. ela*, pak byla identifikační analýza bezchybná.



Obr. 6.4 Sumy ND vektorů dle biochemických vlastností sekvencí druhu *Cheumatopsyche campyla* a *Cheumatopsyche ela*, *W* = 21. Průběhy *Ch. campyla* 3, 4 a 7 jsou ve většině případů překryté průběhem *Ch. ela*.

Chyby u rodově referenčních sekvencí se týkaly téměř vždy dvou sekvencí rodu *Limnephilus*, které byly přiřazovány k druhu *Grammotaulius interrogationis* nebo *Asynarchus mutatus* s hodnotami vzdáleností velmi podobnými hodnotám korektně přiřazených rodově referenčních sekvencí k druhu stejného rodu. Od velikosti výpočetního okna *W*=17 nukleotidů se k chybám přidala sekvence rodu *Hydroptila*, která byla přiřazována k druhu *Asynarchus mutatus* s hodnotou vzdálenosti nepřekračující obvyklé hodnoty správně zařazených rodově referenčních sekvencí. U nejdelšího okna došlo ještě k chybě u sekvence rodu *Brachycentrus*, která byla přiřazena k druhu *Agrypnia deflata*.

Při identifikaci pouze na základě porovnání sum ND vektorů purinových (A+G) nukleotidů, byly všechny rodové sekvence přiřazeny správně. Identifikace druhů na základě purinových ND vektorů byla totožná s výsledky při použití všech ND vektorů, mimo nejkratší výpočetní okno *W*=3 nukleotidy, kdy byla jedna sekvence přiřazena ke komárům rodu *Anopheles*.

Ostatní sekvence, které nebyly ani rodově referenční, byly přiřazovány vždy k některému druhu chrostíků.

Soubor DSTRI

Soubor DSTRI byl použit jen k identifikaci. Obsahuje, kromě 55 jiných sekvencí, 245 sekvencí odpovídajících referenčním druhům a 240 sekvencí patřících alespoň do referenčního rodu. Identifikace sekvencí do druhů byla naprosto úspěšná pro všechny délky výpočetního okna i v případě purinové identifikace.

Všechny sekvence rodově referenčních druhů *Limnephilus infernalis* a *L. kennicotti* byly u všech délek výpočetního okna zařazeny do druhu *Grammotaulius interrogationis*

nebo *Asynarchus mutatus*, což jsou druhy ze stejné čeledě, tudíž blízce příbuzné. Průměrné hodnoty vzdáleností pro přiřazení byly o přibližných 40 % vyšší než hodnoty u správného přiřazení rodově referenčních sekvencí. S narůstající délkou výpočetního okna se přidávaly i další sekvence jiných druhů rodu *Limnephilus* a druhu *Asynarchus rossi*, které byly přiřazeny blízce příbuznému druhu *Agrypnia colorata*. V případě identifikace pouze na základě porovnání purinových ND vektorů bylo dosaženo 100% úspěšnosti i na rodově referenčních sekvencích pro délky výpočetního okna od W=13 nukleotidů výše, u nižších oken se vyskytovalo několik chybných přiřazení.

Soubor GBANO

Větší část jedinců (1007 sekvencí) souboru GBANO sloužila k tvorbě referencí a druhá část pouze k identifikaci, kdy z nich je 1002 sekvencí druhově referenčních a dalších 84 sekvencí rodově referenčních. Soubor obsahuje pouze sekvence blízce příbuzných druhů komárů rodu *Anopheles*. Bylo vytvořeno 56 druhových referencí, přičemž 8 z nich patřilo různým variantám druhu *A. longirostris*. Je nutno poznamenat, že soubor vznikl extrakcí dat z GenBank databáze, tudíž nejde o sekvence, které byly primárně získány pro DNA barcoding.

V tomto souboru se oproti ostatním vyskytuje množství sekvencí podstatně delších než je obvyklý úsek o délce cca 650 bp genu *coxI* pro barcoding používaný. Tyto sekvence byly v souboru ponechány, což vedlo k tomu, že příslušné referenční ND vektory byly dlouhé tak jako nejdelší sekvence daného druhu. To se týkalo především druhů: *A. albimanus* (1058 bp), *A. arabiensis* (1107 bp), *A. benarrochi* (1510 bp), *A. janconnae* (1470 bp), *A. costai* (1510 bp), *A. darlingi* (1510 bp) a *A. triannulatus* (1510 bp). Některé byly naopak podstatně kratší: *A. baimaii* (438 bp), *A. dirus* (438 bp), *A. goeldii* (489 bp), *A. scaloni* (435 bp), *A. sundaicus* (441 bp) a *A. vagus* (471 bp).

Nejhorší výsledky identifikace byly získány pro nejkratší výpočetní okno W=3 nukleotidy. Úspěšnost identifikace sekvencí, z kterých byly tvořeny reference, byla 55,39 % a úspěšnost identifikace druhově referenčních sekvencí byla pouhých 45,40 %. Při purinové identifikaci byly hodnoty úspěšností o několik procent nižší a ve vysoké míře (16,45 %) byly sekvence chybně přiřazovány k druhům ze skupiny obojživelníků nebo ryb. U ostatních velikostí výpočetního okna bylo přiřazování k nepříbuzným druhům vzácné. Dále uváděné výsledky platí pro výpočetní okna od W=5 nukleotidů.

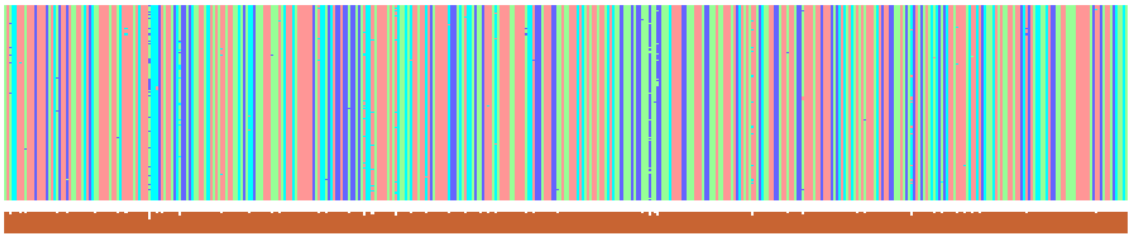
Verifikační identifikace sekvencí, které sloužily k vytvoření druhových referencí, dosahovala úspěšnosti v přibližném rozmezí 78 – 85 %. Většina chyb se vyskytovala formou záměny druhů:

A. antunesi – A. pristinus – A. marajoara,
A. arabiensis – A. gambiae,
A. baimaii – A. dirus,
A. dacidae – A. messeae,
A. epiroticus – A. sundaicus.

Identifikace druhově referenčních sekvencí nepoužitých k tvorbě referencí byla obdobná.

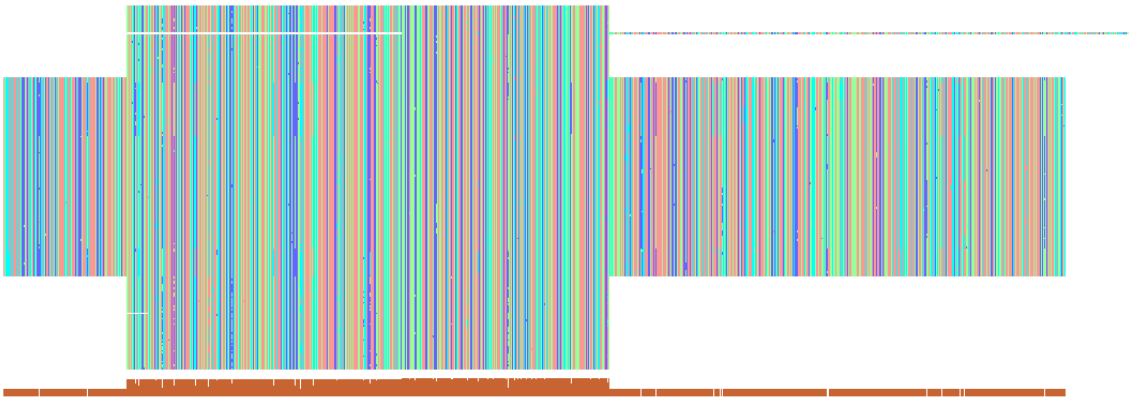
Druh *A. albimanus* mající dlouhou referenci je zastoupen 204 jedinci, z nichž polovina sloužila k tvorbě reference. Pro délky výpočetního okna W=5 a W=9 nukleotidů byla identifikace všech těchto sekvencí bezchybná. Pro ostatní délky výpočetního okna kromě toho nejdelšího byly všechny delší sekvence tříděny k jiným druhům chrostíků, především *A. benarrochi* či *A. darlingi*. Při vyhledání homologních sekvencí v databázi GenBank byly výsledkem pouze sekvence *A. albimanus*, tudíž pravděpodobnost, že by šlo opravdu o jiné sekvence, je velmi malá.

U druhů s kratšími referencemi (351sekvencí pro tvorbu referencí a 350 druhově referenčních sekvencí) byly bezchybně identifikovány sekvence druhů *A. goeldi, A. scaloni* a *A. vagus* (216 sekvencí celkem). U ostatních docházelo často k chybám typu *A. baimai* identifikovány jako *A. dirus* a opačně, *A. sundaicus* jako *A. epiroticus*. Porovnání referenčních ND vektorů a vícenásobného zarovnání (viz **Obr. 6.5**) sekvencí druhů *A. baimai* a *A. dirus* ukázalo, že tyto jsou až na několik hodnot/bodových substitucí totožné. Žádná publikace vztahující se k tomuto souboru dat nebyla vydána, neboť jde o data získaná extrakcí z databáze GenBank, tj. data byla pravděpodobně nashromážděna množstvím pracovišť v delším časovém úseku a nejednotnou metodikou. Proto jsou sekvence také nejednotné délky. Nelze tady určit, nedošlo-li k záměně druhů *A. baimai* a *A. dirus*. Pokud však byli jedinci, kterým sekvence patří, správně taxonomicky zařazeni, pak jsou tyto druhy z hlediska daného úseku mitochondriálního genu *coxI* nerozeznatelné. Stejně tak to platí pro druhy *A. epiroticus* a *A. sundaicus*, kdy druhý jmenovaný má sekvence kratší a daný úsek je téměř totožný s částí reference prvně jmenovaného druhu.



Obr. 6.5 Vícenásobné zarovnání sekvencí druhu *Anopheles baimai* a *Anopheles dirus*. Není patrný žádný předěl oddělující oba druhy.

Obdobná shoda byla zjištěna mezi druhy *A. arabiensis* a *A. gambiae* a druhy *A. dacidae* a *A. messeae*. Pokud tedy přiřazení sekvencí všech výše zmíněných druhů nebudeme brát jako chybné, pak původně udané rozmezí úspěšnosti 78 – 85 % se změní na 90 – 98 %, kde většina chyb se týká nesprávného přiřazování sekvencí druhu *A. albimanus*, který vykazuje vysokou vnitrodruhovou variabilitu (viz **Obr. 6.6**) související s výskytem tohoto malarického komára téměř v celé karibské oblasti.



Obr. 6.6 Vícenásobné zarovnání sekvencí druhu *Anopheles albimanus* ze souboru GBANO.

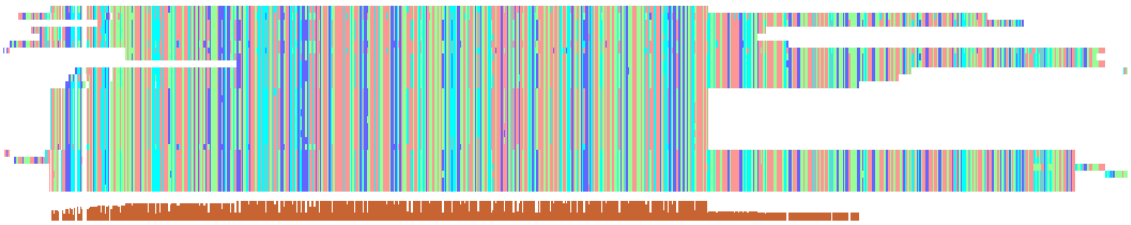
Identifikační analýza odhalila další sporné sekvence, které byly pro každou délku výpočetního okna vždy přiřazovány k jinému ale vždy k tomu stejnému druhu, než jak byly sekvence označeny. Jedná se o tři sekvence *A. barbirostris*, které byly vždy přiřazeny k druhu *A. campestris*. Prohledání databáze GenBank algoritmem BLAST nalezlo jen ty samé sekvence *A. barbirostris* následované 21 nejbližších výsledků (98% podobnost) patřících druhu *A. campestris*. Další chyby se týkaly dvou sekvencí *A. campestris*, které jsou ve vícenásobném zarovnání vůči ostatním sekvencím druhu značně posunuté. BLAST nalezl jako nejpodobnější dvě sekvence *A. barbirostris* s podobností 96 % a následně různé sekvence rodu *Anopheles* s podobností kolem 91 %. Tudíž ani v tomto případě nešlo o identifikační chyby, naopak navržená metoda identifikace odhalila sekvence, jejichž druhové zařazení může být sporné, či došlo k jiné procedurální chybě při získávání sekvencí.

Podobně další tři sekvence druhu *A. marajoara* byly přiřazovány k druhu *A. janconnae*. Tyto tři sekvence jsou podstatně delší než reference daného druhu. Analýza BLASTem dala nejednoznačné výsledky. Na předních pozicích nejpodobnějších sekvencí se vyskytovaly sekvence druhu *A. albitarsis* a *A. janconnae*. S vysokou pravděpodobností by se tedy nemělo jednat o sekvence druhu *A. marajoara*, jak byly sekvence v souboru označeny.

Už z vícenásobného zarovnání bylo patrné, že jedna sekvence *A. pristinus* je značně odlišná od ostatních stejného druhu. Přiřazována byla k druhu *A. antunesi*. I BLAST nalezl jako nejpodobnější sekvence *A. antunesi*. Podobně tomu bylo i u dvou sekvencí *A. rangeli*, které byly přiřazovány k *A. benarrochi*, což potvrdil opět i BLAST.

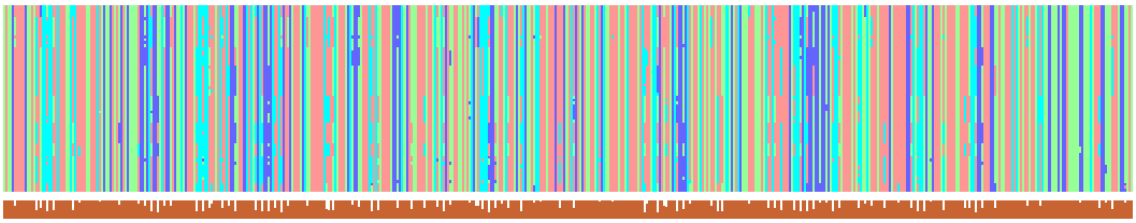
Značně problémové se ukázaly být sekvence druhu *A. subpictus*. Z vícenásobného zarovnání (viz **Obr. 6.7**) je zřejmé, že sekvence jsou značně rozdílné. Takováto vnitrodruhová variabilita by v porovnání s ostatními druhy rodu *Anopheles* byla nestandardní. Analýza algoritmem BLAST ukázala, že jen zhruba polovina ze sekvencí měla mezi deseti nejpodobnějšími sekvencemi pouze zástupce stejného druhu, ostatní měly výsledky nejednoznačné. Druhová reference vytvořená z první poloviny sekvencí

tudíž nemůže být validní a tedy samozřejmě i výsledky identifikační analýzy. Z uvedeného důvodu byly tyto sekvence z výsledků analýzy vyjmuty.



Obr. 6.7 Vícenásobné zarovnání sekvencí druhu *Anopheles subpictus*.

Všechny varianty druhu *A. longirostris* mající referenci byly pro všechny délky výpočetního okna od sebe odlišeny. Viz **Obr. 6.8** vícenásobného zarovnání níže je vidět, že sekvence vykazují variabilitu odpovídající samostatným druhům. Tyto sekvence byly z větší části správně odlišeny i na základě purinových ND vektorů, které spíše vedou k menší úspěšnosti rozlišení druhů, ale lépe rozlišují rody, neboť pozice purinových nukleotidů je více konzervována. Na základě tohoto poznatku by si druh *A. longirostris* zasloužil hlubší analýzu, protože je pravděpodobné, že nejde o jeden druh s několika variantami ale několik druhů.



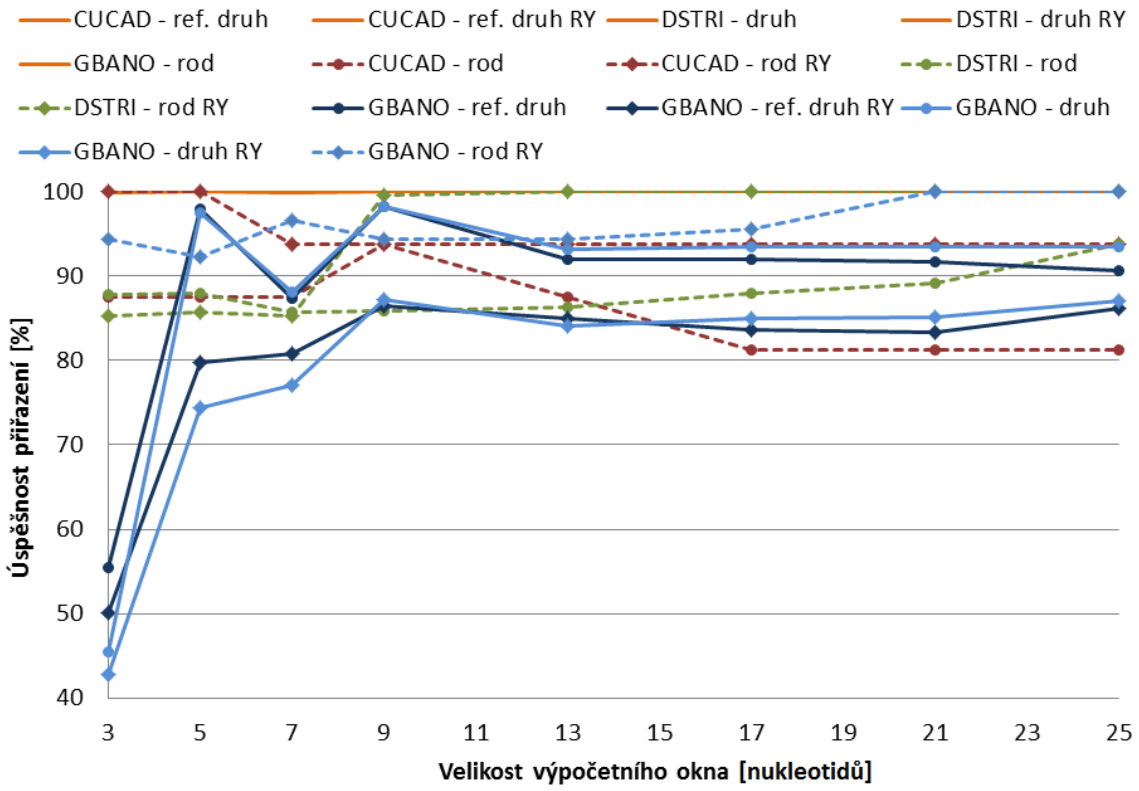
Obr. 6.8 Vícenásobné zarovnání sekvencí variant druhu *Anopheles longirostris*.

Dále byla provedena i identifikace pouze porovnáním purinových ND vektorů. Úspěšnost identifikace druhově referenčních i rodově referenčních sekvencí poklesla o několik procent (viz **Obr. 6.9**). Chyby této identifikace nebyly náhodné, ale soustředily se na některé z druhů; jiné druhy byly soustavně identifikovány správně. Nesprávné přiřazení druhově referenčních sekvencí pro všechny délky výpočetního okna se týkalo především záměn druhů:

(*A. antunesi* – *A. pristinus* – *A. marajoara*),
(*A. arabiensis* – *A. gambiae*),
(*A. baimaii* – *A. dirus*),
(*A. barbirostris* – *A. campestris*),
(*A. dacidae* – *A. messeae*),
A. dunhami – *A. goeldii* – *A. nuneztovari*,
(*A. epiroticus* – *A. sundaicus*),
A. lesteri – *A. peditaeniatus*,

přičemž druhy uvedené v závorkách nebylo možné odlišit i na základě identifikace porovnáváním všech ND vektorů a tyto druhy měly také nízké hodnoty mezidruhových vzdáleností, jejichž výpočet předcházel identifikaci, viz kapitola 6.1. Na základě záměn mezi druhy lze rámcově usuzovat, že tyto druhy mají velmi blízkého společného předka, čemuž odpovídá konzervované pozice adeninových a guaninových nukleotidů.

Obr. 6.9 zobrazuje úspěšnosti identifikace pro všechny testované velikosti výpočetního okna pro soubory CUCAD, DSTRI a GBANO. Verifikační identifikace sekvencí kanadských chrostíků (soubor CUCAD), ze kterých byly tvořeny druhové reference, byla po odstranění nejasných sekvencí druhu *Cheumatopsyche campyla* bezchybná jak při porovnávání všech ND vektorů tak jen pro purinové ND vektory, stejně jako identifikace druhově referenčních sekvencí ze souboru DSTRI a rodově referenčních sekvencí souboru GBANO.



Obr. 6.9 Úspěšnosti identifikační analýzy v závislosti na velikosti výpočetního okna pro soubory hmyzu CUCAD, DSTRI a GBANO. RY – purinová identifikace. Průběhy pro CUCAD – ref. druh, CUCAD – ref. druh RY, DSTRI – druh, DSTRI – druh RY a GBANO – rod se překrývají na hodnotě úspěšnosti 100 %.

6.2.2 RYBY

Soubor FCFP

Tento soubor se 188 sekvencemi celkem sloužil k vytvoření 38 druhových referencí, ze 175 sekvencí. V souboru se nevyskytují žádné výrazně delší nebo kratší sekvence. Verifikační analýza tohoto souboru identifikovala všech 175 sekvencí jedinců

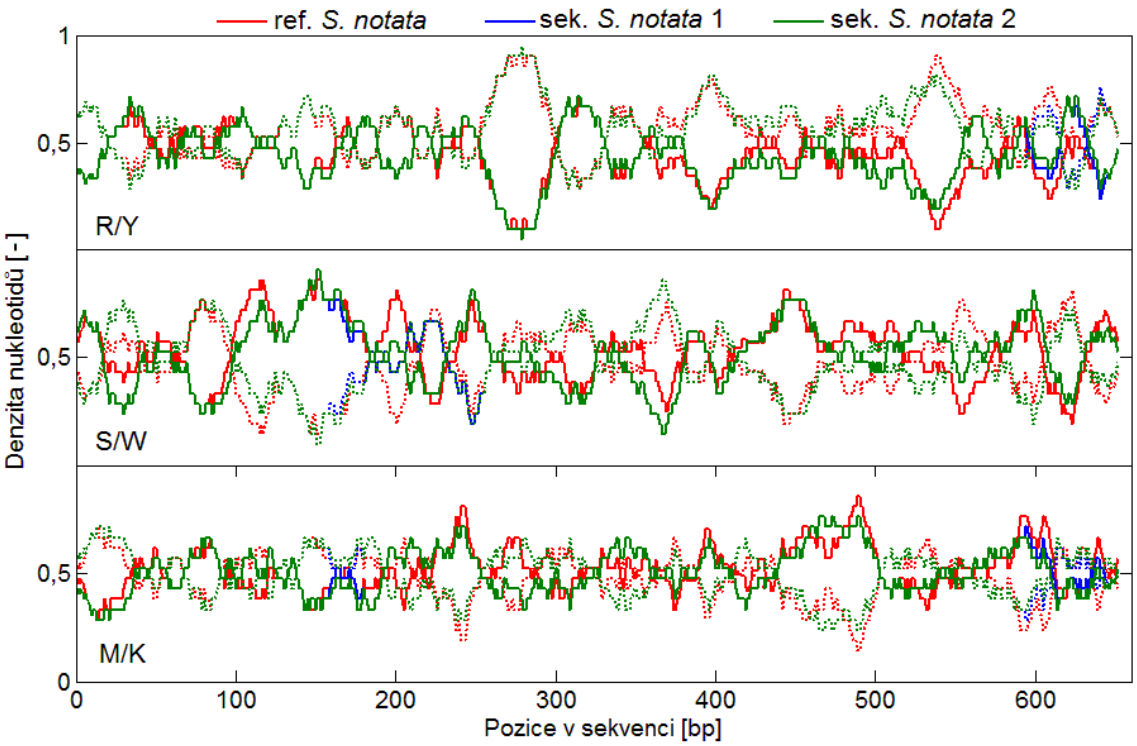
referenčních druhů stoprocentně pro všechny velikosti výpočetního okna. Dalších 8 rodově referenčních sekvencí bylo také vždy bezchybně přiřazeno k druhu správného rodu. Při identifikaci purinovými ND vektory byly všechny sekvence přiřazovány také bezchybně, kromě jedné referenční sekvence při okně *W*=3 nukleotidy.

Soubor FCFPS a FCFPW

Sekvence stejných druhů jako z referenčního souboru FCFP mořských ryb z portugalského pobřeží obsahují ještě soubory FCFPS a FCFPW. V souboru FCFPS je 124 sekvencí patřících do referenčních druhů a také 25 sekvencí rodově referenčních. Pro velikost výpočetního okna *W*=3 bp byly dvě sekvence druhu *Scorpaena notata* přiřazeny k obojživelníkům stejně jako 5 rodově referenčních sekvencí rodu *Microchirus*. Pro výpočetní okno *W*=5 nukleotidů se chyby v přiřazování druhově referenčních sekvencí týkaly jedné sekvence druhu *Macroramphosus scolopax* a dvou sekvencí druhu *Scorpaena notata*. U ostatních velikostí výpočetního okna a pro purinovou identifikaci byla sekvence *M. scolopax* přiřazována správně. Vícenásobným zarovnáním a porovnáním ND vektorů (viz **Obr. 6.10**) pěti sekvencí druhu *Scorpaena notata* použitých k vytvoření druhové reference a dvou špatně přiřazovaných sekvencí tohoto druhu ze souboru FCFPS bylo zjištěno, že tyto sekvence jsou značně odlišné. Purinová identifikace dávala pro tyto sekvence stejné výsledky. Algoritmus BLAST našel v databázi GenBank jako nejpodobnější k těmto sekvencím sekvence druhu *Scorpaena porcus* s podobností 84 %, což je malá hodnota na to, aby tyto dvě problémové sekvence patřily k danému druhu. I odborný článek popisující sběr dat v souborech FCFP, FCFPS a FCFPW uvádí, že tyto dvě sekvence by mohly patřit jinému druhu [167]. Možná jde o druh morfologicky kryptický. Zajímavá je i výraznější rozdílnost mezi purinovými ND vektory, která svojí hodnotou odpovídá spíše sekvencím patřícím do jiného rodu.

U rodově referenčních sekvencí se vyskytovaly chyby pouze v přiřazení sekvencí druhu *Microchirus azevia* a *Microchirus boscanion*. Sekvence těchto druhů vykazují nízké hodnoty podobností s homologními sekvencemi vyhledanými BLASTem obdobně jako v předchozím případě a tyto homologní sekvence ve větší části patří druhům jiných rodů než *Microchirus*. Bližší informace k těmto sekvencím nejsou v publikaci [167] uvedeny, jen to, že vykazují vysokou vnitrodruhovou variabilitu. Při purinové identifikaci byly všechny rodově referenční sekvence přiřazovány ke správnému rodu.

Soubor FCFPW obsahuje 130 druhově referenčních sekvencí, 24 rodově referenčních sekvencí a 53 jiných sekvencí. Všechny druhově a rodově referenční sekvence byly pro všechny délky výpočetního okna správně přiřazeny, stejně tak pro purinovou identifikaci.



Obr. 6.10 Sumy ND vektorů dle biochemických vlastností pro referenci druhu *Scorpaena notata* a dvou problémových sekvencí stejného druhu, *W*=21. Modrý průběh je převážně překryt zeleným průběhem.

Soubor FFNA

Soubor FFNA obsahuje 2108 sekvencí, z kterých byly vytvořeny druhové reference, dále 1733 druhově referenčních sekvencí, 343 rodově referenčních sekvencí a 121 ostatních sekvencí.

Verifikační analýza sekvencí, z kterých byly vytvořeny druhové reference, a následná identifikační analýza druhově referenčních sekvencí odhalily nesrovnalosti u některých druhů. Ukázalo se, že některé druhy od sebe nelze na základě úseku *coxI* genu spolehlivě odlišit. Tyto výsledky byly potvrzeny vícenásobným zarovnáním sekvencí a BLASTem. Jedná se o tyto páry či skupiny druhů:

Ammocrypta beanii – Ammocrypta bifascia,
Carpiodes carpio – Carpiodes cyprinus – Carpiodes velifer,
Cottus bairdii – Cottus cognatus – Cottus caeruleomentum – Cottus rhotheus,
Cyprinella callisema – Cyprinella garmani,
Cyprinella camura – Cyprinella galactura,
Esox americanus americanus – Esox niger; *Esox americanus vermiculatus* je díky několika konzervovaným substitucím odlišený,
rod *Etheostoma* viz text níže,
Ichthyomyzon bdellium – Ichthyomyzon greeleyi,
Ichthyomyzon fossor – Ichthyomyzon unicuspis,
Ichthyomyzon gagei – Ichthyomyzon castaneus,

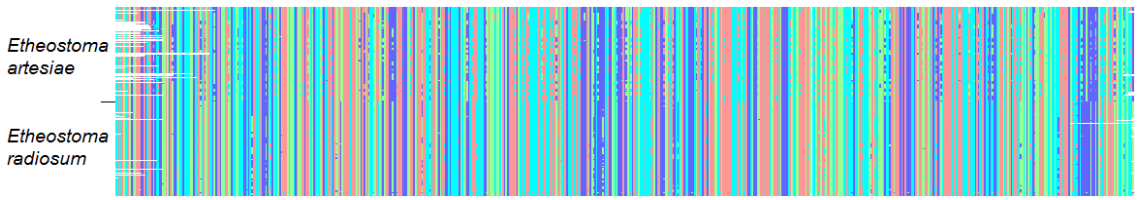
Lepomis marginatus – Lepomis megalotis,
Lythurus fumeus – Lythurus roseipinnis,
Lythurus fasciolaris – Lythurus umbratilis,
Notropis atherinoides – Notropis stilbius,
Notropis buchanani – Notropis volucellus,
Percina caprodes – Percina macrolepida – Percina kathae.

Rod *Cottus* s 20 druhy v souboru (celkem 174 sekvencí) se vyznačuje malou mezidruhovou variabilitou. Na základě analýzy dostupných sekvencí bylo zjištěno, že některé druhy od sebe nelze spolehlivě oddělit vůbec (např. *C. bairdii – C. cognatus – C. caeruleomentum – C. rhotheus*), nebo jsou od sebe odlišeny tak malým počtem konzervovaných pozic, že je toto odlišení značně nespolehlivé. Hodnoty mezidruhové variability se pohybovaly okolo 1 – 3 %, což jsou hodnoty odpovídající vnitrodruhovým variabilitám.

Rod *Etheostoma* je v souboru zastoupen celkem 1256 sekvencemi 139 druhů, přičemž druhová reference byla vytvořena pro 97 druhů. Tento rod obsahuje velké množství druhů, z nichž většina se vyskytuje na území Spojených států a mnoho druhů je endemických pro určitou řeku nebo jezero. Jak vyplynulo z analýzy, mnohé druhy jsou si v úseku genu *coxI* natolik podobné, že je není možné od sebe s jistotou odseparovat:

E. aquali – E. cinereum,
E. bellum – E. camurum – E. wapiti,
E. caeruleum – E. uniporum,
E. lynceum – E. zonale,
E. nigrum – E. podostemone,
E. percnurum – E. sitikuens,
E. spectabile – E. burri – E. lawrencei.

Sekvence druhu *E. artesiae* byly přiřazovány převážně k druhům *E. swaini, uniporum* a *whipplei*, přičemž posledně zmíněný druh by měl být nejblíže příbuzný. *E. artesiae* byl v souboru zastoupen 88 jedinci a oproti dalším druhům ze stejného rodu vykazují sekvence značnou vnitrodruhovou variabilitu, viz **Obr. 6.11** v porovnání s dalším sekvenčně početným druhem. Sekvence *E. artesiae* přiřazované k jinému druhu byly takto přiřazovány i BLASTem s vysokou mírou podobnosti.

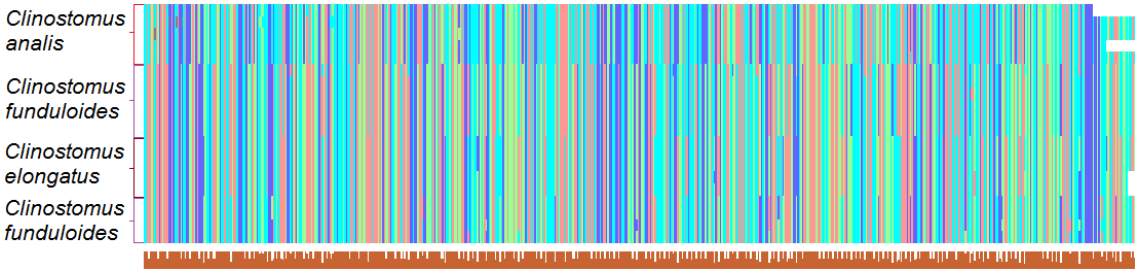


Obr. 6.11 Vícenásobné zarovnání sekvencí *Etheostoma artesiae* a *Etheostoma radiosum*.

V souboru se vyskytovalo množství sekvencí, u kterých identifikační analýza odhalila nesrovnalosti. Např.: jedna sekvence druhu *E. smithi* přiřazována k *E. striatulum*, stejně dopadla kontrola BLASTem. Je málo pravděpodobné, že by jedna sekvence druhu s nízkou vnitrodruhovou variabilitou byla skutečně sekvencí ucelené populace daného druhu. Bohužel odborný článek publikující sběr dat souboru FFNA je značně stručný a neexistuje ani fotografická dokumentace vzorků jedinců. Nelze tedy dohledat, jestli sekvence třeba nepochází z jiné separované populace, podruhu, či došlo k nějaké procesní chybě při taxonomické klasifikaci nebo při zpracování vzorků v laboratoři.

Pro všechny délky výpočetního okna byla sekvence druhu *Campostoma anomalum* přiřazována k druhu *C. oligolepis*. K této sekvenci byly algoritmem BLAST také nalezeny jako nejpodobnější sekvence právě druhu *C. oligolepis*.

Čtyři sekvence druhu *Clinostomus funduloides* byly vždy přiřazeny k druhu *C. elongatus*. Vícenásobné zarovnání i BLAST ukázaly, že tyto sekvence jsou skutečně podobnější sekvencím druhu *C. elongatus*, ale úplně shodné nejsou. Ostatní sekvence druhu *C. funduloides* a dalšího druhu rodu *Clinostomus* v souboru přítomného byly od sekvencí druhu *C. elongatus* na první pohled dobře rozlišitelné, viz **Obr. 6.12**. Purinová identifikace buď tyto sekvence přiřazovala správně nebo jako příslušníky druhu *Richardsonius balteatus*, který patří do stejné podčeledi.



Obr. 6.12 Vícenásobné zarovnání sekvencí druhů rodu *Clinostomus*.

Dvě sekvence z 12 druhu *Erimyzon oblongus* byly při všech délkách výpočetního okna přiřazovány k druhu *Erimyzon sucetta*. Kontrola BLASTem také nalezla jako nejpodobnější sekvence druhu *E. sucetta* s podobností 98 % a sekvence *E. oblongus* s podobností 94 %. Jelikož jsou oba druhy morfologicky dobře rozlišitelní, pravděpodobnost záměny druhů by měla být nízká. Purinová identifikace byla shodná, jen pro nejdelší okno W=25 nukleotidů byly sekvence přiřazeny k druhu *E. tenuis*.

Druhým druhově nejpočetnějším rodem v souboru je rod *Notropis* s 53 referenčními druhy. Z těchto druhů od sebe není možné na základě dostupných sekvencí jedinců jednoznačně separovat druhy *Notropis atherinoides – Notropis stilbius* a *Notropis buchanani – Notropis volucellus*. Při purinové identifikaci se k první neodlišitelné dvojici ještě přidal druh *N. amoenus* a k druhé dvojici druh *N. spectrunculus*. Dále purinová identifikace neodliší druhy *Notropis texanus – Notropis xaenocephalus*.

Rodově referenční sekvence

Při použití výpočetního okna *W*=3 nukleotidy byla většina chybných přiřazení k některému druhu obojživelníků. Následující popis výsledků se týká pro výpočetní okna od *W*=5 nukleotidů a výše.

U rodově referenčních sekvencí se soustavně vyskytovalo chybné přiřazení u dvou sekvencí druhu *Cyprinella gibbsi*, tří sekvencí *Cyprinella monacha* a dvou *Cyprinella trichroistia.* Přiřazovány byly k druhům rodu *Notropis* či *Ericymba*. I BLAST v GenBank databázi nalezl vysokopravděpodobnostní shodu se sekvencemi druhů zmíněných rodů, přičemž všechny tři rody jsou blízce příbuzné a náleží ke stejné podčeledi. Při purinové identifikaci byly tyto sekvence buď přiřazovány ke správnému rodu, ale také se vyskytlo přiřazování k dalším dvěma rodům *Hybopsis* a *Hybognathus* ze stejné podčeledě.

Dále sekvence druhu *Elassoma evergladei* byly přiřazovány buď k naprosto špatné skupině ryb, nebo dokonce k obojživelníkům. Z vícenásobného zarovnání byla patrná značná odlišnost od dalších sekvencí rodu *Elassoma*. BLAST jako nejpodobnější určil sice sekvence stejného rodu ale s podobností jen 85 %. Identifikace jen purinovými ND vektory však ve většině případů tyto sekvence zařadila ke správnému rodu.

Dvě sekvence druhu *Eudontomyzon hellenicus* byly přiřazovány k druhu *Caspiomyzon wagneri*. Odlišnost těchto sekvencí od ostatních sekvencí rodu *Eudontomyzon* byla dobře patrná z vícenásobného zarovnání. Algoritmus BLAST k těmto sekvencím našel jako nejpodobnější také sekvence druhu *Caspiomyzon wagneri* s podobností 93 %. Také jedna sekvence druhu *Eudontomyzon mariae* byla přiřazována k druhu *Lampetra planeri*, což s 99% podobností potvrdil i BLAST.

Dvě sekvence druhu *Gila coerulea* byly přiřazovány k druhu *Ptychocheilus umpquae* a tři sekvence *Gila nigrescens* k *Ptychocheilus oregonensis*. Porovnání sekvencí s BLASTem nedává jednoznačné výsledky, protože sekvence jsou si podobné s několika různými druhy různých rodů. Dále dvě rodově referenční sekvence *Hybopsis hypsinotus* byly přiřazovány k druhu *Notropis buchanani* a stejný výsledek dal i BLAST. Dvě druhově referenční sekvence *Hybopsis wincheli* byly pro délky výpočetního okna do *W*=13 nukleotidů přiřazovány k druhu *H. amblops* a tato větší podobnost byla také potvrzena BLAStem, avšak pro ostatní délky výpočetního okna byly sekvence přiřazovány ke správnému druhu.

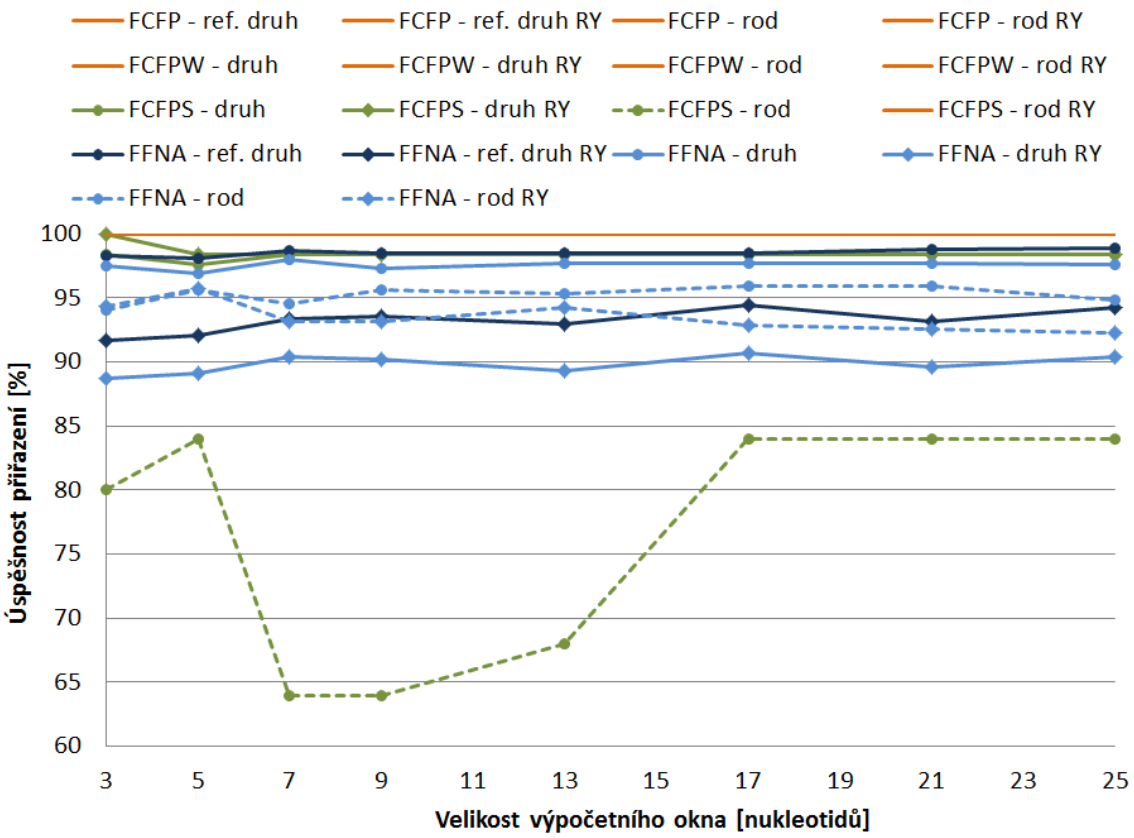
Sekvence druhů *Lampetra geminis, L. richardsoni* a *L. zanandreai* byly přiřazovány k druhům *Entosphenus tridentatus, Entosphenus hubbsi* a *Eudontomyzon lanceolata,* avšak rod *Entosphenus* je synonymem rodu *Lampetra*, tudíž jde o správné rodové přiřazení.

Dvě sekvence *Poecilia reticulata* byly přiřazovány k obojživelníkům. Z vícenásobného zarovnání s dalšími rodovými sekvencemi byla patrná značná

odlišnost. BLAST nalezl množství velmi podobných sekvencí *P. reticulata*, sekvence dalších druhů rodu *Poecilia* měly podobnost nanejvýše 87 %.

U ostatních sekvencí byla vysoká míra chybného přiřazení k obojživelníkům. Celkem bylo chyb pro všechny velikosti výpočetního okna 350; přičemž nejvíce chyb bylo pro okno *W*=3 nukleotidy 80 chyb a s délkou okna počet chyb klesal až na 21 pro okno *W*=25 nukleotidů. Identifikace na základě purinových ND vektorů počet těchto chyb radikálně snížil na celkový počet 42.

Graf závislosti úspěšnosti identifikace na velikosti výpočetního okna pro jednotlivé soubory ryb je zobrazen na **Obr. 6.13**.



Obr. 6.13 Úspěšnosti identifikační analýzy v závislosti na velikosti výpočetního okna pro soubory ryb FCFP, FCFPS, FCFPW a FFNA. Všechny průběhy pro soubory FCFP a FCFPW a průběh FCFPS – rod RY se překrývají na hodnotě úspěšnosti 100 %.

6.2.3 OBOJŽIVELNÍCI

Soubor GBAP se sekvencemi obojživelníků byl sestaven ze sekvencí z databáze GenBank, tudíž nejde o sekvence, které byly primárně sekvenované pro DNA barcoding. Celkem je v souboru 1515 sekvencí, ale mnohé druhy jsou zastoupeny méně než čtyřmi jedinci, takže z nich nebylo možné vytvořit druhové reference, kdy je potřeba minimálně tří sekvencí pro referenci a jedna sekvence pro identifikaci. Této podmínce vyhovovalo 74 druhů, přičemž druhové reference byly vytvořeny z celkového

počtu 490 sekvencí a pro samotnou identifikaci bylo použito dalších 428 sekvencí. Dále je v souboru 188 sekvencí rodově referenčních a 409 ostatních.

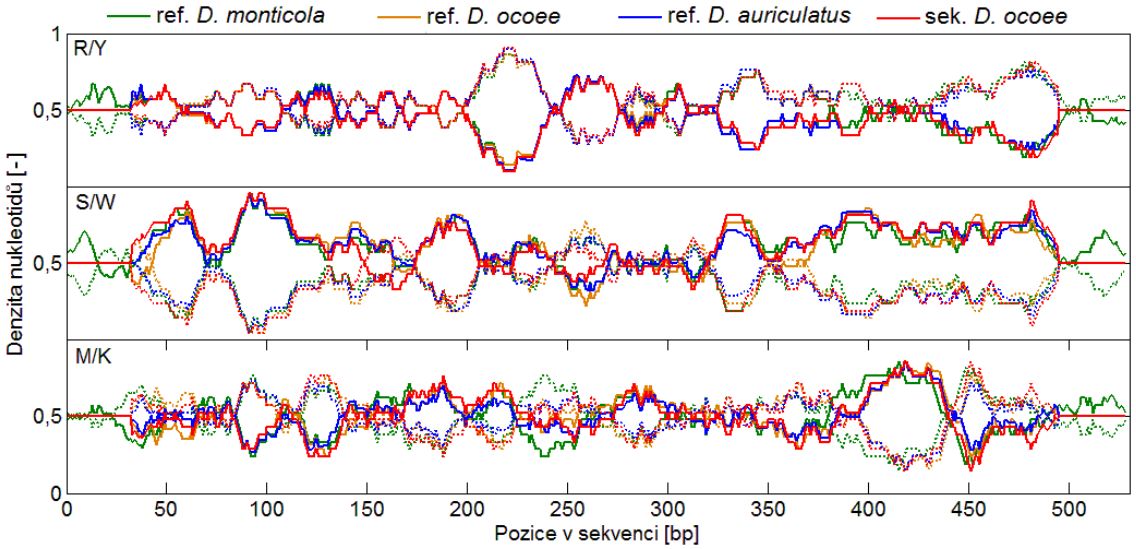
I v tomto souboru se opět projevila nízká úspěšnost identifikace pro výpočetní okno délky *W*=3 nukleotidy s vysokou mírou identifikace k druhům jiné skupiny organismů, zde konkrétně k druhům komárů rodu *Anopheles*. Pro všechny ostatní délky výpočetního okna byla jedna druhově referenční sekvence druhu *Allobates talamancae* přiřazována k druhu *Hyla meridionalis* s hodnotou vzdálenosti několikanásobně vyšší než správně přiřazené ostatní sekvence stejného druhu. BLAST v databázi GenBank nalezl jako nejpodobnější sekvence druhů z rodu *Bombina* a to s nejvyšší podobností jen 85 %. Z vícenásobného zarovnání byla patrná naprostá odlišnost od ostatních sekvencí, tudíž se pravděpodobně nejedná o sekvenci úseku *coxI*. Podobně byla jedna sekvence druhu *Dendrobates auratus* přiřazována k *Allobates caeruleodactylus*. Vícenásobné zarovnání i BLAST potvrdili, že tato sekvence patří druhému zmíněnému druhu. Tyto dva druhy jsou morfologicky dobře rozpoznatelné a v tomto ohledu nemohlo dojít k jejich záměně, pokud nešlo o vzorek z raného počátečního vývojového stádia. Samozřejmě mohlo dojít i k procesní chybě při zpracování vzorků nebo ukládání do databáze.

Dále byla vždy jedna sekvence druhu *Desmognathus ocoee* přiřazována k příbuznému druhu *Desmognathus auriculatus*. V tomto případě nalezl BLAST jako nejpodobnější sekvence druhu *Desmognathus monticola* s podobností 94 % a další druhy stejného rodu. Porovnáním ND vektorů (viz **Obr. 6.14**) je ale patrné, že daná sekvence *D. ocoee* má v některých úsecích blíže k referenci *D. auriculatus* než k referenci vlastního druhu či *D. monticola*. Reference *D. monticola* je v některých úsecích od ostatních odlišná (že je tato reference delší nemá při výpočtu vzdálenosti vliv, neboť se berou v úvahu pouze signálově zarovnané vektory). Analýza dále ukázala, že druh *Desmognathus auriculatus* má vysokou vnitrodruhovou variabilitu a s druhem *Desmognathus fuscus* jsou od sebe na úseku *coxI* nerozlišitelné.

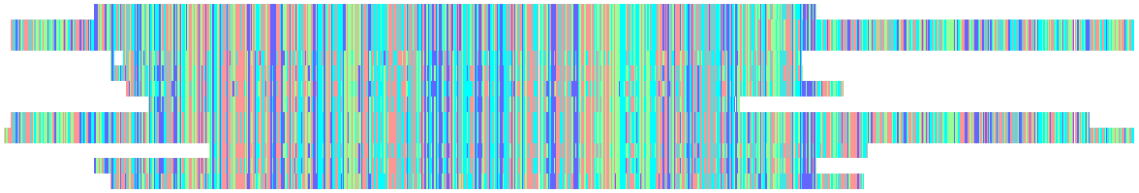
Algoritmem BLAST a vícenásobným zarovnáním byla také potvrzena příslušnost dalších sekvencí, které byly identifikační analýzou přiřazovány k jiným druhům, než jak byly tyto sekvence označeny: 2 sekvence *Bombina bombina* k *Bombina fortinuptialis*, 2 sekvence *Heterixalus tricolor* k *Heterixalus variabilis*, 4 sekvence *Onychodactylus fischeri* k *Onychodactylus zhaoermii* a 4 další sekvence k *Onychodactylus zhangyapingi*.

Ze 12 sekvencí několika druhů rodu *Fejervarya* byla pro druh *Fejervarya limnocharis* s 5 sekvencemi vytvořena druhová reference. Následná analýza však ukázala, že sekvence byly až na dvě přiřazovány k různým druhům různých rodů. Sekvence, z nichž byla vytvořena reference, jsou značně rozdílné, stejně jako ostatní rodové sekvence (viz **Obr. 6.15**). Dle literatury [168] vykazují široce a hojně se vyskytující druhy rodu *Fejervarya* neočekávaně vysokou variabilitu mitochondriálního genomu, z čehož se usuzuje na možný parafyletický původ některých druhů a

morfologicky kryptické druhy. Z tohoto důvodu nebyly výsledky identifikační analýzy těchto sekvencí započítány do celkového přehledu úspěšnosti identifikace souboru GBAP.



Obr. 6.14 Sumy ND vektorů referencí pro druhy *Desmognathus monticola*, *Desmognathus ocoee* a *Desmognathus auriculatus* a problémovou sekvenci druhu *Desmognathus ocoee*, *W*=21.



Obr. 6.15 Vícenásobné zarovnání sekvencí rodu *Fejervarya*.

Rodově referenční sekvence

Je nutné poznamenat, že taxonomie žab tropických oblastí především čeledě *Dendrobatidae* není ustálená a stále dochází k přesunům druhů mezi rody a některé rody mají i více synonymních názvů. V případě rodově referenčních druhů v souboru GBAP jde o rody či jejich synonyma: *Adelphobates*, *Allobates*, *Ameerega*, *Anomaloglossus*, *Aromobates*, *Colostethus*, *Dendrobates* a *Epipedobates*. Jde o rody blízce příbuzné s nejasnou příslušností mnoha druhů a i identifikační analýza často sekvence přiřazovala různě v rámci této skupiny. Při vyhodnocování úspěšnosti identifikace rodově referenčních sekvencí bylo bráno přiřazování v rámci této skupiny rodů jako správné.

Následující sekvence mimo výše zmíněné rody byly přiřazovány k druhu jiného rodu a toto přiřazení potvrdil i BLAST (tzn. sekvence má více podobných znaků v úseku *coxI* s jiným rodem než jak je taxonomicky zařazen).

Jedna sekvence *Batrachuperus mustersi* byla přiřazována k různým druhům, převažoval *Pachyhynobius shangchengensis* a k tomuto druhu byla sekvence

přiřazována i na základě purinových ND vektorů. BLAST udal jako nejpodobnější také sekvence různých druhů s nízkými hodnotami podobnosti. Ostatní sekvence rodu *Batrachuperus* v souboru mají stejný charakter purinových ND vektorů. Na základě nízkých podobností s ostatními druhy z rodu by si tento kriticky ohrožený druh mloka vyskytující se na malém území v Afghánistánu zasloužil podrobnější analýzu mitochondriálního genomu.

Dvě sekvence *Batrachuperus yenyuanensis* byly přiřazovány k příbuznému mloku *Lius tsipaensis*. Purinové ND vektory však na rozdíl od předcházejícího druhu zařadily tyto sekvence do správného rodu. Obě tyto sekvence mají délku přes 1500 bp stejně jako reference pro *Lius tsipaensis*. Reference druhů pro rod *Batrachuperus* a *Lius* jsou si velmi podobné.

Jedna sekvence *Eleutherodactylus mariposa* a 1 *Eleutherodactylus ronaldi* byly přiřazovány k *Allobates femoralis*; v případě purinových ND vektorů v polovině případů k některému z druhů rodu *Bufo*. BLAST nalezl 100% shodu se sekvencemi druhu *Bufo fustiger* a *Bufo peltocephalus*, které ale nejsou v souboru GBAP zastoupeny žádnou sekvencí. Rod *Bufo* patří k jiné čeledi než *Eleutherodactylus* a tyto rody jsou z hlediska morfologie dobře rozeznatelné. Přirozená shoda těchto sekvencí je nepravděpodobná, zvláště když odstup referenčních denzit pro *Eleutherodactylus auriculatus* a druhy *Bufo* je dostatečný a ostatní sekvence rodu *Eleutherodactylus* byly přiřazovány správně.

Sekvence rodu *Hyla* jsou oproti ostatním sekvencím značně posunuté; jejich prvních cca 80 bází odpovídá koncům standartního úseku *coxI*. Osm sekvencí tohoto rodu odpovídalo standardnímu úseku a nebyly proto přiřazovány k referencím daného rodu, ale různým jiným druhům.

Dvě sekvence *Ichthyophis glutinosus* byly přiřazovány k různým druhům. Tento rod je zastoupen ještě čtyřmi sekvencemi druhu *Ichthyophis bannanicus* mající druhovou referenci. Výše zmíněné dvě sekvence se od reference významně liší a i BLAST určil jako nejpodobnější sekvence různých druhů s podobností nejvýše 83 % a mezi prvními byly sekvence hlodavců a dalších savců. Identifikační analýza do velikosti výpočetního okna W=17 nukleotidů určovala sekvence jako příslušníky druhů některého z mloků, kam sekvence skutečně patří, pak mezi žáby. Při identifikaci purinovými ND vektory byly sekvence přiřazovány dokonce k rodovému druhu *Ichthyophis bannanicus* pro velikosti výpočetního okna W=5 až W=21 nukleotidů, dále k mlokům.

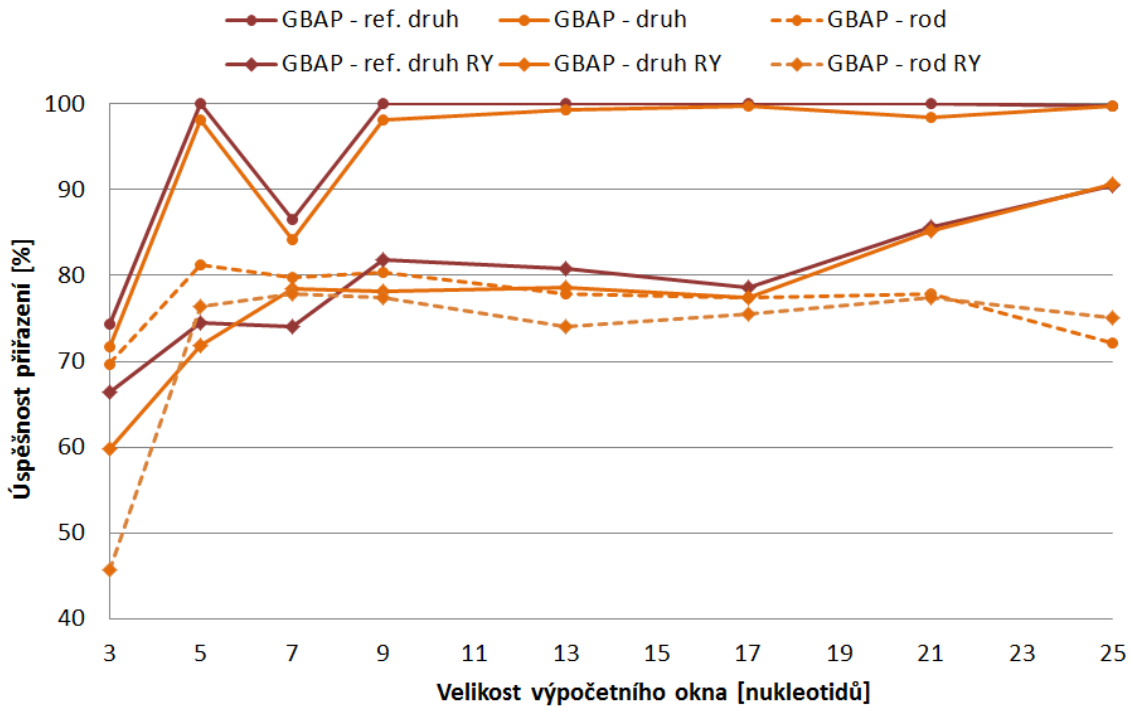
Všech šest rodově referenčních sekvencí rodu *Limnonectes* byly přiřazovány k různým druhům. Většina těchto sekvencí byla podstatně delší než standartní úsek *coxI*. BLAST nalezl jako nejpodobnější také různé sekvence, kde mezi prvními byly i ptáci či ryby. Nalezené rodové sekvence měly podobnost ne vyšší než 85 %. Identifikace purinovými ND vektory přiřadila 4 z těchto sekvencí správně pro výpočetní okna delší než W=13 nukleotidů.

Rodově referenční sekvence *Litoria caerulea* byla přiřazována k různým druhům i v případě purinových ND vektorů, stejně tak BLAST k sekvenci nenalezl jako nejpodobnější sekvence rodových příslušníků. Další nejasnosti u rodově referenčních sekvencí patří po jedné sekvenci druhů *Pseudohynobius puxiongensis* a *Pseudohynobius shuichengensis*. Tyto sekvence jsou dvakrát tak delší než ostatní a byly přiřazovány k druhům, jejichž reference byla také tak dlouhá a současně patřily do stejné čeledě. I BLAST udával jako nejpodobnější sekvence jiných druhů stejné čeledě, i když v případě, že by se sekvence zkrátily o přečnívající úseky, tak jsou s ostatními sekvencemi stejných druhů téměř totožné.

Posledními problémovými rodově referenčními sekvencemi jsou některé sekvence rodů *Rana* a *Ranitomeya* a byly přiřazovány k různým druhům stejné čeledě i při purinové identifikaci. K některým sekvencím byly BLASTem nalezeny jako nejpodobnější sekvence rodu *Maxomys* (krysa) s podobností 83 % a další sekvence hlodavců či ryb.

Úspěšnost přiřazování rodově referenčních sekvencí je velmi nízká a vypovídá především o nedostatečném informačním obsahu krátkých DNA barcodingových sekvencí k přesnému odlišení jednotlivých rodů a také klasické taxonomické rozdělení na základě morfologických znaků nemusí korespondovat s rozdělením dle molekulárních znaků.

Graf závislosti úspěšnosti identifikace na velikosti výpočetního okna je zobrazen na **Obr. 6.16**.



Obr. 6.16 Úspěšnosti identifikační analýzy v závislosti na velikosti výpočetního okna pro soubor obojživelníků GBAP.

6.2.4 PTÁCI

Soubor MNCN

Referenční soubor MNCN obsahuje 758 sekvencí a pouze dvě z nich nebyly použity pro vytvoření referencí 42 druhů. Verifikační identifikace sekvencí měla 100% úspěšnost pro všechny délky výpočetního okna, viz **Obr. 6.17**. U purinové identifikace docházelo k záměnám 5 sekvencí druhu *Thalurania furcata* za *Thalurania fannyi*.

Soubor BRAS

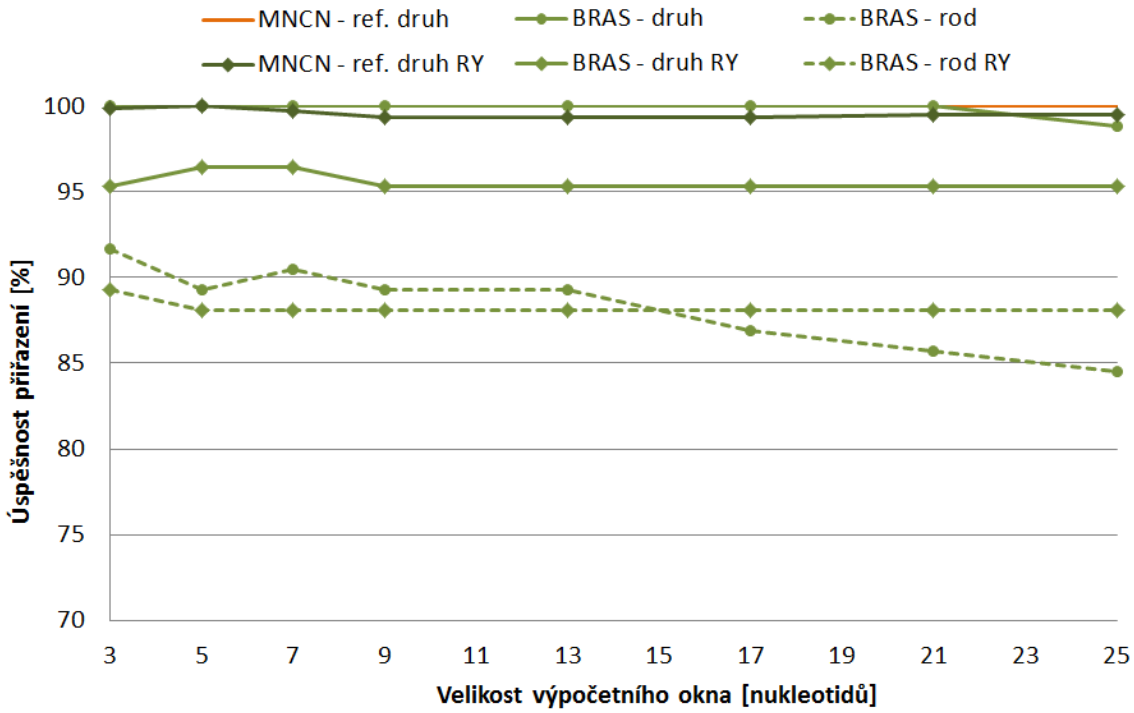
Soubor BRAS sloužící pouze k identifikaci obsahuje z celkového počtu 637 sekvencí jen 85 sekvencí druhově referenčních a 84 rodově referenčních sekvencí. Identifikace druhově referenčních sekvencí byla bezchybná, pouze pro velikost výpočetního okna W=25 nukleotidů byla jedna sekvence druhu *Myiobius barbatus* přiřazována k druhu *Schiffornis turdina*. Identifikace purinovými ND vektory přiřadila tyto sekvence správně vždy a také docházelo, stejně jako v případě souboru MNCN, k záměně mezi druhy *Thalurania furcata* a *Thalurania fannyi*.

Identifikace rodově referenčních sekvencí bezchybná nebyla. Jedna sekvence druhu *Campylopterus macrourus* (synonymum *Eupetomena macroura*) byla u všech délek výpočetního okna přiřazena k příbuznému druhu kolibříků *Thalurania furcata*. Mezi 50 nejpodobnějšími sekvencemi nalezenými BLASTem nefiguruje žádná další sekvence rodu *Campylopterus* nebo *Eupetomena*, ale jde o druhy různých rodů kolibříků a často se zde objevuje *Thalurania furcata* i s nejvyšší podobností 92 %. K tomuto druhu byla sekvence přiřazována i identifikací purinovými ND vektory.

Další rodově referenční sekvence *Hypocnemis hypoxantha* byla přiřazována k *Hylophylax poecilinotus* patřící do stejné čeledě *Thamnophilidae*. Identifikace purinovými ND vektory sekvenci identifikovala správně k rodu *Hypocnemis*. BLAST nalezl jako nejpodobnější směsici druhů dané čeledě s podobnostmi v rozsahu 90 – 93 %. V databázích GenBank i BOLD se nachází pouze tato jedna sekvence daného druhu.

Pět rodově referenčních sekvencí různých druhů rodu *Myrmotherula* byly přiřazovány téměř vždy k druhům blízce příbuzného rodu *Hylophylax*. Jenda z nich byla pro okna W=9 a W=13 nukleotidů přiřazována k správnému rodu. BLAST také přiřazoval sekvence se stejnou hodnou podobnosti 90 % k různým druhům čeledě *Thamnophilidae*, převážně rodu *Hylophylax* nebo *Hypocnemis*. Všechny tyto sekvence se v databázi vyskytují v jednom či dvěma exemplářích. Identifikace purinovými ND vektory sekvence přiřazovala stejně jako identifikace všemi ND vektory.

A jako poslední dvě rodově referenční sekvence rodu *Pipra* byly přiřazovány buď k druhu *Dixiphia pipra* (stejná čeleď) nebo *Automolus ochrolaemus* (jiná čeleď). Sekvence druhu *Dixiphia pipra* byly převažující mezi výsledky vyhledávání BLASTem. Identifikace purinovými ND vektory obě sekvence také přiřazovala k *D. pipra* nebo k *Pipra erythrocephala*.



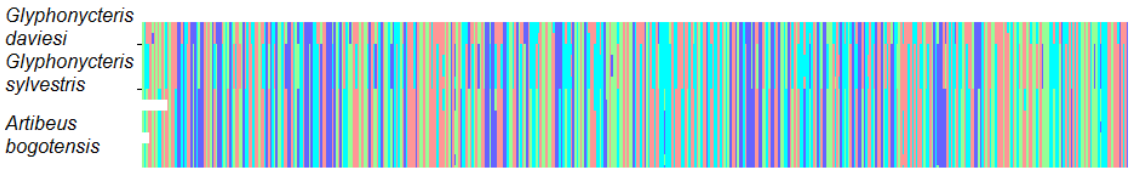
Obr. 6.17 Úspěšnosti identifikační analýzy v závislosti na velikosti výpočetního okna pro soubory ptáků MNCN a BRAS.

6.2.5 SAVCI

Soubor BCBNC

Soubor projektu BCBNC obsahuje 840 sekvencí, přičemž z 807 sekvencí bylo vytvořeno 70 druhových referencí. Referenční druhy netopýrů patřily k těmto řádům: *Emballonuridae*, *Molossidae*, *Mormoopidae*, *Noctilionidae*, *Phyllostomidae* a *Vespertilionidae*, přičemž k řádu *Phyllostomidae* patřilo 49 druhů. Verifikační identifikace druhů byla stoprocentní pro všechny délky výpočetního okna. Při identifikaci purinovými ND vektory byla většina chyb typu záměna druhů *Artibeus lituratus* – *Artibeus planirostris*, *Carollia brevicauda* – *Carollia perspicillata* a *Molossus molossus* – *Molossus rufus*.

Z 22 rodově referenčních sekvencí byly nesprávně přiřazovány 2 sekvence druhu *Glyphonycteris daviesi* a to do druhu *Artibeus bogotensis*. Z vícenásobného zarovnání (viz **Obr. 6.18**) je dobře patrná značná odlišnost těchto sekvencí od druhově referenčních sekvencí stejného rodu. Purinová identifikace však až do okna 17 nukleotidů včetně tyto sekvence zařazovala správně, pak k druhu *Sturnira lilium*. BLAST jako nejpodobnější nalezl kromě sekvencí stejného druhu sekvence rodů *Hylonycteris*, *Micronycteris* a *Vampyrodes* s podobností 84 %. Mezi stovkou nejpodobnější sekvencí nebyla ani jedna rodu *Glyphonycteris*.



Obr. 6.18 Vícenásobné zarovnání sekvencí druhů *Glyphonycteris daviesi, Glyphonycteris sylvestris a Artibeus bogotensis.*

Dalšími špatně přiřazovanými rodově referenčními sekvencemi byly 2 sekvence druhu *Pteronotus parnellii*, které byly převážně zařazeny k *Chiroderma villosum*. V databázi se nachází přes stovku barcodingových sekvencí *P. parnellii*, jejichž podobnost daná BLASTem je mezi 90 – 100 %. Dále byly jako nejpodobnější nalezeny sekvence *Tonatia saurophila* s podobností 84 %. Purinová identifikace, stejně jako v předchozím případě, přiřadila sekvence k druhové referenci stejného rodu až do okna *W*=17 nukleotidů.

Jedna sekvence *Saccopteryx bilineata* byla zařazována k nejbližšímu příbuznému druhu *Saccopteryx leptura*, stejné výsledky dal BLAST s podobností 87 %. Druhá sekvence daného druhu byla správně zařazována do okna *W*=13 nukleotidů, pak byla přiřazena převážně k *Myotis riparius*. Další dvě rodově referenční sekvence byly přiřazovány k různým druhům. Identifikace purinovými ND vektory první dvě sekvence přiřazovala správně vždy, druhé dvě až do okna *W*=13 nukleotidů.

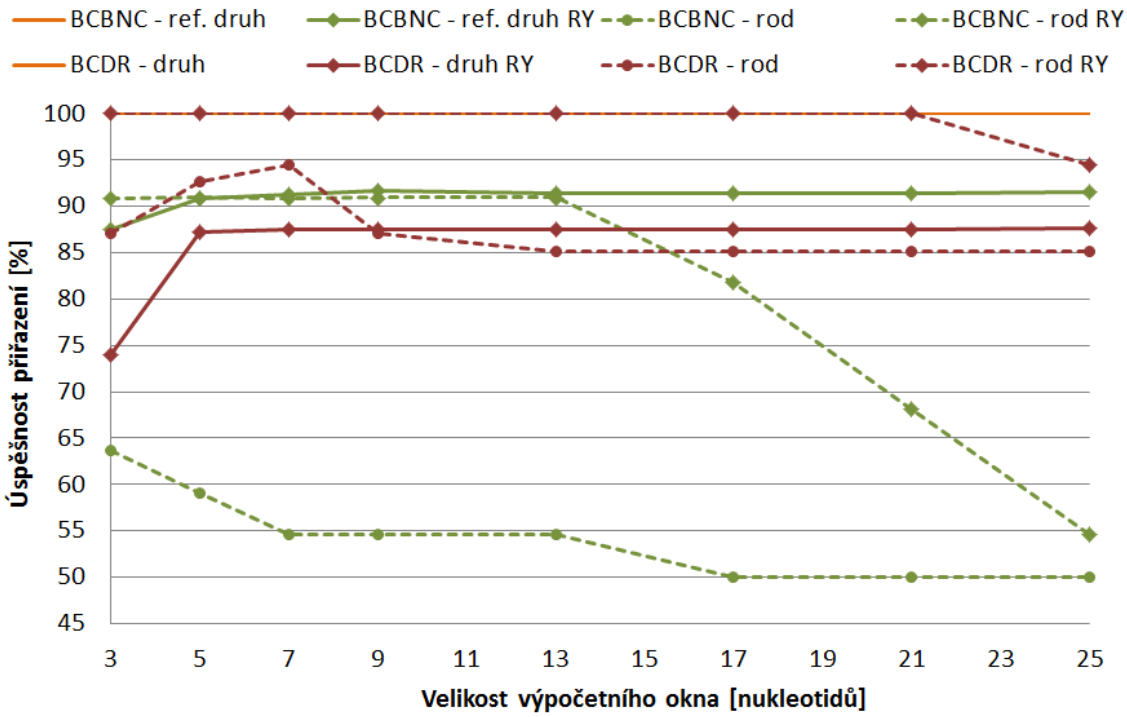
Posledními nesprávně přiřazovanými sekvencemi byly dvě sekvence druhu *Vampyressa thyone* a to do druhu *Mesophylla macconnelli* i pro všechny délky výpočetního okna při identifikaci purinovými ND vektory. Všechny rody tribusu *Stenodermatini*, kam *Vampyressa* i *Mesophylla* patří, jsou si purinových ND vektorech velmi podobné, ale v ostatních vektorech jsou odlišnosti, které umožňují identifikovat jednotlivé druhy, ale zároveň tyto odlišnosti nejsou společné rodově. BLAST nalezl jako homologní sekvence druhů *Vampyressa pusilla* (podobnost 89 %) a *Platyrrhinus helleri* (podobnost 88 %).

Soubor BCDR

Soubor BCDR obsahuje 4134 sekvencí 68 druhů, přičemž 4051 sekvencí odpovídá referenčním druhům, 54 sekvencí je rodově referenčních a 29 sekvencí patří druhům nereferenčním. Druhově referenční sekvence byly identifikovány vždy bezchybně; v případě identifikace purinovými ND vektory docházelo ke stejným druhovým záměnám jako v souboru BCBNC viz text výše.

Z rodově referenčních sekvencí byly nesprávně přiřazovány pro nejkratší výpočetní okno *W*=3 nukleotidy 4 sekvence ze 7 druhu *Pteronotus parnellii*, pro ostatní okna byly všechny sekvence přiřazovány stejně jako v souboru BCBNC viz text výše. Purinová identifikace byla úspěšná do okna *W*=21 nukleotidů, pak byly 3 sekvence ze 7 zařazovány špatně.

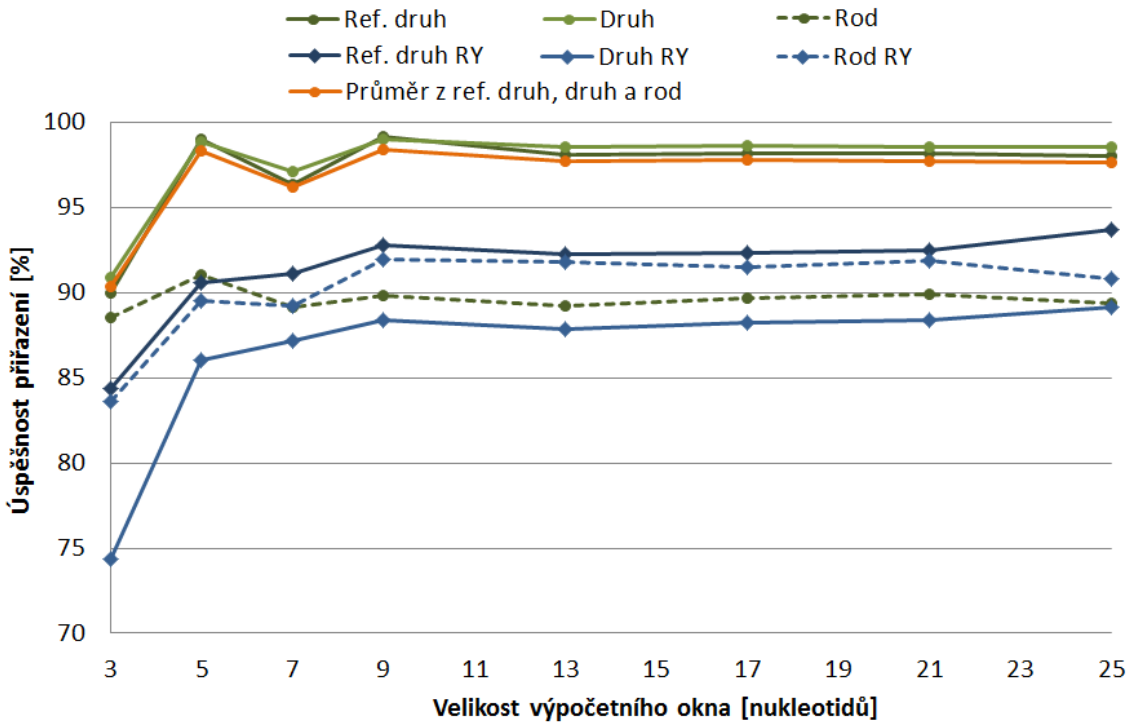
Graf závislosti úspěšnosti identifikace na velikosti výpočetního okna je zobrazen na **Obr. 6.19**.



Obr. 6.19 Úspěšnosti identifikační analýzy v závislosti na velikosti výpočetního okna pro soubory netopýrů BCBNC a BCDR. Průběhy pro BCBNC – ref. druh a BCDR - druh se překrývají na hodnotě úspěšnosti 100 %.

6.3 SHRNUTÍ

Identifikace sekvencí do referenčních druhů byla pro téměř všechny testované skupiny organismů velmi úspěšná s dosaženou shodou v úrovni téměř 99 % a ve sledovaném rozsahu velikostí výpočetního okna, kromě nejkratšího, byly úspěšnosti identifikace na velikosti okna jen málo závislé. Nejmenší hodnoty úspěšnosti identifikace byly získány při použití výpočetního okna W=3 nukleotidy, při kterém také docházelo k vyšší míře přiřazování k druhům jiné skupiny organismů, kdežto toto chybné přiřazování druhově a rodově referenčních sekvencí bylo u ostatních délek výpočetního okna ojedinělé. V průměru z úspěšností pro druhově a rodově referenční sekvence byly nejlepší výsledky dosaženy pro okno délky W=5 a W=9 nukleotidů. **Obr. 6.20** zobrazuje vážené průměry úspěšností identifikace pro všechny analyzované délky výpočetního okna. Byly provedeny další analýzy pro okna délek W=29 až W=59 nukleotidů, jejichž výsledky zde nejsou podrobně diskutovány, avšak úspěšnosti identifikace od okna W=25 nukleotidů vytrvale mírně klesají. Níže je uvedeno stručné shrnutí výsledků pro rozsah délek výpočetního okna W=3 až W=25 nukleotidů.



Obr. 6.20 Vážené průměry úspěšností identifikační analýzy v závislosti na velikosti výpočetního okna pro všechny soubory.

Tab. 6.3 shrnuje průměrné hodnoty a směrodatné odchylky normalizovaných Euklidovských vzdáleností na 1000 nukleotidů správně a nesprávně přiřazených sekvencí k druhům a rodům a také ostatních sekvencí správně přiřazených alespoň do skupiny organismů. Vzdálenosti pro rod jsou vždy několikrát větší než pro druh včetně započítání odchylek a vzdálenosti pro ostatní sekvence jsou většinou větší oproti druhovým o jeden řád.

Tab. 6.3 Průměrné hodnoty s odchylkami Euklidovských vzdáleností správně a nesprávně přiřazených sekvencí do druhů a rodů, *W*=5.

Projekt	Druh dobře	Druh špatně	Rod dobře	Rod špatně	Ostatní
CUCAD	0,9829±1,2865	-	7,8161±2,0384	8,1140±0,4528	9,4151±1,1761
DSTRI	1,5178±0,7846	-	7,3289±1,4018	8,0453±0,1801	10,5643±1,2928
GBANO	2,0508±0,9600	6,2315±1,6290	6,1241±1,8355	-	-
FCFP	0,7417±0,7384	-	7,2573±1,6933	-	9,4492±0,9034
FCFPS	1,0822±1,2176	-	7,5180±1,5427	9,9495±0,0560	9,7976±0,9311
FCFPW	0,9578±0,7901	-	6,7062±1,1999	-	9,4088±0,9787
FFNA	1,6861±1,4387	4,1555±1,1877	5,3711±2,2699	8,6417±1,0460	9,4910±1,1264
GBAP	2,4867±1,7703	3,8351±3,4277	7,5707±1,9775	9,1239±2,0241	9,7158±1,7513
MNCN	1,9806±1,0631	-	-	-	8,6290±0,1628
BRAS	3,1962±1,4791	-	6,4334±1,6062	7,1332±0,4913	7,9048±0,8104
BCBNC	1,3123±0,8995	-	8,0306±1,4822	9,5588±0,5700	10,0099±0,4121
BCDR	1,4327±0,7328	-	8,8766±0,4575	10,0880±0,1922	10,0737±0,0155

Druhové reference byly vytvořeny z celkového počtu 6016 sekvencí, pro které samozřejmě vycházely nejvyšší hodnoty úspěšnosti kontrolní identifikace. Nezávislá identifikace druhově referenčních sekvencí byla provedena na celkovém počtu 7778 sekvencí. Úspěšnosti této identifikace byla jen mírně nižší než kontrolní identifikace. Dále byla hodnocena úspěšnost přiřazení 1094 rodově referenčních sekvencí k referenčnímu druhu stejného rodu. Nejlepší vážený průměr úspěšnosti identifikace do druhů pro všechny soubory byl získán pro okno *W*=5 nukleotidů a to 98,77 % a pro okno *W*=9 nukleotidů 98,97 %. Pro rodově referenční sekvence bylo maximální úspěšnosti 86,48 % dosaženo pro okno velikosti *W*=13 nukleotidů. U purinové identifikace bylo průměrného maxima úspěšnosti pro druh dosaženo pro okno *W*=25 nukleotidů a to 91,44 %. Ačkoliv z **Obr. 6.20** se může zdát, že úspěšnost purinové identifikace do druhu má rostoucí tendenci v závislosti na velikosti okna, pro okna nad *W*=25 nukleotidů úspěšnosti vytrvale klesají. Purinová identifikace rodově referenčních sekvencí byla úspěšnější než při porovnávání všemi denzitními vektory. Maxima 90,22 % bylo dosaženo pro okno *W*=13 nukleotidů.

V případě zástupců hmyzu, kanadští chrostíci (soubor CUCAD a DSTRI), se problémy s identifikací týkaly jen rodově referenčních sekvencí rodů *Asynarchus* a *Limnephilus*; u druhého zmíněného rodu byly sekvence přiřazovány k druhům blízce příbuzných rodů ze stejné čeledě. Identifikační analýza odhalila dvě sekvence označené jako druh *Cheumatopsyche campyla*, které však byly přiřazovány k druhu *Cheumatopsyche ela*. Tuto příslušnost potvrdil i algoritmus BLAST. Druhové reference byly vytvořeny pro 32 druhů z 21 různých rodů z 10 čeledí, což znamená, že většina referenčních druhů je od sebe dostatečně evolučně vzdálena a to se také projevuje na distinkčních DNA barcodingových sekvencí.

Opačný případ představuje soubor GBANO, který obsahuje sekvence druhů jediného rodu a to komárů rodu *Anopheles*. Identifikační analýza odhalila 11 druhů, které není možné na základě úseku genu *coxI* spolehlivě oddělit. Z toho lze usuzovat, že tyto druhy byly evolučně separovány teprve nedávno. Bylo také nalezeno několik sekvencí různých druhů, které konzistentně, tj. u všech délek výpočetního okna a purinové identifikace, byly přiřazovány k jinému druhu, přičemž i BLAST u těchto sekvencí určil jako nejpodobnější sekvence přiřazovaného druhu. Je zajímavé, že druh *A. longirostris* byl zastoupen osmi variantami označenými písmeny A-H a všechny tyto varianty byly od sebe spolehlivě odlišeny a v případě purinové identifikace byla polovina z těchto variant také správně odlišena. To vede k úvaze, jestli jde skutečně o jeden druh nebo několik samostatných druhů, zejména když vezmeme v úvahu, že několik samostatných druhů je na stejném úseku mtDNA neodlišitelných.

U souborů mořských ryb z pobřeží Portugalska bylo dosaženo skvělých výsledků. Reference byly vytvořeny pro 38 druhů 23 různých rodů z 10 čeledí. Z toho plyne, že zastoupené druhy byly od sebe dostatečně evolučně vzdáleny, aby byla jejich identifikace bezproblémová. Byly však odhaleny dvě sekvence označené jako druh *Scorpaena notata*, které však vykazují od ostatních sekvencí druhu značnou odlišnost.

BLAST udal jako nejpodobnější sekvence druhu *Scorpaena porcus* avšak s nízkou hodnotou podobnosti. Charakter průběhů ND vektorů vede k hypotéze, že se jedná o naprosto jiný druh, možná dokonce i jiného rodu než *Scorpaena*, neboť sekvence byly přiřazovány k druhům jiných čeledí a dokonce i jiných skupin organismů, jak se stává u sekvencí, které nejsou ani rodově referenční.

Severoamerické sladkovodní ryby ze souboru FFNA, který vznikl extrakcí dat z databáze GenBank, jsou zastoupeny 436 referenčními druhy (95 rodů, 20 čeledí, 11 řádů). Nejvíce druhů patří rodům *Etheostoma* (97), *Notropis* (53) a *Percina* (29). U rodu *Etheostoma* byly všechny sekvence 77 druhů bezchybně identifikovány při všech délkách výpočetního okna; u ostatních druhů docházelo k chybám pouze v rámci rodu, přičemž u 16 z těchto druhů byla zjištěna taková shoda v *coxI* sekvencí, že druhy není možné od sebe odlišit. U zbylých 3 druhů, ve kterých se vyskytlo nesprávné přiřazení, bylo další analýzou vícenásobným zarovnáním a BLASTem potvrzeno, že příslušnost dané sekvence k druhu, pod kterým byla v databázi uložena, je sporná. Tedy veškeré další chyby se soustředily na jeden druh, který se vyznačoval neobvykle vysokou vnitrodruhovou variabilitou. Pouze dva druhy z rodu *Notropis* vykazovaly špatné přiřazování, které bylo způsobené malou odlišností v *coxI* těchto dvou druhů mezi sebou. U rodu *Percina* se také nalezly 3 neodlišitelné druhy a kromě jednoho druhu se pár chyb vyskytlo pouze u dvou nejkratších výpočetních oken. Purinová identifikace zhoršila o několik procent druhovou i rodovou identifikaci.

Druhových referencí pro soubor obojživelníků GBAP bylo vytvořeno 73 (29 rodů, 15 čeledí, 3 rody). Výsledky pro druhově referenční sekvence jsou srovnatelné s ostatními skupinami živočichů, ale velmi špatně byly přiřazovány rodově referenční sekvence, ačkoliv většina těchto nesprávných přiřazení byla alespoň v rámci čeledě. Purinová identifikace sice vedla k zlepšení identifikace rodově referenčních sekvencí, ale současně výrazně poklesla úspěšnost přiřazení druhově referenčních sekvencí. Bylo nalezeno mnoho sekvencí se spornou příslušností k druhu a i BLAST často potvrdil příslušnost těchto sekvencí k jinému druhu stejného rodu.

Jihoameričtí tropičtí ptáci (soubory MNCN a BRAS) byly, co se týče druhů, perfektně identifikovány až na jednu sekvenci pro dvě nejdelší výpočetní okna. Druhových referencí bylo vytvořeno 42 z 36 rodů (16 čeledí, 6 řádů, převládá řád *Passeriformes*). To znamená, že se v souborech vyskytovalo jen velmi málo blízce příbuzných druhů, u nichž by mohla být větší pravděpodobnost nesprávné identifikace. Pouze u purinové identifikace docházelo právě k identifikačním chybám mezi dvěma blízce příbuznými druhy. Identifikace rodově referenčních sekvencí bezchybná nebyla a purinová identifikace výsledky nezlepšila.

Savci byli v analýze zastoupeni pouze tropickými netopýry (soubory BCBNC a BCDR). Vytvořeno bylo 70 druhových referencí (42 rodů, 6 čeledí), přičemž 9 z nich byly různé druhové varianty. Identifikace druhů byla pro oba soubory bezchybná i pro druhové varianty. Při purinové identifikaci docházelo k záměnám mezi dvěma

dvojicemi blízce příbuzných druhů. Pro rodově referenční sekvence ze souboru BCBNC je úspěšnost relativně nízká z toho důvodu, že sekvencí bylo jen 22 a tak každá chybně přiřazená sekvence výrazně snižuje procento úspěšnosti. Purinová identifikace pro krátká okna úspěšnost výrazně zlepšila. U rodově referenčních sekvencí druhého souboru byla úspěšnost akceptovatelná a purinová identifikace dosahovala dokonce téměř vždy 100 %. Současně ale purinová identifikace nebyla schopna od sebe odlišit některé druhy stejných rodů.

6.4 DISKUZE

Vedlejším produktem identifikační analýzy nově navrženými metodami je množství biologicky významných informací o sekvencích v použitých souborech. Vzhledem k tomu, že byla k testování použita reálná data, jejichž vlastnosti mají daleko k ideálním vlastnostem, které by DNA barcodingové sekvence měly mít, došlo tedy současně k testování vlastností těchto sekvencí. V mnohých případech identifikace odhalila sekvence, které pravděpodobně patří jinému druhu, než pod kterým jsou uložené. Tyto výsledky byly ve většině případů potvrzeny i prohledáváním databáze GenBank algoritmem BLAST. Tento algoritmus hledá v databázi homologní sekvence na základě lokální podobnosti a vypočítává statistickou významnost shod. BLAST je víceméně jediný nástroj, kterým lze podobné sekvence v databázích hledat. Je však nutné zdůraznit, že výsledky BLASTu nemusí vždy korespondovat se skutečně homologními sekvencemi [118]. Především v případě rychle mutujícího mitochondriálního genomu, kdy dochází k saturaci mutací, tj. vzdáleně příbuzné organismy mohou mít velmi podobné sekvence, může BLAST dávat špatné výsledky, neboť počítá pouze s celkovou podobností (skóre zarovnání), ale už nebere v potaz umístění mutací v sekvencích. Výsledky také mohou být ovlivněny výběrem skórovacích parametrů. Naproti tomu v případě nukleotidových denzitních vektorů ovlivňuje bodová mutace charakter vektoru v závislosti na svém umístění a také je tato změna ovlivněna výskytem nukleotidů v okolí daném velikostí výpočetního okna.

Z hlediska metody identifikace druhů pomocí referenčních nukleotidových denzitních vektorů se neprojevil žádný faktor, který by měl na výsledky výrazně negativní vliv a to ani velikost výpočetního okna kromě velikosti W=3 nukleotidy. Jediným projeveným faktorem byla nevěrohodnost dat. Každý měsíc přibyde do databáze BOLD 70 – 150 tisíc nových sekvencí včetně DNA barcode sekvencí pro rostliny a houby. Při tomto množství dat vzniká otázka, jestli byl každý organismus, ze kterého byla sekvence získána, správně identifikován zkušeným taxonomem, a jestli byla každá sekvence verifikována prohledáním databáze, případně zda je skutečně podobná případným dalším jedincům stejného druhu. Samotná verifikace dat je totiž velmi pracná a časově náročná oproti dnes již vcelku dobře zvládnutým postupem sekvenace DNA barcode sekvencí, a proto jsou obavy o věrohodnosti dat na místě.

Ačkoliv byla identifikační analýza provedena na relativně malém souboru dat vzhledem k celkové rozsáhlosti databáze BOLD, bylo objeveno nemalé množství sekvencí, které se nějakým způsobem vymykají z charakteru dalších sekvencí stejného rodu nebo sekvencí velmi blízce příbuzných. Množství takovýchto sekvencí devalvuje hodnotu databáze. Množství odborníků volá po revizi přístupu k DNA barcodingu a vyzívají především k pečlivému získávání sekvencí z jednoznačně identifikovaných organismů namísto horečnatého sběru bez jasných pravidel, ke kterému dochází.

Bylo publikováno několik prací zabývajících se i jinými přístupy k identifikaci druhů než jsou distanční a znakové metody. Počátkem roku 2014 byla publikována práce, ve které byly vyzkoušeny čtyři přístupy reprezentované algoritmy strojového učení na souboru dat různých skupin organismů. Celkem byly metody testovány na 6764 sekvencí bezobratlých, obratlovců a několika desítek sekvencí hub a řas. Průměrná úspěšnost metod pro všechny testované skupiny organismů byla v rozmezí 88 až 94 % [115]. Kontrolně byla data identifikována i pomocí algoritmu BLAST s průměrnou úspěšností přibližně 89 %, což je ve všech případech menší úspěšnost než u navržené metody založené na porovnávání nukleotidových denzitních vektorů, která byla 98,97 % (avšak použita byla jiná data). K identifikaci byla také testována tří-vrstvá neuronová síť se zpětným šířením chyby, která byla testován na 159 sekvencích střevlíků a 407 sekvencí tropických motýlů, přičemž k trénování sítě bylo použito 79 a 132 sekvencí a zbylých 80 a 275 bylo nezávisle identifikováno. Úspěšnost neuronové sítě na těchto datech byla 97,5 % a 95,6 % [169].

7 IDENTIFIKACE DO ČELEDÍ

DNA barcoding by měl především sloužit k přiřazení neznámého vzorku organismu k druhu. Někteří autoři navrhovali, že by šlo DNA barcodingu využít také pro taxonomii, tj. roztřiďovat druhy do nadřazených taxonomických skupin a zkoumat tak fylogenezi druhů. Tento přístup k DNA barcodingu však byl částí odborníků záhy zamítnut jako nevhodný (viz kapitola 2.2.2).

Jak ukázala identifikační analýza, referenční nukleotidové denzitní vektory dobře reprezentovaly jednotlivé druhy testovaných skupin organismů a byly specifické i pro sekvence rodově referenční. Ačkoliv byla úspěšnost přiřazování rodově referenčních sekvencí k rodu vyhodnocována na základě přiřazení k referenčnímu druhu daného rodu a nebyly tedy vytvářeny samostatné reference pro rody, tj. nevytvářely se reference ze všech sekvencí různých druhů daného rodu, myslím si, že není potřeba samostatné analýzy rodové identifikace. Soustředila jsem se na problémovější analýzu přiřazování k čeledím.

Základním předpokladem, aby bylo možné jednotlivé analyzované sekvence přiřazovat k čeledím, je, že DNA barcodingové sekvence části genu *coxI* nesou dostatek informace o taxonomické příslušnosti sekvence založené na analýze morfologických, behaviorálních a ekologických znaků. To znamená, dostatečné množství nukleotidů je na svých pozicích v rámci čeledě konzervovaných a nedochází k jejich mutacím. Pokud DNA barcodingové sekvence nesou dostatek informace o příslušnosti do vyšších taxonomických jednotek, měla by metoda porovnávání sekvencí s referenčními nukleotidovými denzitními vektory být stejně úspěšná jako v případě druhové identifikace. Jelikož je metoda vytváření referencí a následné identifikace pomocí reprezentace DNA sekvencí nukleotidovými denzitními vektory testována na reálných datech, jsou výsledky silně ovlivněny právě vlastnostmi dat.

Byly testovány tři varianty vytváření referenčních nukleotidových denzitních vektorů pro čeledě a to metoda mediánu ND vektorů, metoda průměru ND vektorů a metoda průměru jen purinových ND vektorů ze signálově zarovnaných ND vektorů sekvencí patřící do téže čeledě, viz kapitola 5.3. Jelikož zatím neexistuje žádný volně dostupný a standardně používaný nástroj na identifikaci DNA barcodingových sekvencí do taxonomických skupin kromě vytváření fylogenetických stromů z vybraného souboru sekvencí, byla k porovnání s navrženou metodikou použita kombinace standardně používaných přístupů k vytvoření reference a následné identifikaci. Standardní reference pro čeleď je tvořena konsenzuální sekvencí, která byla určena z vícenásobně zarovnaných sekvencí téže čeledě. Identifikace pak probíhá tak, že analyzovaná sekvence je globálně zarovnána Needlemanovým-Wunchovým algoritmem se všemi referenčními konsenzuálními sekvencemi a jsou spočítány evoluční K2P distance. Sekvence je přiřazena k čeledi, s jejíž konsenzuální sekvencí má nejmenší K2P distanci.

7.1 REFERENČNÍ DATABÁZE

Databáze referencí čeledí byla vytvořena ze sekvencí živočišné třídy ptáků. Databáze BOLD shromažďuje data ze všech kontinentů a současně je DNA barcoding ptáků využíván řadou výzkumných skupin. Vhodných dat (sekvencí ptáků z různých částí světa) je proto v databázi uloženo značné množství. Pokud bude metoda identifikace do vyšších taxonomických jednotek opravdu robustní a výše zmíněný předpoklad o dostatečném informačním obsahu sekvencí platný, měla by být metoda schopna správně zatřídit sekvence ptáků z referenčních čeledí, i když pocházejících z druhů vyskytujících se na jiném kontinentu. Druhý důvod, proč byla vybrána třída ptáků, je, že ptáci mají dobře propracovanou taxonomii, s kterou můžeme výsledky porovnávat. Samozřejmě je nutné vzít v úvahu, že neustále probíhají drobné revize, co se týče čeledí a druhů do nich patřících. V současné době je počet čeledí žijících ptáků podle různých odborných ornitologických skupin od 172 [170] do 232 [171]. Klasické taxonomické třídění však nemusí nutně korespondovat s taxonomií založenou na molekulárních znacích.

Výběr samotného projektu se sekvencemi pro vytvoření referencí byl náhodný a bez odůvodnění vhodnosti tohoto projektu k vytváření referenční databáze. Od metody pro identifikaci do čeledí neznámého vzorku organismu očekáváme robustnost, proto byl referenční soubor vybrán víceméně náhodně, jediným kritériem byla dostatečná velikost souboru. Reference čeledí byly vytvořeny obdobným způsobem jako druhové reference. V tomto případě však nelze využít automatické extrakce sekvencí přes aplikaci DPT. Sekvence dané taxonomické skupiny je potřeba na základě známého taxonomického zařazení vybírat v aplikaci DPT ručně do datasetu a ten uložit pod názvem příslušné taxonomické skupiny. Takto připravené datasety se následně zpracovávají aplikací RND stejným způsobem jako v případě druhových souborů, je však možné využít všechny tři metody výpočtu reference (medián ND vektorů, průměr ND vektorů, průměr jen purinových ND vektorů) a „standardní“ metoda konsenzuální sekvence.

7.1.1 DATA PRO REFERENCE ČELEDÍ

Reference čeledí byly vytvořeny ze sekvencí části mitochondriálního genu *coxI* ptáků z DNA barcodingového projektu Birds of Argentina – Phase I (dále jen BARG), který byl jako FASTA soubor stažen z databáze BOLD. Celkem se v souboru nachází 1588 sekvencí. Pro vytváření referencí byly vybrány pouze sekvence delší nebo rovny 600 bp. V **Tab. 7.1** je uveden soupis ptačích řádů a čeledí, které byly zařazeny do referenční databáze. Aby byla čeleď zařazena mezi reference, musely být v BARG projektu alespoň tři sekvence různých druhů z dané čeledě. Celkem obsahují referenční databáze 38 čeledí vytvořených ze 445 druhů o celkovém počtu 1450 sekvencí.

Tab. 7.1 Soupis řádů a čeledí projektu BARG, z nichž byly vytvořeny reference čeledí.

Řád	Čeleď	Počet druhů	Počet jedinců
Accipitriformes	Accipitridae	19	37
Falconiformes	Falconidae	15	15
Anseriformes	Anatidae	22	81
Caprimulgiformes	Caprimulgidae	5	14
Columbiformes	Columbidae	13	47
Coraciiformes	Cerylidae	3	10
Cuculiformes	Cuculidae	6	15
Gruiformes	Rallidae	10	21
Charadriiformes	Charadriidae	7	22
	Scolopacidae	6	15
	Sternidae	5	9
	Thinocoridae	2	9
Passeriformes	Cardinalidae	8	30
	Emberizidae	6	29
	Formicariidae	3	8
	Fringillidae	7	20
	Furnariidae	54	224
	Hirundinidae	4	12
	Icteridae	19	73
	Mimidae	4	22
	Motacillidae	3	10
	Parulidae	7	21
	Pipridae	3	12
	Thamnophilidae	6	21
	Thraupidae	49	166
	Tityridae	4	10
	Troglodytidae	3	25
	Turdidae	9	39
	Tyrannidae	70	220
Pelecaniformes	Ardeidae	9	21
Phoenicopteriformes	Phoenicopteridae	3	7
Piciformes	Picidae	17	57
	Ramphastidae	3	4
Podicipediformes	Podicipedidae	4	5
Psittaciformes	Psittacidae	14	37
Strigiformes	Strigidae	9	20
Tinamiformes	Tinamidae	8	19
Trochiliformes	Trochilidae	14	43
Celkem		442	1450

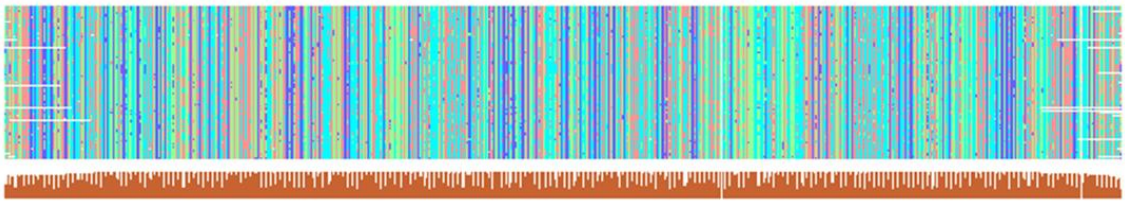
Druhy ptáků byly zatříděny do řádů a čeledí na základě klasické taxonomie, kterou bereme jako referenční. Klasická taxonomie založená na morfologických, behaviorálních a ekologických informací však nemusí plně odpovídat skutečné příbuznosti druhů. Zatřídění několika druhů vyskytujících se v projektu BARG je dokonce sporné, jedná se o některé drobné pěvce z řádu Passeriformes, kde se hranice mezi některými čeleděmi prolínají. Taxonomické zatřídění nám tudíž do referencí vnáší chybu. Řád Falconiformes (sokolovití) podle některých odborníků neexistuje a je součástí řádu Accipitriformes (dravci) jako čeleď Falconidae. Avšak podle nového výzkumu retroposonů (krátké specifické úseky DNA) je čeleď Falconidae příbuznější s řádem Psittaciformes (papoušci) a Passeriformes (pěvci) než s ostatními čeleděmi dravců a měla by tedy být od řádu Accipitriformes oddělena [172].

Čeleď Ardeidae (volavkovití) dle starších zdrojů patří k řádu Ciconiiformes (brodiví), novější zdroje však uvádějí, že dle genetického výzkumu obsahuje tento řád dvě nepříbuzné skupiny. Až na čeleď Ciconiidae (čápovití) byly všechny ostatní čeledě včetně Ardeidae přesunuty k řádu Pelecaniformes (veslonozí) [173].

Nukleotidová variabilita čeledí

Variabilitu sekvencí je obtížné jednoznačně kvantifikovat. Často se používá následující přístup. Sekvence všech jedinců z dané čeledě byly vícenásobně zarovnány, pak byla vytvořena konsenzuální sekvence a pomocí ní byl spočítán počet rozdílných nukleotidů v bloku zarovnaných sekvencí vůči konsenzuální sekvenci. Tento počet byl podělen počtem sekvencí v čeledi. Byla tak získána průměrná hodnota počtu rozdílných nukleotidů na sekvenci vůči konsenzuální sekvenci. V ideálním případě bychom pro co nejpřesnější kvantifikaci variability potřebovali co největší počet sekvencí dané čeledě. Jelikož mají všechny konsenzuální sekvence čeledí stejnou délku 694 bp nebylo nutné provádět normalizaci na jednotku délky sekvencí. **Tab. 7.2** shrnuje hodnoty průměrných variabilit jednotlivých čeledí ze souboru BARG. Pro porovnání jsou uvedeny i počty sekvencí v dané čeledi.

Na **Obr. 7.1** je pro představu znázorněno vícenásobné zarovnání sekvencí čeledě Tyrannidae, které odpovídá druhá nejvyšší variabilita. Hnědě zobrazený hřebínek pod zarovnáním představuje míru zastoupení nukleotidů z konsenzuální sekvence na dané pozici.



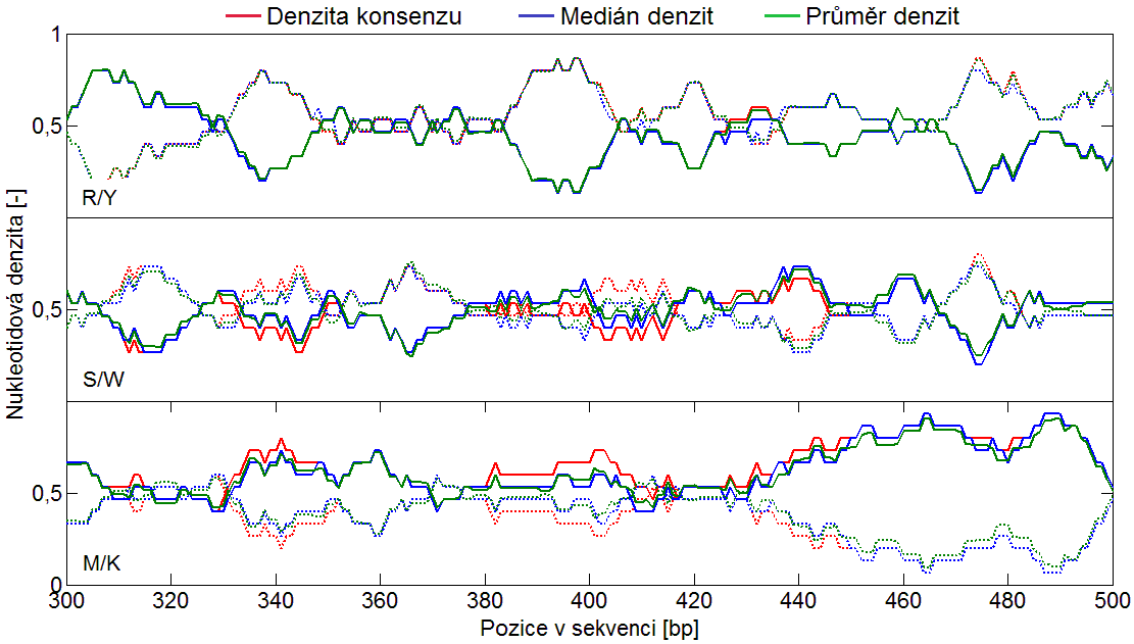
Obr. 7.1 Vícenásobné zarovnaní sekvencí pro čeleď Tyrannidae.

Tab. 7.2 Průměrné variability v rámci referenčních čeledí ze souboru BARG pro délku sekvencí 694 bp.

Název čeledi	Variabilita	Sekvencí	Název čeledi	Variabilita	Sekvencí
Accipitridae	47,05	37	Phoenicopteridae	13,86	7
Anatidae	50,31	81	Picidae	55,95	57
Ardeidae	49,57	21	Pipridae	41,92	12
Caprimulgidae	46,79	14	Podicipedidae	31,60	5
Cardinalidae	46,93	30	Psittacidae	53,43	37
Cerylidae	45,00	10	Rallidae	47,33	21
Columbidae	58,98	47	Ramphastidae	41,25	4
Cuculidae	65,00	15	Scolopacidae	56,87	15
Emberizidae	42,72	29	Sternidae	39,67	9
Falconidae	55,73	15	Strigidae	67,70	20
Formicariidae	61,13	8	Thamnophilidae	62,81	21
Fringillidae	35,80	20	Thinocoridae	24,00	9
Furnariidae	67,72	224	Thraupidae	52,42	166
Hirundinidae	47,08	12	Tinamidae	83,26	19
Charadriidae	55,86	22	Tityridae	63,40	10
Icteridae	43,88	73	Troglodytidae	25,68	25
Mimidae	13,55	22	Trochilidae	59,72	43
Motacillidae	28,30	10	Turdidae	28,74	39
Parulidae	39,14	21	Tyrannidae	73,60	220

Referenční ND vektory pro čeledě vypočítané metodami mediánu a průměru ND vektorů se od sebe navzájem liší a liší se také ND vektory konsenzuální sekvence pro čeleď. Rozdíly vznikly díky variabilitě v rámci čeledí, kterou každá z metod reprezentuje jiným způsobem. Na **Obr. 7.2** jsou znázorněny ND vektory vypočítané těmito metodami pro část sekvence mitochondriálního genu *coxI* čeledě dravců Accipitridae (jestřábovití). Je dobře znatelné, že suma ND vektorů pro purinové nukleotidy jsou značně konzervované a referenční konsenzuální, mediánové a průměrné denzity se tedy liší neznatelně. Nejvíce rozdílů je patrných v sumách ND vektorů amino/keto skupiny.

Rozdíly mezi referenčními ND vektory úzce souvisí s variabilitou čeledí. Čím větší variabilita čeledí, tím větší rozdíly mezi referenčními ND vektory jednotlivých metod. Hodně variabilní čeledě mají signifikantní rozdíly i v sumě denzit puriny/pyrimidiny. Jelikož konsenzuální sekvence obsahuje nukleotidy s nejvyšší frekvencí výskytu, zkresluje celkový charakter sekvencí čeledě v případě porovnatelných frekvencí výskytu nukleotidů na určité pozici. Tento fakt může ovlivňovat identifikaci sekvencí do čeledí.



Obr. 7.2 Rozdíly mezi referenčními ND vektory pro čeleď Accipitridae vytvořené třemi metodami. Medián a průměr ND vektorů (denzit) pro W=15.

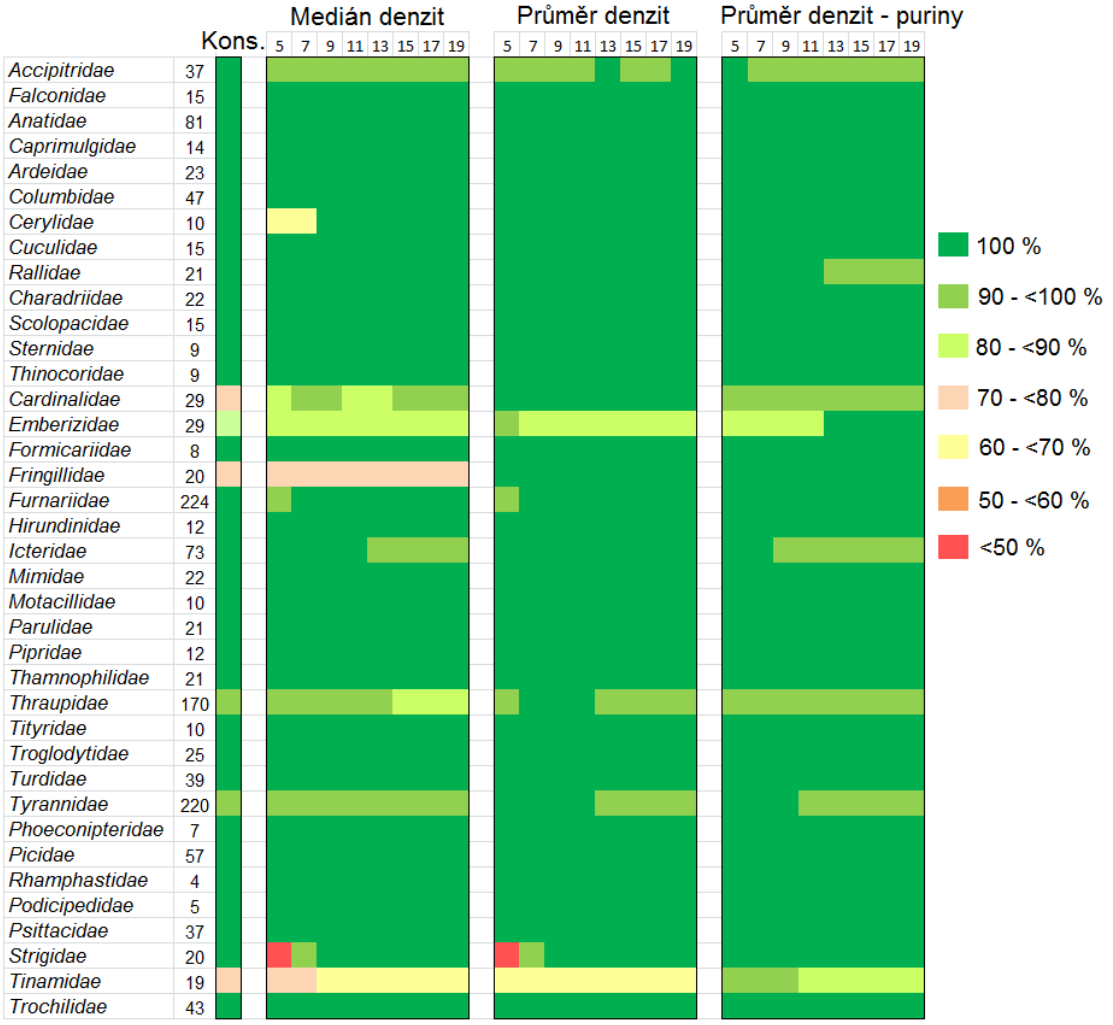
7.1.2 VERIFIKACE REFERENCÍ PRO ČELEDĚ

Verifikace metod výpočtu referencí pro čeledě spočívá v identifikaci do čeledí těch stejných sekvencí, z kterých byly vytvořeny reference čeledí. Pokud by úspěšnost identifikace na těchto sekvencích byla nízká, byly by reference dané metody ještě méně úspěšné pro jiné sekvence, a tudíž pro identifikaci nevhodné.

Verifikovány byly reference čeledí tvořené konsenzuálními sekvencemi, mediánovými ND vektory, průměrnými ND vektory a průměrnými ND vektory jen purinových nukleotidů. Každá sekvence z projektu BARG byla porovnána s referencemi vytvořených dle všech metod. Sekvence byly zatříděny do čeledí podle nejmenší hodnoty K2P distance v případě konsenzuální reference (RK) a Euklidovské vzdálenosti pro mediánové ND reference (RM), průměrné ND reference (RP) a průměrné ND reference jen purinových nukleotidů (RP-RY). Pro metody RM a RP se Euklidovská vzdálenost počítá dle vzorce (5.2) a pro metodu RP-RY dle vzorce (5.3). Analýza byla provedena pro velikosti výpočetního okna od W=5 do W=19 nukleotidů pro výpočty nukleotidových denzitních vektorů. Velikost výpočetního okna W=3 nukleotidy nebyla vzhledem k nízké míře úspěšnosti identifikace sekvencí do druhů použita.

Přehled úspěšnosti identifikace sekvencí do čeledí porovnáváním s referencemi je zobrazen na následujících obrázcích **Obr. 7.3** až **Obr. 7.5**. Úspěšnost identifikace do řádu byla získána tak, že se hodnotilo přiřazení k čeledi, která patřila do daného řádu (pouze pro řády Charadriiformes, Passeriformes a Piciformes, které měly více než jednu referenční čeleď). Např.: pokud sekvence byla přiřazena k čeledi, do které nepatří, pak byla její identifikace do čeledi neúspěšná, ale pokud přiřazená čeleď patří do řádu, kam

daná sekvence spadá, pak byla její identifikace do řádu úspěšná. Pokud řád byl v referenční databázi zastoupen pouze jednou čeledí, pak jsou samozřejmě úspěšnosti identifikace do řádu a čeledi shodné.



Obr. 7.3 Grafické znázornění procentuální úspěšnosti identifikace do čeledí metodami konsenzus (Kons. = RK), medián denzit (RM), průměr denzit (RP) a průměr denzit jen purinových nukleotidů (RP-RY) při délkách výpočetního okna *W*=5 až *W*=19 nukleotidů (vertikálně) pro soubor BARG.

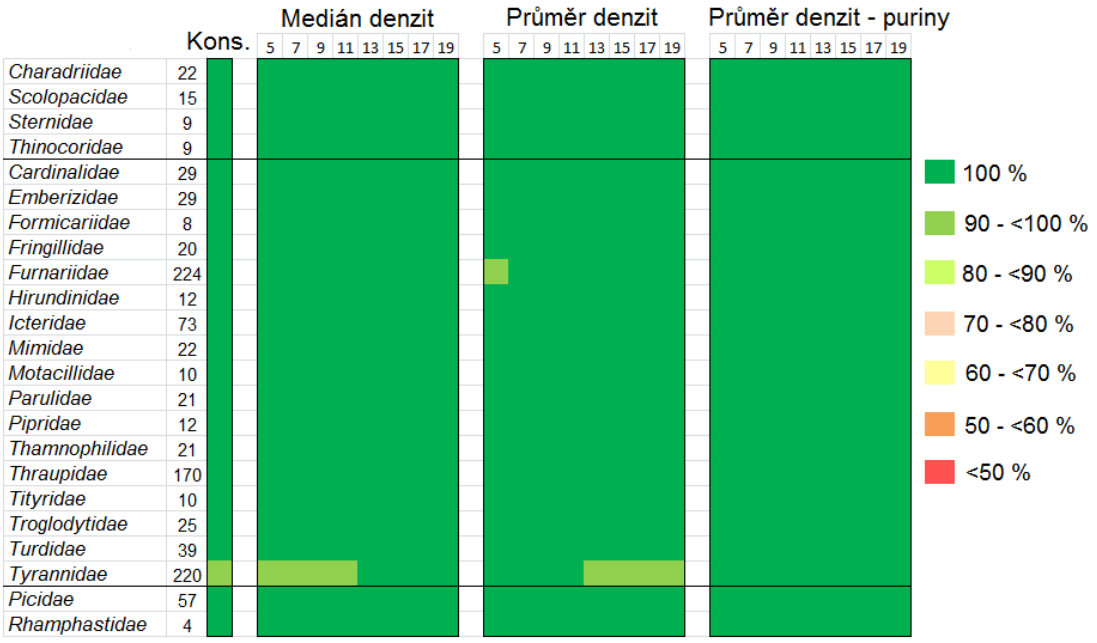
Obr. 7.3 zjednodušeně graficky znázorňuje procentuální úspěšnost identifikace do čeledí při verifikaci všemi metodami využívající nukleotidové denzitní vektory v závislosti na délce výpočetního okna a porovnání s metodikou konsenzuální sekvence. Vertikální směr odpovídá pořadí čeledí; horizontální směr odpovídá délce výpočetního okna *W*=5, 7,…, 19 nukleotidů ve směru zleva doprava. Je vidět, že metoda průměru ND vektorů (RP) má celkově nejlepší výsledky pro délky výpočetního okna *W*=9 a *W*=11 nukleotidů. Vážený průměr úspěšnosti identifikace čeledí metodou RK byl 96,49 %, nejlepší hodnota pro metodu RM byla 98,08 %, pro metodu RP 99,31 % shodně pro výpočetní okna délek *W*=9 a *W*=11 nukleotidů a pro metodu RP-RY 99,04 5 pro *W*=5 nukleotidů.

K méně úspěšným při identifikaci patří několik čeledí z řádu Passeriformes (pěvci), který je co do počtu čeledí a celkově druhů velmi obsáhlý. Mnoho druhů pěvců je obtížné správně klasifikovat do čeledí díky jejich morfologické podobnosti a blízké příbuznosti jednotlivých čeledí. U všech metod v případě ne stoprocentní úspěšnosti, byly sekvence z čeledí Cardinalidae, Emberizidae a Fringillidae nejčastěji zatříděny do čeledě Thraupidae. Současně druhy z čeledě Thraupidae byly klasifikovány jako Cardinalidae nebo Emberizidae. Všechny tyto čeledě patří do nadčeledě Passeroidea. Na základě těchto výsledků můžeme vyslovit hypotézu, že druhy patřící do nadčeledě Passeroidea jsou v případě části genu *cox1* těžko charakterizovatelní především na základě konsenzuální sekvence. Metoda průměru ND vektorů ale byla schopná sekvence těchto čeledí převážně odlišit. Komplikovaná příbuznost nadčeledě Passeroidea byla studována i na základě mitochondriálního genu *cytochrome b* a NADH dehydrogenase subunit 2 [87].

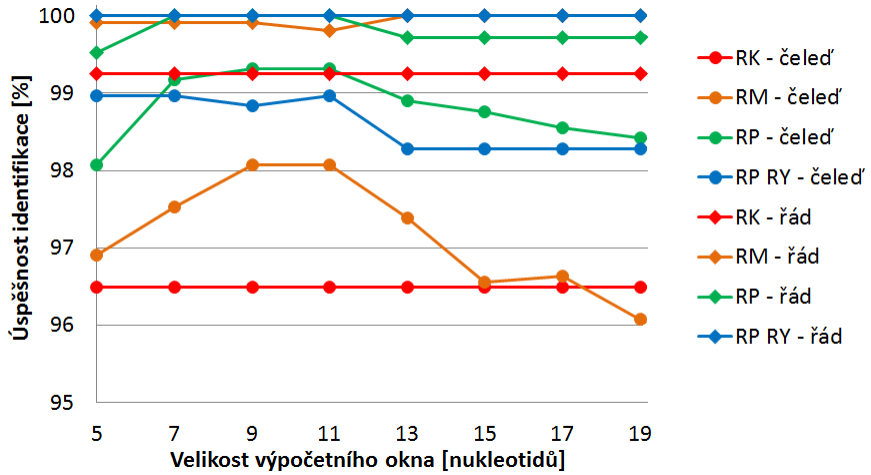
Nízkou úspěšnost u většiny metod a délek okna měla čeleď Tinamidae (tinamovití) a to nejčastěji 73,68 % kromě metody RP-RY, která s touto čeledí dosáhla úspěšnosti téměř 95 %. Vždy byly špatně identifikovány stejné sekvence. Tito ptáci patří mezi jedny z nejstarších doposud žijících ptáků a patří do samostatné podtřídy Palaeognathae, kam patří ještě pštrosi. Všichni ostatní ptáci patří do podtřídy Neognathae. Vysoká variabilita mezi druhy je tedy vysvětlena dlouhou evolucí příslušných druhů, kdy docházelo k hromadění bodových mutací v mitochondriálním genomu, což také zapříčiňuje nízkou hodnotu úspěšnosti identifikace, neboť ke genetickému oddělení jednotlivých druhů mohlo dojít dříve než o oddělování jednotlivých řádů a čeledí podtřídy Neognathae.

Zajímavá je velmi nízká úspěšnost identifikace čeledě Strigidae metodami RM a RP pro výpočetní okno W=5 nukleotidů, která byla pod 50 %, ale pro výpočetní okna od W=11 nukleotidů byla úspěšnost již 100%. Metoda RP-RY s těmito sekvencemi nechybovala vůbec.

Obr. 7.4 podobně znázorňuje úspěšnost identifikace do řádu pouze sekvencí náležících do čeledí, kterých bylo daného řádu mezi referencemi více. Úspěšnost je pro všechny metody ve většině případů stoprocentní, ale opět se ukázalo, že metoda průměru ND vektorů jen purinových nukleotidů dosahuje nejlepších výsledků. RP-RY přístup zlepšil identifikaci do řádů na 100 % pro téměř všechny délky výpočetního okna, ale identifikace do čeledí byla mírně nižší než v případě porovnávání denzitních vektorů všech nukleotidů. Nejvýraznějšího zlepšení při identifikaci do čeledí pomocí purinové identifikace bylo dosaženo pro čeleď Strigidae, kdy úspěšnost byla stoprocentní pro všechny délky výpočetního okna, přičemž u klasické identifikace metodou RP byla úspěšnost pro okno délky W=5 nukleotidů jen 45 %. A také se zvýšila úspěšnost pro čeleď Tinamidae, konkrétně došlo ke zvýšení pro okna délky W=5 a W=9 nukleotidů na necelých 95 % a pro ostatní délky okna na 84 %. Naopak u čeledí nadčeledě Passeroidea došlo k mírnému poklesu úspěšnosti v důsledku záměn mezi těmito velmi blízce příbuznými čeleděmi.



Obr. 7.4 Grafické znázornění procentuální úspěšnosti identifikace do řádů metodami konsenzus (Kons. = RK), medián denzit (RM), průměr denzit (RP) a průměr denzit jen purinových nukleotidů (RP-RY) při délkách výpočetního okna *W*=5 až *W*=19 nukleotidů (vertikálně) pro soubor BARG. Řády jsou odděleny horizontálami v pořadí: Charadriiformes, Passeriformes a Piciformes.



Obr. 7.5 Vážené průměrné procentuální úspěšnosti identifikace sekvencí BARG do čeledí a řádů v závislosti na velikosti výpočetního okna pro metodu mediánové denzity (RM), průměrné denzity (RP) a průměrné denzity purinových nukleotidů (RP-RY) v porovnání s metodou konsenzuální sekvence (RK), která se v okně nepočítá.

Obr. 7.5 zobrazuje pro všechny testované metody vážené průměry úspěšností identifikace pro všechny čeledě a řády v závislosti na velikosti výpočetního okna pro metody založené na nukleotidových denzitních vektorech. Všechny metody založené na výpočtu ND vektorů mají v určitém rozsahu délek výpočetního okna lepší hodnoty úspěšnosti přiřazení do čeledí i řádů. Vážený průměr úspěšnosti identifikace řádů metodou RK byl 99,25 %, nejlepší hodnota pro metodu RM byla 100,0 % pro délky výpočetního okna *W*=13 až *W*=19 nukleotidů, pro metodu RP byla úspěšnost 100,0 %

pro výpočetní okna délek W=7 až W=11 nukleotidů a W=21 nukleotidů a pro metodu RP-RY 100 % až do okna W=19 nukleotidů. Nejlepších výsledků pro identifikaci do čeledí dosáhla metoda RP s hodnotou 99,31 % pro okna W=9 a W=11 nukleotidů.

7.2 VÝSLEDKY IDENTIFIKACE DO ČELEDÍ

Po otestování metod identifikace na stejných datech, která byla použita pro vytvoření referencí čeledí, byly navržené metody použity na datech odlišných. Použily se další DNA barcodingové projekty se sekvencemi ptáků z různých kontinentů. **Tab. 7.3** obsahuje výčet použitých projektů s počtem sekvencí a druhů v nich obsažených spolu s počtem sekvencí, které byly přiřazovány do čeledí. Tyto projekty byly staženy jako FASTA soubory se sekvencemi delšími než 600 bp z databáze BOLD Systems.

Tab. 7.3 Vybrané DNA barcodingové projekty pro identifikaci do čeledí.

Kód projektu	Název projektu	Počet sekvencí	Počet druhů	Ref. sekvencí	Ref. čeledí
KBBI	DNA Barcoding Korean Birds	254	102	180	18
TZBNA	Birds of North America	410	247	295	27
BNACA	Birds of North America – Canadian geese	137	2	137	1
BNABS	Birds of North America – Canadian passerines	114	37	102	7
BNAUS	Birds of North America – General sequences	1695	551	1165	30
BEPAL	Birds of the eastern Palearctic	1665	398	813	20
SWEBI	Birds of Scandinavia – Swedish birds	477	258	281	20
NORBI	Birds of Scandinavia – Norwegian birds	480	254	270	19

Ref. sekvencí – počet sekvencí patřících do čeledí majících referenci
Ref. čeledí – počet čeledí majících referenci

Všechny sekvence z vyjmenovaných projektů byly přiřazovány do referenčních čeledí vytvořených ze sekvencí souboru BARG pomocí metod konsenzuální sekvence, mediánu ND vektorů, průměru ND vektorů a průměru ND vektorů jen purinových nukleotidů. K vyhodnocení úspěšnosti identifikace do čeledí bylo nutné každou sekvenci zatřídit dle klasické taxonomie do řádů a čeledí. Sekvence, které patřily do jiného řádu, než které byly obsažené mezi referencemi, nebyly hodnoceny. Sekvence náležící k řádu, jehož alespoň jedna čeleď je referenční, byly hodnoceny na úspěšnost přiřazení alespoň do patřičného řádu.

V následujících kapitolách jsou popsány výsledky identifikace do čeledí a řádů jednotlivě pro analyzované soubory. Většina sekvencí v těchto souborech náleží jiným druhům ptáků, než ze kterých byly vytvořeny reference čeledí. **Tab. 7.4** shrnuje počty sekvencí z analyzovaných souborů, které odpovídají druhům, jež byly přítomny i v projektu BARG použitém pro tvorbu referencí. V souboru BNACA není žádný

společný druh zastoupen. Ostatní projekty obsahují několik sekvencí stejných druhů; jedná se o jednotky procent z celkového počtu sekvencí v souboru až na soubor KBBI, kde 15 % sekvencí je druhově shodných s referenčním souborem BARG. U těchto sekvencí lze předpokládat o něco menší úspěšnost identifikace než při verifikaci, neboť ptáci, ať druhů obsažených v referencích, jsou z jiného kontinentu odlišnější. Odlišnost je dána délkou trvání separace populací na kontinentech.

Tab. 7.4 Počty sekvencí stejných druhů jaké byly použity pro tvorbu referencí.

	Celkem		Společné s BARG	
Soubor	**Druhů**	**Jedinců**	**Druhů**	**Jedinců**
KBBI	66	180	8	27
TZBNA	174	295	14	30
BNACA	2	137	0	0
BNABS	32	102	2	6
BNAUS	394	1165	39	103
BEPAL	208	813	6	16
SWEBI	150	281	6	11
NORBI	150	270	6	9

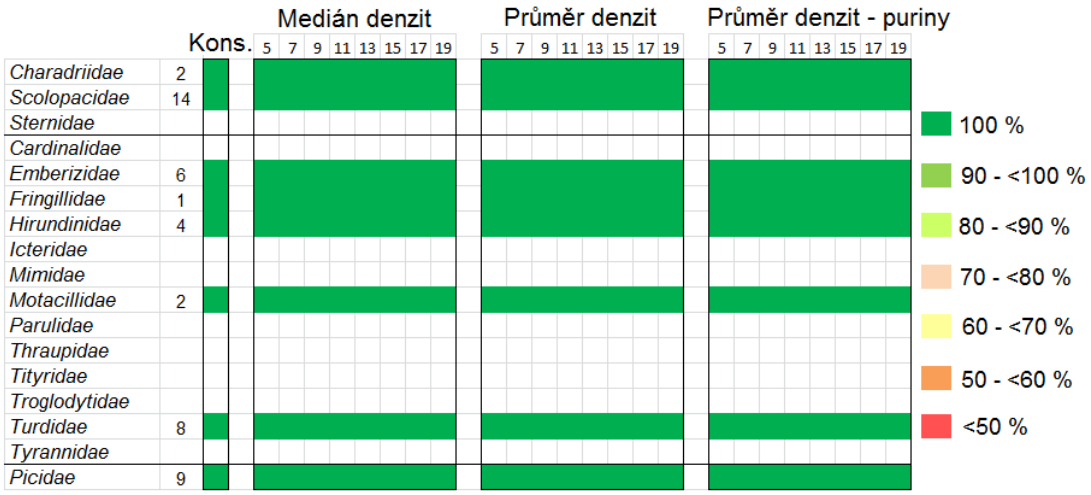
V případě identifikace čeledí mají pojmy druhově a rodově referenční sekvence jiný význam než v případě identifikace do druhů. Druhově referenční znamená, že analyzovaná sekvence patří k druhu, jehož sekvence jiného jedince byla součástí sekvencí, z kterých byla vytvořena reference čeledě. Rodově referenční pak znamená, že analyzovaná sekvence patří k druhu, který nebyl zastoupen mezi sekvencemi pro tvorbu reference čeledě, ale patřil tam jiný druh stejného rodu.

7.2.1 SOUBOR KBBI

Soubor KBBI (DNA Barcoding Korean Birds) obsahuje celkem 254 sekvencí, z nichž 180 náleží druhům patřícím do referenčních čeledí. Na **Obr. 7.6** je graficky znázorněna procentuální úspěšnost identifikace sekvencí do referenčních čeledí. Na **Obr. 7.7** je pak graficky znázorněna procentuální úspěšnost identifikace do řádů, která byla v případě všech metod 100 %.

V tomto souboru se vyskytuje 21 sekvencí čeledě Accipitridae 5 různých rodů, ale žádný z druhů neodpovídá druhům zastoupených v referencích. I přesto byly sekvence perfektně přiřazeny do správné čeledě. I u čeledě Falconidae, která má ze 12 sekvencí společnou s referencí jednu druhovou sekvenci, byly všechny sekvence také identifikovány bezchybně.

Obr. 7.6 Grafické znázornění procentuální úspěšnosti identifikace do čeledí sekvencí ze souboru KBBI.



Obr. 7.7 Grafické znázornění procentuální úspěšnosti identifikace do řádů sekvencí ze souboru KBBI.

Čeleď Anatidae má v souboru 21 sekvencí, z nichž 18 je rodově referenčních. Všechny sekvence byly správně identifikovány. Stejného dobrého výsledku bylo dosaženo u čeledě Ardeidae, která má v souboru 4 sekvence referenčního druhu, který však byl při tvorbě referencí zastoupen pouze jednou sekvencí. Ostatních 23 sekvencí této čeledě patří rodům nevyskytujících se v referenci až na 3 sekvence z rodu Egretta.

Dvě sekvence čeledě Caprimulgidae odpovídají jednomu z rodů obsažených v referenci avšak jiného druhu. Tyto dvě sekvence byly vždy správně identifikovány.

Čeleď Columbidae obsahuje v souboru 4 sekvence stejné druhu zastoupeného i v referenci jednou sekvencí a dalších 8 sekvencí, které nejsou v referenci zastoupeny ani rodově. I tak byla identifikace velmi úspěšná stejně jako pro devět sekvencí čeledě Picidae a jednu čeledě Podicipedidae.

Pouhá jedna sekvence v souboru patří čeledi Cuculidae, a i když druh, kterému sekvence náleží, není v referenci přítomen ani rodově, byla sekvence ve všech případech správně identifikována.

Dvě ze tří sekvencí čeledě Rallidae jsou druhově shodné s datasetem pro referenční čeleď a jejich identifikace byla všemi metodami bezchybná. Chyby v identifikaci nastaly při metodách RM a RP právě u sekvence, která se v sekvencích svým rodem nevyskytovala, avšak metoda purinové identifikace správně přiřadila i tuto sekvenci. Metody RM a RP tuto sekvenci pro delší okna přiřazovaly k čeledi Cuculidae.

V souboru byly ze 4 čeledí řádu Charadriiformes zastoupeny 2 a to čeleděmi Charadriidae a Scolopacidae. Dvě sekvence z prvně zmíněné čeledě jsou rodově přítomny v referenci. Druhá čeleď je v analyzovaném souboru zastoupená 13 sekvencemi 5 různých rodů, kde 7 těchto sekvencí je i rodově přítomno v referenci. U metody RM došlo u jedné sekvence čeledě Scolopacidae, jejíž druh nepatří k referenčním, k chybnému přiřazení k čeledě Charadriidae pro dvě nejdelší velikosti výpočetního okna a taktéž u purinové identifikace pro nejdelší okno.

U řádu Passeriformes byly metody RK, RM a RP málo úspěšné s čeleděmi Emberizidae, Fringillidae a Motacillidae. Jednalo se celkově o pouhých 9 sekvencí, které nepatřily k rodům, z nichž se tvořily reference. Metoda purinové identifikace RP-RY selhala pouze v případě jediné sekvence čeledě Fringillidae. Čeleď Hirundinidae byla zastoupena 4 sekvencemi jednoho referenčního druhu a tyto sekvence byly vždy správně přiřazeny. Čeleď Turdidae se 7 sekvencemi nereferenčního rodu a 1 sekvencí referenčního rodu dosáhla také 100% úspěšnosti pro všechny metody a velikosti výpočetního okna. Ve všech případech chybné čeleďové identifikace byly sekvence přiřazeny správně alespoň do řádu Passeriformes.

Osmnáct z 27 sekvencí čeledě Strigidae nebylo rodově zastoupeno v referenci. Metoda RK všechny sekvence druhu *Otus lempiji* přiřazovala do čeledě Podicipedidae. Metoda RM do délky okna W=13 špatně identifikovala pouze sekvence druhu *Ninox scutulata*, jehož rod není v referenci. Metoda RP od velikosti výpočetního okna W=13 měla 100% úspěšnost identifikace do správné čeledě a všechny nesprávné identifikace pro kratší okna se týkaly pouze sekvencí druhu *Ninox scutulata*. Tyto sekvence byly přiřazovány k čeledím Charadriidae nebo Accipitridae. Purinová identifikace byla bezchybná.

Soubor KBBI také obsahuje 62 sekvencí nereferenčních čeledí náležejícím k řádům, jejichž alespoň jedna čeleď patřila k referenčním. **Tab. 7.5** shrnuje průměry úspěšnosti identifikace těchto sekvencí alespoň do řádu. Metoda konsenzuální sekvence má nejnižší hodnoty úspěšnosti. U ostatních metod je zajímavá vysoká úspěšnost identifikace sekvencí nereferenčních čeledí do správného řádu Charadriiformes a Passeriformes. Tyto dva řády byly zastoupeny více než dvěma referenčními čeleděmi s větším počtem sekvencí. Řád Gruiformes byl zastoupen pouze jednou referenční čeledí Rallidae, což jsou menší chřástalovití ptáci, kdežto hodnocená sekvence patří do čeledě Gruidae, což jsou ptáci jeřábovití náležející do jiného podřádu. Metoda purinové identifikace přiřazovala sekvenci řádu Gruiformes vždy k čeledi Podicipedidae a v případě sekvencí řádu Passeriformes chybovala pouze u sekvencí jednoho druhu z čeledě Corvidae, které přiřazovala k čeledi Cerylidae z řádu Coraciiformes. Druhé dvě metody nebyly v chybném přiřazování tak konzistentní.

Tab. 7.5 Průměrné procentuální úspěšnosti identifikace do řádů pro sekvence nereferenčních čeledí souboru KBBI.

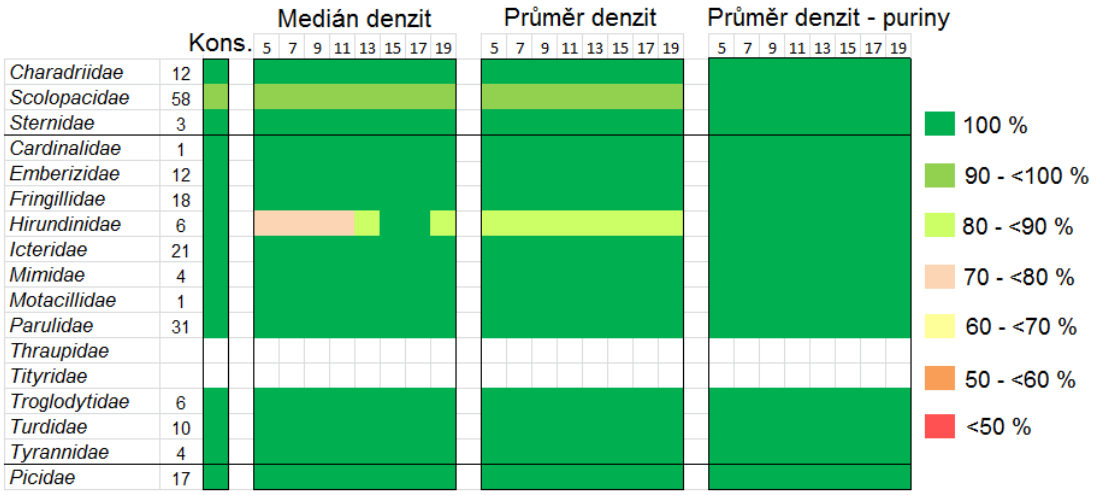
		Úspěšnost [%]			
Řád	**Počet**	**RK**	**RM**	**RP**	**RP-RY**
Gruiformes	1	0,00	0,00	0,00	0,00
Charadriiformes	5	60,00	96,92	64,62	100,0
Passeriformes	56	92,86	99,31	99,31	94,37

7.2.2 SOUBOR TZBNA

Soubor TZBNA (Birds of North America) obsahuje celkem 410 sekvencí, z nichž 295 náleží druhům patřícím do referenčních čeledí. Na **Obr. 7.8** je graficky znázorněna úspěšnost identifikace těchto sekvencí do čeledí a na **Obr. 7.9** pak do řádů.

Čeleď Accipitridae je v souboru zastoupena celkem 17 sekvencemi z 10 rodů, přičemž 12 sekvencí náleží k rodům referenčním. Žádná z 32 sekvencí čeledě Anatidae nepatří k referenčním druhům, ale 14 z nich je referenčních rodově. Pouhé 3 sekvence v souboru jsou z čeledě Caprimulgidae a patří stejnému rodově referenčnímu druhu. Čeleď Ardeidae i Columbidae jsou zastoupeny také po třech sekvencích různých druhů i rodů a všechny jsou rodem referenční a čeleď Cuculidae má v souboru pouze jednu rodově referenční sekvenci. Čeleď Falconidae má v souboru sekvencí rodu Falco a 5 z nich přímo odpovídá druhům obsaženým v referenci. Čeleď Rallidae je v souboru zastoupena 3 sekvencemi, z nichž jedna je rodově přítomna v referenci. Všechny tyto zmíněné čeledě dosáhly u všech metod a velikostí výpočetního okna 100% úspěšnosti identifikace do čeledí až na jednu špatně identifikovanou sekvenci čeledě Ardeidae pro metodu RP a okno W=5 nukleotidů.

Obr. 7.8 Grafické znázornění procentuální úspěšnosti identifikace do čeledí sekvencí ze souboru TZBNA.



Obr. 7.9 Grafické znázornění procentuální úspěšnosti identifikace do čeledí sekvencí ze souboru TZBNA.

Čeleď Charadriidae má v souboru 10 sekvencí jednoho rodu, který byl v referenci zastoupen 4 sekvencemi různých druhů. Čeleď Sternidae má jen 3 sekvence a dvě z nich nejsou referenční ani rodem. U obou čeledí bylo dosaženo perfektní identifikace u všech metod nezávisle na délce výpočetního okna. Méně úspěšná byla identifikace pro čeleď Scolopacidae. Reference této čeledě byla vytvořena z pouhých 6 sekvencí tří rodů. V souboru TZBNA je 58 sekvencí této čeledě celkem, z toho 29 jich rodově

odpovídá referenci a z nich jich 14 dokonce je shodných druhově. Chyby v identifikaci do čeledí se týkaly sekvencí rodově nezastoupených v referenci. Nejúspěšnější v identifikaci sekvencí této čeledě byla metoda RP-RY, pro okna do délky W=17 nukleotidů byla metoda bezchybná.

U obsáhlého řádu Passeriformes bylo pro některé čeledě dosaženo výborných výsledků identifikace, u některých zase byly výsledky velmi špatné. Čeleď Cardinalidae je v souboru zastoupena pouze jednou rodově nereferenční sekvencí. Metoda RK si s touto sekvencí neporadila a metoda RM ji byla schopna správně určit pouze při nejdelším výpočetním okně, avšak sekvence byla alespoň přiřazena do správného řádu. Metoda RP ji špatně identifikovala jen jednou a metoda RP-RY byla úspěšná vždy. Dvanáct sekvencí čeledě Emberizidae také není rodem referenční a metoda RP s nimi byla úspěšnější než RM, RP-RY a RK. Všechny metody byly málo úspěšné s čeledí Fringillidae. Dvě sekvence čeledě Hirundinidae jsou druhově referenční, dvě rodově a další dvě nebyly v referenci zastoupené. Metody RM i RP špatně identifikovaly do čeledě pouze ty dvě referenci neznámé sekvence.

Patnáct z 21 sekvencí čeledě Icteridae je rodově referenčních (dvě i druhově). Všechny identifikační chyby u všech metod se týkají rodově referenčních sekvencí. Čeleď Mimidae obsahuje v souboru TZBNA pouze sekvence, které nemají v referenci zastoupení ani rodově. Několik chyb v identifikaci patřilo k sekvencím rodu *Dumetella,* druhý rod *Toxostoma* byl identifikován bezchybně. V souboru je z čeledě Motacillidae pouze jedna rodově referenční sekvence, která byla bezchybně identifikována. Z 31 sekvencí čeledě Parulidae jsou pouze tři rodově zastoupeny v referenci a i přesto byly všechny metody s těmito sekvencemi úspěšné.

Všechny sekvence čeledě Troglodytidae mají v referenci rodového zástupce, stejně tak čeleď Turdidae až na 3 sekvence z 10, naopak z čeledě Tyrannidae má rodového zástupce v referenci 1 sekvence ze 4. U všech těchto čeledí bylo dosaženo 100% správné identifikace do čeledí všemi třemi metodami.

Ze 17 sekvencí čeledě Picidae je pouze jedna rodově nereferenční a identifikace byla bezchybná stejně jako téměř ve všech případech dvou druhově referenčních sekvencí čeledě Podicipedidae a čtyř rodově nereferenčních sekvencí čeledě Psittacidae.

Šest sekvencí ze 14 čeledě Strigidae nemá rodového zástupce v referenci. Metody RM a RP měly nižší míru úspěšnosti identifikace pro nereferenční sekvence druhu *Aegolius funereus* a rodově referenční druh *Glaucidium*, naproti tomu metody RK a RP-RY byly bezchybné.

Čeleď Trochillidae má v souboru tři sekvence, které nemají v referenci rodové zastoupení. I přesto byly všemi metodami správně identifikovány.

Co se týká identifikace nereferenčních čeledí referenčních řádů, výsledky jsou shrnuty v **Tab. 7.6**. Řád Charadriiformes obsahuje 38 sekvencí druhů z nereferenčních čeledí a řád Passeriformes 52 sekvencí. Až na pár případů, především u metody RK,

byly tyto sekvence z řádu Charadriiformes správně zatříděny do tohoto řádu. Jednalo se nejvíce o sekvence z čeledí Alcidae a Laridae, které patří do stejného podřádu jako referenční čeleď Sternidae, avšak sekvence nebyly častěji přiřazovány k této nejbližší referenci. U řádu Passeriformes patří sekvence převážně k čeledím Corvidae a Vireonidae, u kterých se také neprojevila výrazná inklinace k nějaké čeledi. Sekvence řádu Gruiformes patří stejnému rodu jako sekvence v předchozím případě souboru KBBI a stejně tak nebyla ani v jednom případě správně klasifikována. Nereferenční sekvence řádu Falconiformes patří druhu *Pandion haliaetus* čeledě *Pandionidae*. Reference obsahovala pouze sekvence čeledě Falconidae. Sekvence nebyla ani v jednom případě klasifikována ani do velmi blízkého řádu Accipitriformes.

Tab. 7.6 Průměrné úspěšnosti identifikace do řádů pro sekvence nereferenčních čeledí souboru TZBNA.

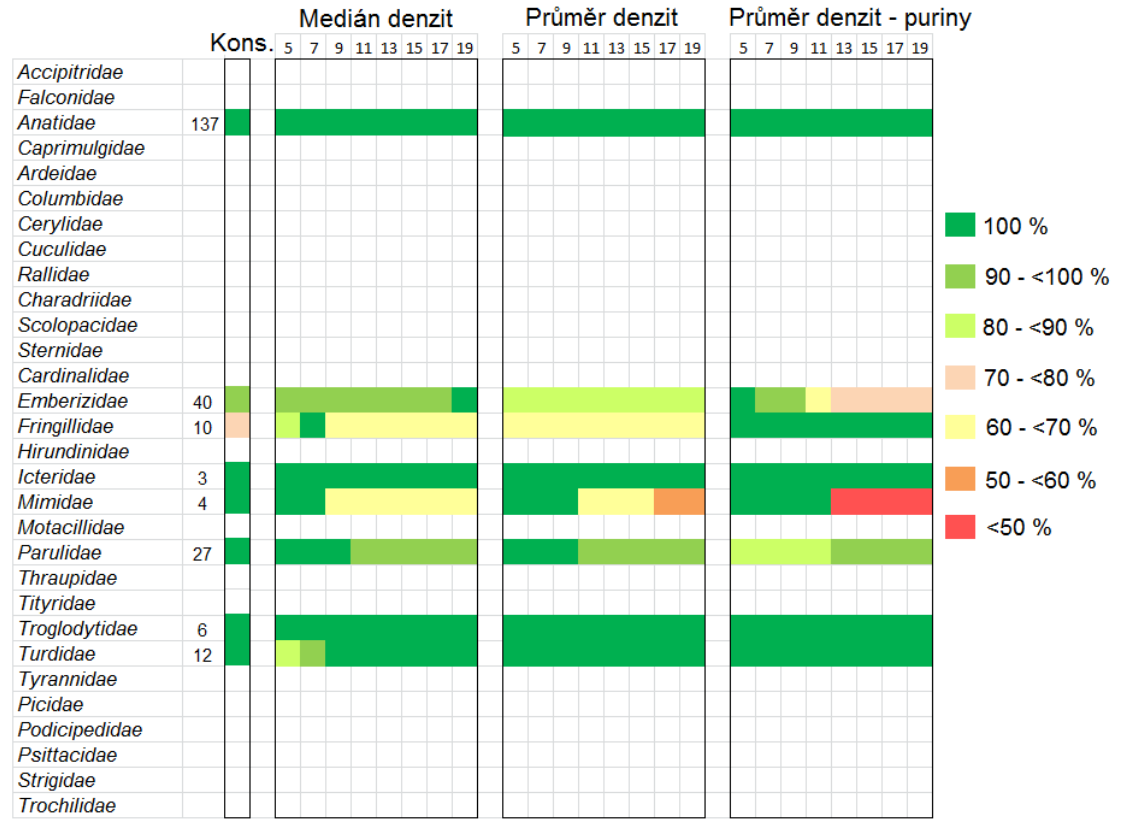
		Úspěšnost [%]			
Řád	**Počet**	**RK**	**RM**	**RP**	**RP-RY**
Gruiformes	1	0,00	0,00	0,00	0,00
Charadriiformes	38	73,68	92,71	96,56	100,0
Passeriformes	52	80,77	82,84	92,01	93,20
Pelecaniformes	3	0,00	0,00	0,00	0,00

7.2.3 SOUBOR BNACA A BNABS

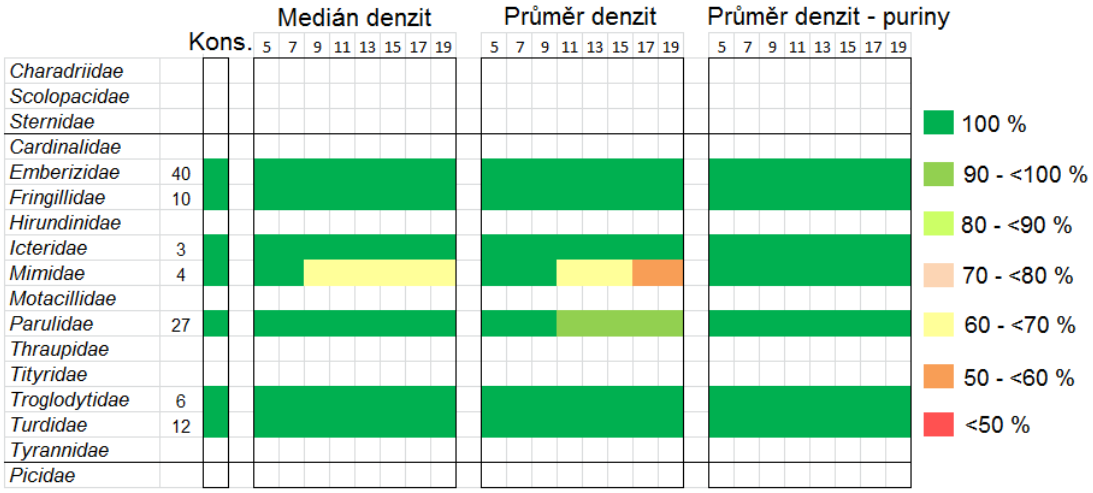
Soubor BNACA (Birds of North America – Canadian geese) obsahuje 137 sekvencí vrubozobých ptáků a to pouze dvou druhů bernešky velké (*Branta canadensis*) a bernešky aljašské (*Branta hutchinsii*). Všechny čtyři metody identifikovaly tyto sekvence do správné čeledě neomylně, přičemž rod bernešek není v referenčních sekvencích obsažen.

Soubor BNABS (Birds of North America – Canadian passerines) obsahuje 114 sekvencí řádu Passeriformes, kde 12 z nich náleží nereferenčním čeledím. Z toho 40 sekvencí patří čeledi Emberizidae šesti různých rodů, přičemž pouze jeden z nich je zastoupen v referenci avšak jinými druhy. Všechny sekvence tohoto rodu byly vždy správně identifikovány. U ostatních sekvencí docházelo k chybám v identifikaci do čeledě, přičemž nejhorší výsledky byly pro metodu RP-RY, ale identifikace do řádu byla bezchybná. Na **Obr. 7.10** je graficky znázorněna úspěšnost identifikace těchto sekvencí do čeledí a na **Obr. 7.11** pak do řádů.

Z čeledě Fringillidae jsou v souboru zastoupeny rody *Carduelis* a *Carpodacus*. Prvně zmiňovaný rod je zastoupen i v referenci avšak jinými druhy, ten druhý zastoupen není. Metoda purinové identifikace správně zařadila všechny sekvence, nejhorší výsledky dala metoda konsenzuální sekvence.



Obr. 7.10 Grafické znázornění procentuální úspěšnosti identifikace do čeledí sekvencí ze souboru BNACA a BNABS.



Obr. 7.11 Grafické znázornění procentuální úspěšnosti identifikace do řádů sekvencí ze souboru BNACA a BNABS.

Čeleď Icteridae má v souboru BNABS pouze tři sekvence rodu *Molothrus*, který je i součástí reference, a všechny byly správně přiřazeny. Zajímavý je postupný pokles úspěšnosti identifikace pro čeleď Mimidae u metod RM a RP a ostrý přechod od 100 % k 0 % u RP-RY; bezchybné nebylo ani přiřazení k řádu. Jedná se však o pouhé 4 sekvence rodu *Dumetella*. Reference čeledě Mimidae je tvořena pouze sekvencemi rodu *Mimus*.

Ptáci čeledě Parulidae jsou zastoupeny 27 sekvencemi 7 rodů, z nichž ani jeden není součástí reference. I přesto byly tyto sekvence identifikovány s vysokou úspěšností, i když u purinové identifikace byla úspěšnost nižší, ale neklesla pod 88 %. Soubor také obsahuje 6 sekvencí řádu Troglodytidae dvou druhů rodu *Troglodytes*. Jeden z nich je zastoupen i v referenci. Všechny sekvence byly vždy správně identifikovány stejně jako sekvence čeledě *Turdidae*, která má v souboru jen sekvence rodů *Catharus* a *Turdus*, jež jsou také oba obsaženy v referenci, přičemž 3 sekvence odpovídají i referenčnímu druhu.

Dvanáct sekvencí ze souboru BNABS patří nereferenčním čeledím Certhidae, Paridae, Regulidae a Vireonidae řádu Passeriformes. Průměrná úspěšnost identifikace těchto sekvencí do řádu metodou RK 100 %, metodou RM byla 85,90 %, metodou RP 91,67 % a metodou RP-RY 80,00 %.

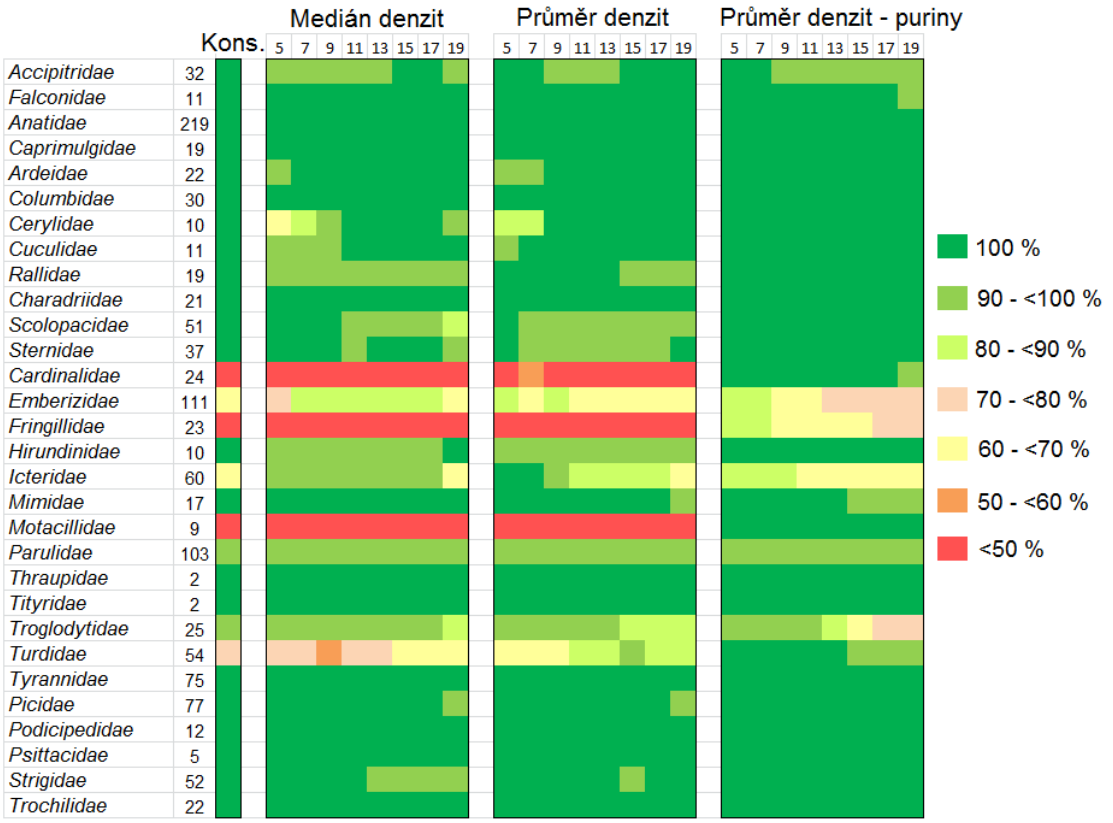
7.2.4 SOUBOR BNAUS

Soubor BNAUS (Birds of North America – General sequences) obsahuje celkem 1695 sekvencí, kde 155 z nich patří do nereferenčních řádů a 375 do referenčních řádů ale nereferenčních čeledí, takže zbývá 1165 sekvencí pro identifikaci. Na **Obr. 7.12** je graficky znázorněna úspěšnost identifikace těchto sekvencí do čeledí a na **Obr. 7.13** pak do řádů.

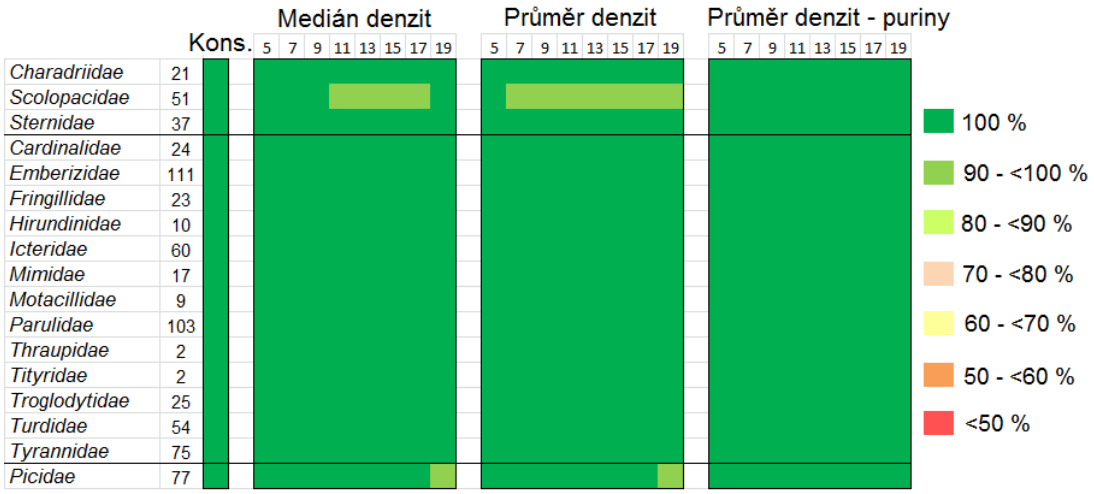
Čeleď Accipitridae má v souboru 32 sekvencí, kde pouze tři z nich nemají rodového zástupce v referenci a dokonce 6 jich je druhově referenčních. Veškeré chyby v identifikaci se týkaly pouze dvou sekvencí druhu *Elanus leucurus*, který byl kupodivu druhově referenční. Všechny sekvence čeledě Falconidae patří rodu *Falco*, který je referenční, a skoro polovina sekvencí je druhově shodná s referenčními sekvencemi. Metoda RP-RY chybovala u delších výpočetních oken, kdežto druhé dvě spíše u kratších oken. Nesprávně přiřazované sekvence obou řádů nebyly přiřazovány recipročně, což svým způsobem podporuje nové poznatky, že tyto skupiny dravých ptáků nejsou blízce příbuzné.

Z 219 sekvencí čeledě Anatidae jich 60 rodem odpovídá referenčním. I přes velké množství rodově nereferenčních sekvencí byly všechny metody v identifikaci bezchybné. Dále bylo bezchybně identifikováno všech 19 sekvencí čeledě Caprimulgidae, z nichž 10 je druhově referenčních a 2 rodově referenční, a 3 rodově referenční sekvence čeledě Columbidae. U čeledě Ardeidae jen 3 sekvence z 22 nemá rodovou referenci, přičemž 8 sekvencí má druhovou referenci a identifikace byla téměř bezchybná, jen pro nejkratší okna se vyskytlo pár chyb pro metody RM a RP.

Sedm z 10 sekvencí čeledě Cerylidae je druhově referenčních. Metody RM a RP byla při identifikaci s krátkými výpočetními okny méně úspěšná s druhově nereferenčními sekvencemi, naproti tomu metody RK a RP-RY byly vždy 100%. Podobně to platilo i pro sekvence čeledě Cuculidae a Rallidae, které jsou z větší části rodově a některé i druhově referenční.



Obr. 7.12 Grafické znázornění procentuální úspěšnosti identifikace do čeledí sekvencí ze souboru BNAUS.



Obr. 7.13 Grafické znázornění procentuální úspěšnosti identifikace do řádů sekvencí ze souboru BNAUS.

Všechny sekvence v souboru čeledě Charadriidae jsou minimálně rodově referenční, a proto byla vysoká úspěšnost identifikace očekávána a bylo jí i dosaženo pro všechny metody a délky výpočetního okna. U čeledě Sternidae je přes polovinu sekvencí rodově či druhově referenčních. Chyby v identifikaci se vyskytovaly pouze u jedné nereferenční sekvence pro metody RM a RP, která však byla alespoň dobře identifikována do řádu. Čeleď Scolopacidae má 27 z 51 sekvencí v souboru minimálně

rodově referenční. Všechny chybně identifikované sekvence této čeledě patřily k nereferenčním sekvencím a metody RM a RP chybovaly i v identifikaci do řádu.

Naproti tomu identifikace do řádu Passeriformes byla pro všech 515 sekvencí vynikající. Naopak identifikace do čeledí byla méně úspěšná. Čeleď Cardinalidae obsahuje v souboru BNAUS 24 sekvencí nemající v referenci rodového zástupce. Výsledky identifikace těchto sekvencí do čeledě byly pro většinu metod velmi špatné: RK < 5 %, RM 4 % – 33 %, a RP 33 % – 50 %; naopak metoda purinové identifikace RP-RY správně přiřadila všechny sekvence pro délky výpočetního okna W=5 až W=17 nukleotidů. Čeleď Emberizidae má v souboru 111 sekvencí, z nichž pouze 15 je rodově referenčních. U žádné z metod nepřesáhla úspěšnost klasifikace 90 %. Další neúspěšně identifikovanými čeleděmi byla čeleď Fringillidae, jejíž dvě třetiny sekvencí také nejsou referenční. Mezi těmito čeleděmi docházelo k nesprávnému přiřazování k čeledím Icteridae a Thraupidae, které patří do stejné skupiny zpěvných ptáků (nine-primaried oscines) nadčeledě Passeroidea. Z 60 sekvencí čeledě Icteridae nejsou dvě třetiny ani rodově referenční a nejméně úspěšné s jejich rozlišením byly metody RK a RP-RY. Jen dvě sekvence jednoho druhu patří v souboru čeledi Thraupidae. Do stejné skupiny patří ještě čeleď Parulidae, která má v souboru 103 sekvencí, z nichž jen 9 je rodově referenčních. Všechny metody identifikovaly tyto sekvence s úspěšností nad 90 %, kdy většina chyb patřila dvěma sekvencím jednoho druhu. Do nadčeledě Passeroidea patří ještě čeleď Motacillidae, s jejímiž sekvencemi si metody RK, RM a RP neporadily a přiřazovaly je výše zmíněným čeledím skupiny Passeroidea, kdežto metoda RP-RY byla bezchybná. Sekvence čeledě Hirundinidae jsou více než z poloviny rodově referenční. Metody RM a RP byly stejně úspěšné ale ne bezchybné, metody RK a RP-RY byly 100%. Ačkoliv jen 2 ze 17 sekvencí čeledě Mimidae jsou rodově referenční, metody RK a RM v identifikaci nechybovaly, ostatní dvě metody chybovaly několikrát pouze u jedné sekvence. Čeleď Tityridae má v souboru jen dvě rodově referenční sekvence a identifikace byla bezchybná. Čeleď Troglodytidae má 25 sekvencí, kde jsou 4 druhově referenční a ostatní nejsou ani rodově referenční. Zajímavé je, že všechna nesprávná přiřazení byla k čeledím z podskupiny Passeroidea, kam však Troglodytidae nepatří. Ze 75 sekvencí čeledě Tyrannidae jich 33 není ani rodově referenčních. Oproti sekvencím čeledě Turdidae s podobným poměrem referenčních a nereferenčních sekvencích, byly tyto sekvence identifikovány bezchybně.

V souboru je dále 77 sekvencí 5 různých rodů čeledě Picidae, přičemž 4 rody jsou referenční. U metody RM byly pro okno W=19 nukleotidů chybně identifikovány sekvence referenčního rodu *Melanerpes*, u kterých chybovala i metoda RP.

Třetina sekvencí z 12 čeledě Podicipedidae není ani rodově referenční. Všechny metody identifikovaly tyto sekvence bezchybně stejně jako 5 sekvencí čeledě Psittacidae, z nichž jsou tři druhově referenční a dvě nejsou referenční ani rodově, a 22 sekvencí čeledě Trochillidae, kde není ani jedna rodově referenční.

Z celkového počtu 52 sekvencí čeledě Strigidae jich 5 nemá rodového zástupce v referenci. Sekvence byly přiřazeny téměř bezchybně kromě pár případů u sekvencí druhu *Aegolius acadicus* (nereferenční) a *Asio flammeus* (druhově referenční).

V souboru BNAUS bylo i mnoho sekvencí patřících do referenčních řádů ale nereferenčních čeledí. **Tab. 7.7** shrnuje průměrné úspěšnosti identifikace takovýchto sekvencí do správného řádu. V souboru jsou 4 sekvence čeledě Pandionidae z řádu Accipitriformes. Metoda RP byla schopná tyto nereferenční sekvence zatřídit téměř v 79 % případů, kdežto nejhorších výsledků dosáhla metoda RP-RY s 15,38 %.

Nereferenční sekvence řádu Gruiformes patří do jiného podřádu než referenční čeleď Rallidae, kam patří vzrůstově menší chřástalovití ptáci. Nereferenční sekvence patřily velkým jeřábovitým ptákům. Velké morfologické rozdíly mezi těmito čeleděmi jsou pravděpodobně v korelaci s rozdíly mezi sekvencemi a to vedlo k nesprávné identifikaci do řádu ve všech případech.

Dále se v souboru vyskytuje vysoký počet sekvencí nereferenčních čeledí řádu Charadriiformes. Převážná část sekvencí patří čeledím Alcidae a Laridae, které jsou nejblíže příbuzné referenční čeledi Sternidae. Všechny metody byly schopny nereferenční sekvence identifikovat do správného řádu s vysokou úspěšností, metoda RP-RY nesprávně přiřadila jen dvě sekvence pro nejdelší výpočetní okno.

Z řádu Passeriformes je sekvencí z 19 nereferenčních čeledí celkem 199. Všechny tři metody byly s identifikací do řádu vcelku úspěšné. Naopak málo úspěšné byly metody při identifikaci 45 nereferenčních sekvencí řádu Pelecaniformes. U řádu Strigiformes figurovaly 3 sekvence čeledě Tytonidae, které metoda RP byla až na pár případů správně přiřadila.

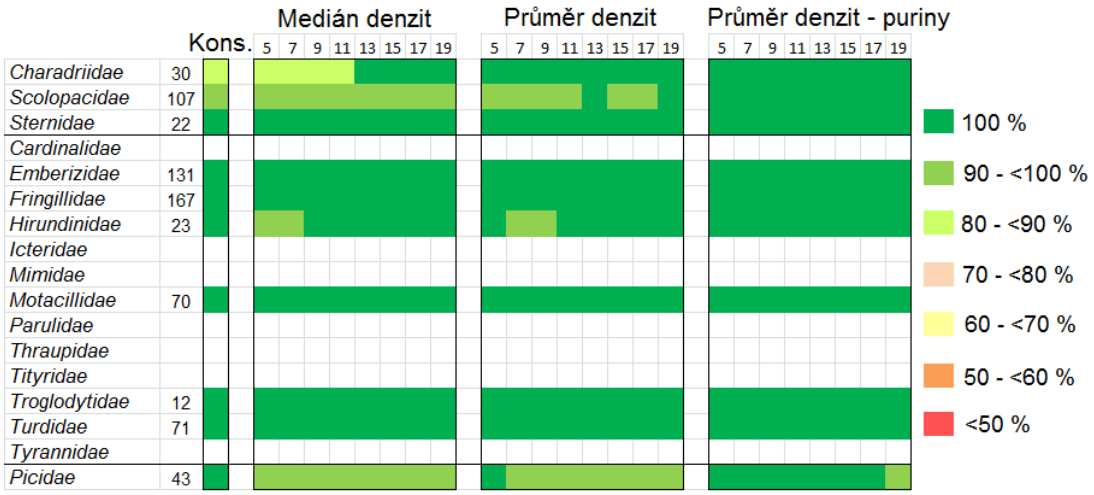
Tab. 7.7 Průměrné úspěšnosti identifikace do řádů pro sekvence nereferenčních čeledí souboru BNAUS.

		Úspěšnost [%]			
Řád	**Počet**	**RK**	**RM**	**RP**	**RP-RY**
Accipitriformes	4	100,0	67,31	78,85	15,38
Gruiformes	6	0,00	0,00	0,00	0,00
Charadriiformes	111	64,86	96,12	97,09	99,86
Passeriformes	199	90,45	90,07	95,36	97,68
Pelecaniformes	45	0,00	8,89	8,21	5,13
Strigiformes	3	33,33	38,46	94,87	33,33

7.2.5 SOUBOR BEPAL

Soubor BEPAL (Birds of the eastern Paleartic) obsahuje celkem 1665 sekvencí, přičemž sekvencí referenčních čeledí je 813. Na **Obr. 7.14** je graficky znázorněna úspěšnost identifikace těchto sekvencí do čeledí a na **Obr. 7.15** pak do řádů.

Obr. 7.14 Grafické znázornění procentuální úspěšnosti identifikace sekvencí do čeledí ze souboru BEPAL.



Obr. 7.15 Grafické znázornění procentuální úspěšnosti identifikace sekvencí do řádu ze souboru BEPAL.

Čeleď Accipitridae má v souboru 12 sekvencí a všechny jsou rodově referenční a čeleď Falconidae má v souboru 12 sekvencí jednoho referenčního rodu 5 různých druhů. Z 37 sekvencí čeledě Anatidae jich 24 nemá rodového zástupce v referenci. Čeleď Caprimulgidae má v souboru 6 sekvencí dvou druhů stejného referenčního rodu. U čeledě Ardeidae jsou 3 sekvence z 5 druhově referenční, čeleď Cuculidae je v souboru zastoupena 9 sekvencemi dvou druhů rodu *Cuculus*, který nebyl použit pro

tvorbu referencí, a čeleď Columbidae má 11 rodově referenčních sekvencí z 19. Všechny sekvence těchto čeledí byly bezchybně identifikovány všemi čtyřmi metodami při všech délkách výpočetního okna.

Čeleď Rallidae má v souboru pouhých 5 sekvencí, přičemž jedna je druhově a jedna rodově referenční. Chybně byly identifikovány metodami RM a RP pouze nereferenční sekvence při některých délkách výpočetního okna.

Všech 30 sekvencí čeledě Charadriidae je rodově referenčních. Metoda RK a metoda RM při kratších výpočetních oknech špatně identifikovaly nejen do čeledě ale i do řádu sekvence druhu *Charadrius dubius*; ostatní dvě metody byly bezchybné. Čeleď Scolopacidae má v souboru 107 sekvencí, z nichž 52 je rodově referenčních. Všechny metody chybovaly pouze v identifikaci nereferenčních sekvencí, především v rodech *Limosa, Numenius* a *Xenus*, přičemž metoda RP-RY chybovala pouze se sekvencemi rodu *Numenius*, ale přiřadila je oproti ostatním dvěma metodám alespoň k příbuzné čeledi Charadriidae. Přesně polovina z 22 sekvencí čeledě Sternidae je rodově referenční a žádná z metod u těchto sekvencí nechybovala.

Čeleď Emberizidae je v souboru zastoupena 131 sekvencemi, z nichž však ani jedna není ani rodově referenční, a jejich identifikace byla značně problémová pro všechny metody. Podobně málo úspěšná byla identifikace sekvencí čeledě Fringillidae. Ze 167 sekvencí jich 57 je rodově referenčních, z nichž některé byly také špatně identifikovány, především druhu *Carduelis flammmea*. Stejně jako v případě předcházejícího souboru, byly sekvence těchto čeledí přiřazovány k blízce příbuzným čeledím Icteridae a Thraupidae. O něco málo lépe dopadla identifikace 23 sekvencí čeledě Hirundinidae, z nichž 5 je druhově referenčních I když úspěšnost metod RK, RM a RP byla vcelku nízká, purinová identifikace byla naopak bezchybná. Sekvence čeledě Motacillidae byly také identifikovány s malou úspěšností až na metodu RP-RY. Přes polovinu sekvencí je rodově referenčních a tyto sekvence byly většinou identifikovány správně. Poslední v souboru přítomné čeledě řádu Passeriformes jsou Troglodytidae a Turdidae. Prvně zmíněný má v souboru 12 sekvencí jednoho rodově referenčního druhu. Kromě případů s nejkratšími výpočetními okny pro RM a RP byly metody v identifikaci 100% úspěšné. Druhá zmíněná čeleď má v souboru 71 sekvencí 12 druhů jednoho referenčního rodu a byly vždy bezchybně identifikovány.

Čeleď Picidae je v souboru zastoupena 43 sekvencemi, kde 10 jich je rodově referenčních. Metoda RM chybovala pouze v jedné sekvenci druhu *Jynx torquilla,* stejně tak metoda RP, kromě případu s nejkratším výpočetním oknem, kde i tuto sekvenci identifikovala dobře, a metoda RP-RY chybovala s touto sekvencí pro okno W=19 nukleotidů.

Ze 7 sekvencí čeledě Podicipedidae není pouze jedna rodově referenční a všechny sekvence byly správně klasifikovány všemi metodami.

Ze 17 sekvencí čeledě Strigidae jsou 3 druhově referenční a 7 rodově referenční. Metoda RK chybovala ve dvou rodově referenčních sekvencí, metoda RM byla

v identifikaci perfektní do velikosti výpočetního okna *W*=17 nukleotidů včetně, následně chybovala v jedné nereferenční sekvenci. Metody RP a RP-RY byly bezchybné.

Z hlediska faktu, že byly analyzovány sekvence euroasijských populací ptáků, přičemž reference byly vytvořeny z jihoamerických populací, tak je výsledek identifikace do čeledí/řádů vynikající především pro metodu purinové identifikace.

Soubor obsahuje množství čeleďově nereferenčních sekvencí především z řádu Passeriformes, kterých je dokonce více než sekvencí odpovídajících referenčním čeledím, viz **Tab. 7.8**. Konkrétně těchto sekvencí bylo 732 a všechny metody identifikace dokázaly tyto sekvence přiřazovat k příbuzným čeledím s vysokou mírou úspěšnosti nad 96 %. Stejně vysoce úspěšné byly metody, kromě RK, s přiřazováním k příbuzným čeledím 39 sekvencí řádu Charadriiformes, kde metoda purinové identifikace byla dokonce při všech délkách výpočetního okna bezchybná. Nulovou úspěšnost vykázaly všechny tři metody se sekvencemi ostatních řádů.

Tab. 7.8 Průměrné úspěšnosti identifikace do řádů pro sekvence nereferenčních čeledí souboru BEPAL.

		Úspěšnost [%]			
Řád	**Počet**	**RK**	**RM**	**RP**	**RP-RY**
Columbiformes	4	0,00	0,00	0,00	0,00
Coraciiformes	22	0,00	0,00	0,00	0,00
Gruiformes	1	0,00	0,00	0,00	0,00
Charadriiformes	39	33,33	93,69	89,15	100,0
Passeriformes	732	98,63	96,06	97,16	98,79
Pelecaniformes	2	0,00	0,00	0,00	0,00

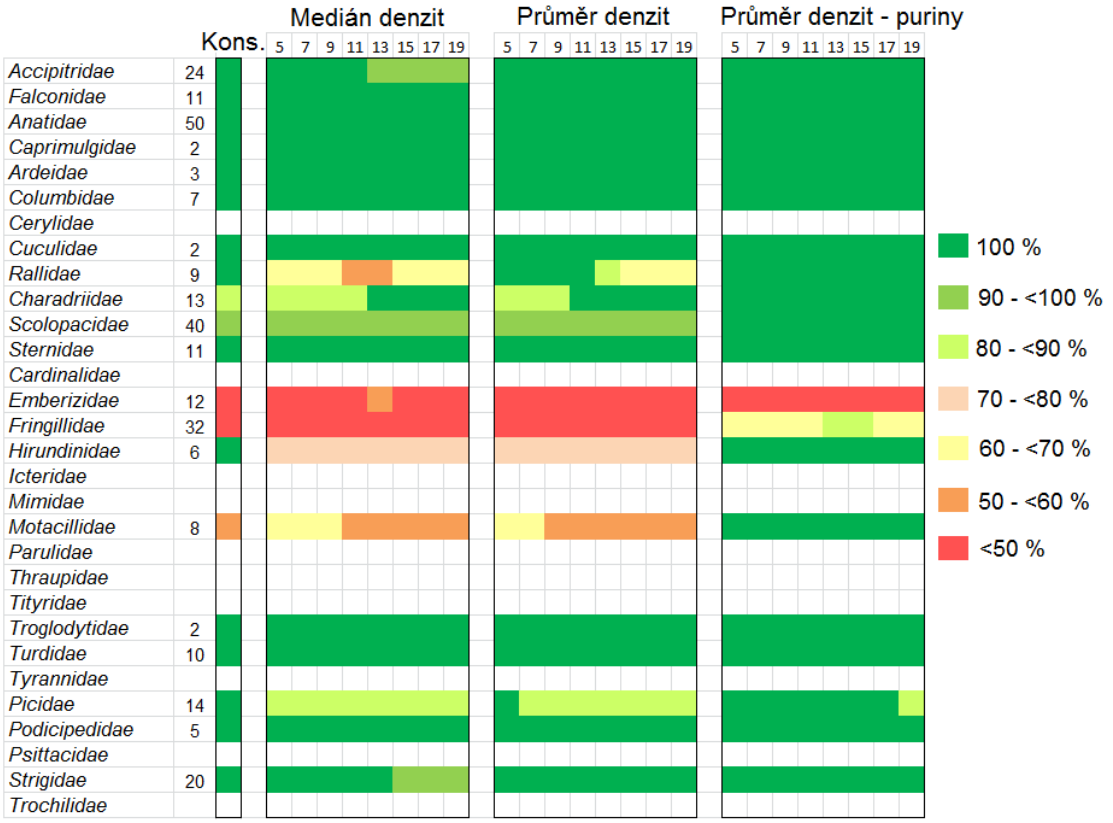
7.2.6 SOUBOR SWEBI

Soubor SWEBI (Birds of Scandinavia – Swedish birds) obsahuje celkem 477 sekvencí, z nichž 25 není ani řádově referenčních. Sekvencí náležejících do referenčních čeledí je celkem jen 281. Na **Obr. 7.16** je graficky znázorněna úspěšnost identifikace těchto sekvencí do čeledí a na **Obr. 7.17** pak do řádů.

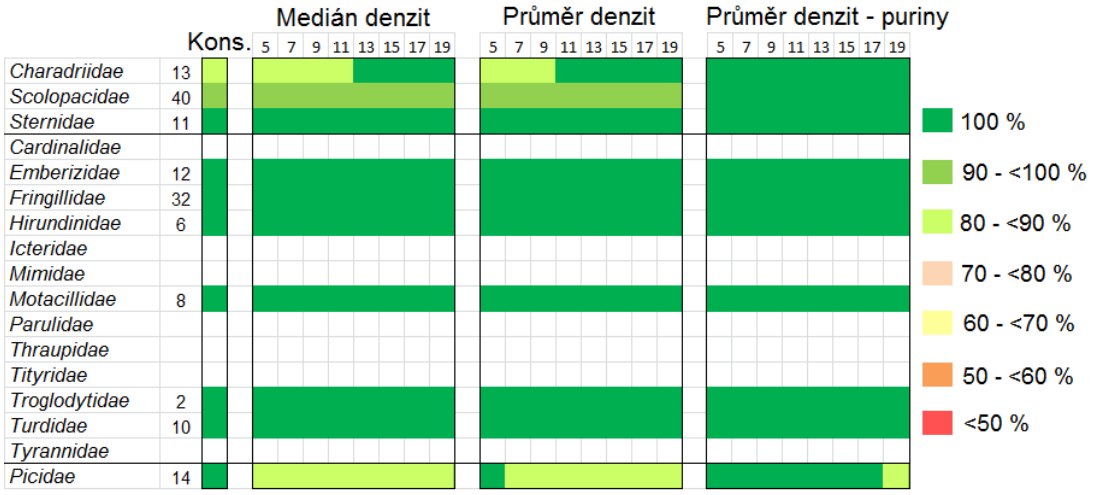
Přes polovinu z 24 sekvencí čeledě Accipitridae je rodově referenčních. Metoda RM chybovala pouze v jedné až dvou sekvencích nereferenčního druhu *Pernis apivorus* při delších výpočetních oknech.

Žádná z metod v identifikaci do čeledí nechybovala u následujících čeledí. Jedenáct sekvencí čeledě Falconidae náleží 6 druhům referenčního rodu Falco. Z 50 sekvencí čeledě Anatidae jich 20 je rodově referenčních. Čeleď Caprimulgidae má v souboru pouze dvě sekvence, čeleď Ardeidae tři sekvence a čeleď Columbidae 7 sekvencí, které

jsou všechny rodově referenční až na dvě sekvence z posledně zmíněné čeledě. Čeleď Cuculidae má v souboru dvě nereferenční sekvence.



Obr. 7.16 Grafické znázornění procentuální úspěšnosti identifikace do čeledí sekvencí ze souboru SWEBI.



Obr. 7.17 Grafické znázornění procentuální úspěšnosti identifikace do řádů sekvencí ze souboru SWEBI.

Z 9 sekvencí čeledě Rallidae jsou 4 rodově referenční. Metody RM a RP chybovaly v nereferenčních sekvencích. Všech 13 sekvencí čeledě Charadriidae je rodově referenční. U všech metod patřila většina špatně identifikovaných sekvencí k druhu

Charadrius dubius. Přes polovinu sekvencí čeledě Scolopacidae je také rodově referenčních a všechny metody špatně identifikovaly především sekvence druhu *Numenius arquata*. V případě sekvencí těchto dvou čeledí nebyla pro metody RK, RM a RP bezchybná ani identifikace do řádu. U čeledě Sternidae jich 6 z 11 je rodově referenčních a jedna druhově referenční. Žádná z metod při identifikaci těchto sekvencí nechybovala.

Sekvencí z řádu Passeriformes odpovídajících referenčním čeledím je vcelku málo a stejně jako v případě předchozích souborů, vykazují sekvence čeledí Emberizidae a Fringillidae velmi nízkou úspěšnost identifikace, neboť jsou přiřazovány k jiným blízce příbuzným čeledím. Taktéž čeledě Hirundinidae a Motacillidae byly metodami RM a RP přiřazovány s nízkou úspěšností oproti metodě RP-RY, která přiřazovala sekvence bezchybně. Všechny metody byly maximálně úspěšné se sekvencemi čeledí Troglodytidae a Turdidae.

Ze 14 sekvencí čeledě Picidae jsou jen 4 z nich rodově referenční. Veškeré chyby při identifikaci se týkaly pouze dvou sekvencí nereferenčního druhu *Jynx torquilla*, jako tomu bylo i v souboru BEPAL. Všech 5 sekvencí čeledě Podicipedidae je rodově referenčních a byly bezchybně identifikovány. U čeledě Strigidae jsou z 20 sekvencí 2 druhově a 14 rodově referenčních. Až na jednu sekvenci (druhově referenční) pro okna délek od 15 nukleotidů a metodu RM byla identifikace perfektní.

V souboru je také množství sekvencí nereferenčních čeledí ale referenčních řádů. Stejně jako v předešlých souborech byly všechny metody schopny s vysokou mírou úspěšnosti správně zařadit sekvence k příbuzným čeledím řádů Charadriiformes a Passeriformes. U ostatních řádů byla identifikace víceméně neúspěšná, viz **Tab. 7.9**.

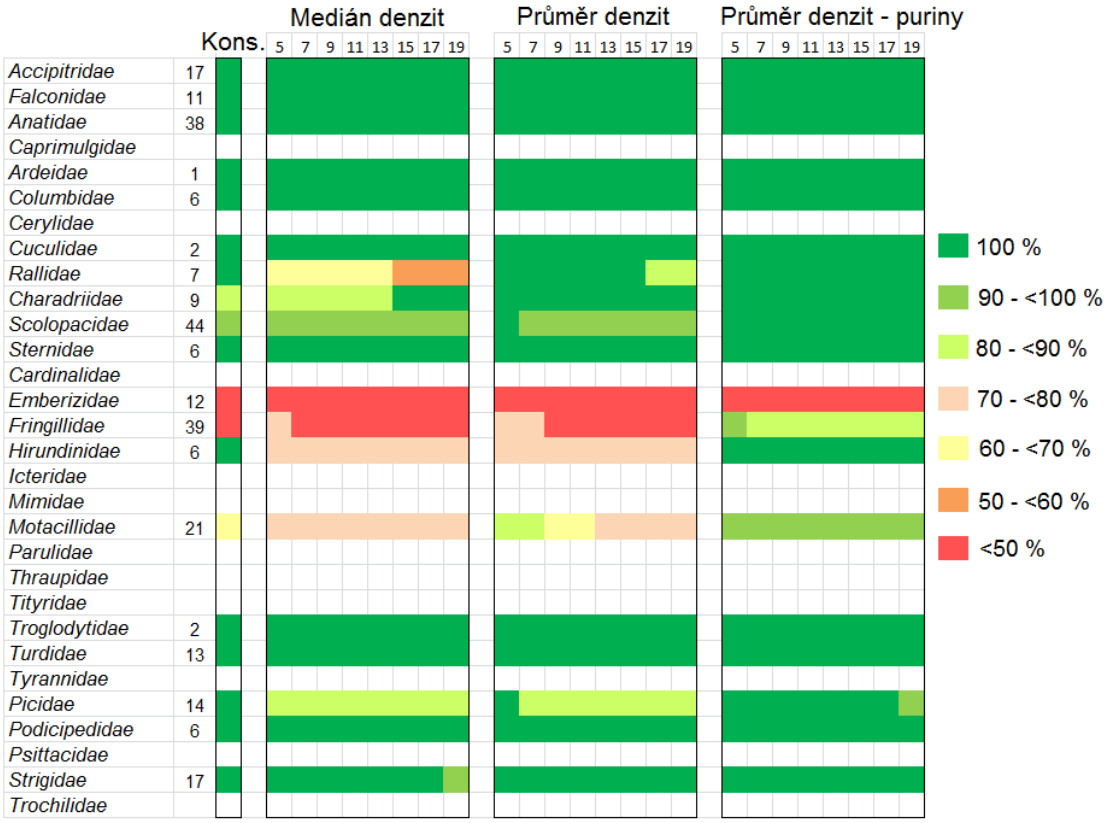
Tab. 7.9 Průměrné úspěšnosti identifikace do řádů pro sekvence nereferenčních čeledí souboru SWEBI.

		Úspěšnost [%]			
Řád	**Počet**	**RK**	**RM**	**RP**	**RP-RY**
Accipitriformes	2	0,00	38,46	53,85	15,38
Coraciiformes	4	0,00	0,00	0,00	0,00
Gruiformes	2	0,00	0,00	0,00	0,00
Charadriiformes	34	70,59	94,34	96,61	100,0
Passeriformes	120	96,67	96,09	97,12	98,72
Pelecaniformes	7	0,00	0,00	0,00	0,00
Strigiformes	2	100,0	7,69	100,0	0,00

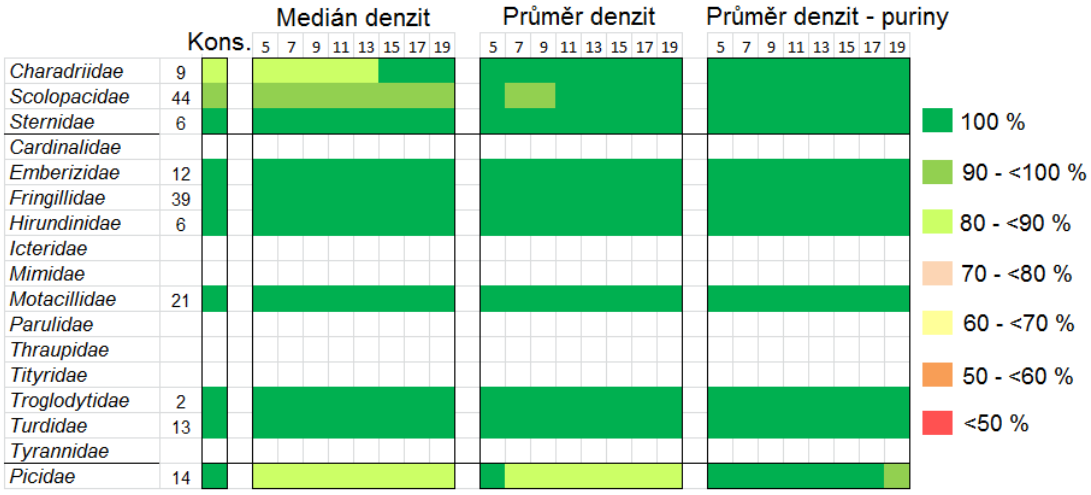
7.2.7 SOUBOR NORBI

Soubor NORBI (Birds of Scandinavia – Norwegian birds) obsahuje celkem 480 sekvencí převážně stejných druhů jako předchozí soubor SWEBI ale jedná se o jiné

populace. Na **Obr. 7.18** je graficky znázorněna úspěšnost identifikace těchto sekvencí do čeledí a na **Obr. 7.19** pak do řádů.



Obr. 7.18 Grafické znázornění procentuální úspěšnosti identifikace do čeledí sekvencí ze souboru NORBI.



Obr. 7.19 Grafické znázornění procentuální úspěšnosti identifikace do řádů sekvencí ze souboru NORBI.

Všechny metody byly schopny perfektně identifikovat pro celý rozsah délek výpočetního okna čeledě Accipitridae, Anatidae, Ardeidae, Columbidae, Cuculidae, Falconidae, Podicipedidae, Sternidae, Troglodytidae a Turdidae. Pro ostatní čeledě, u

jejichž sekvencí se vyskytovalo přiřazování k nesprávné čeledi, platí, že metoda purinové identifikace dávala nejlepší výsledky, až na čeleď Emberizidae, u které selhávaly všechny metody.

Co se týče přiřazování k čeledím jiného řádu, než kam sekvence skutečně patří, tak u řádu Passeriformes se žádná taková chyba nevyskytla a purinová identifikace opět dává obecně nejlepší výsledky.

Úspěšnost přiřazování sekvencí nereferenčních čeledí ale referenčních řádů byla opět jako v předešlých souborech vysoká pro sekvence řádů Charadriiformes a Passeriformes a nízká nebo většinou nulová u ostatních řádů, viz **Tab. 7.10**

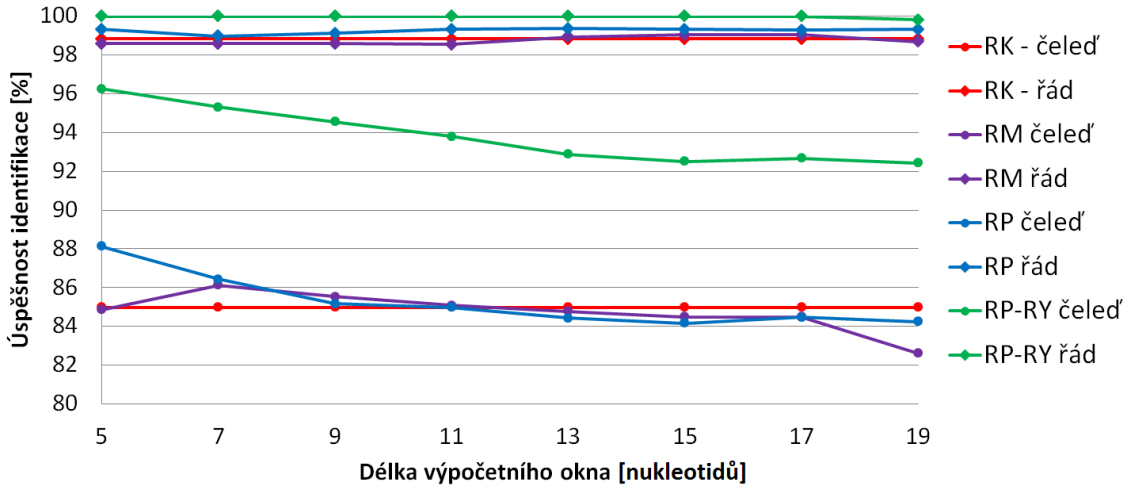
Tab. 7.10 Průměrné úspěšnosti identifikace do řádů pro sekvence nereferenčních čeledí souboru NORBI.

		Úspěšnost [%]			
Řád	**Počet**	**RK**	**RM**	**RP**	**RP-RY**
Accipitriformes	3	66,66	38,89	41,67	20,51
Coraciiformes	1	0,00	0,00	0,00	7,69
Gruiformes	2	0,00	0,00	0,00	0,00
Charadriiformes	35	60,00	96,70	98,90	100,0
Passeriformes	140	97,86	95,99	95,88	99,34
Pelecaniformes	4	0,00	0,00	0,00	5,77

7.3 SHRNUTÍ

Identifikace do čeledí byla v součtu provedena na 3244 sekvencích patřících do 30 různých čeledí. Referencí čeledí bylo 38. Osm z referenčních čeledí nemělo v analyzovaných souborech žádného zástupce. Reference čeledí byly vytvořeny jako konsenzuální sekvence (RK) z vícenásobně zarovnaných sekvencí, medián ND vektorů (RM), průměr ND vektorů (RP) a průměr ND vektorů pouze purinových nukleotidů (RP-RY). Nukleotidové denzitní vektory byly počítány pro rozsah výpočetních oken W=5 až W=19 nukleotidů. **Obr. 7.20** znázorňuje vážené průměry úspěšnosti identifikace do čeledí a řádů v závislosti na velikosti výpočetního okna, kromě metody RK, která se v okně nepočítá. Úspěšnost identifikace do čeledí pro metody RM, RP a RP-RY má mírně klesající závislost na velikosti výpočetního okna, přičemž metody mediánu a průměru ND vektorů mají velmi podobné výsledky. Metoda průměru purinových ND vektorů má přibližně o 8 % vyšší úspěšnost než obě dříve zmíněné metody a to konkrétně pro čeledě 96,24 % pro okno délky W=5 nukleotidů, což je pro čeledě nejlepší výsledek ve váženém průměru pro všechny soubory. Úspěšnost identifikace do řádu se pro všechny tři metody liší jen málo a pro všechny metody jsou úspěšnosti nad

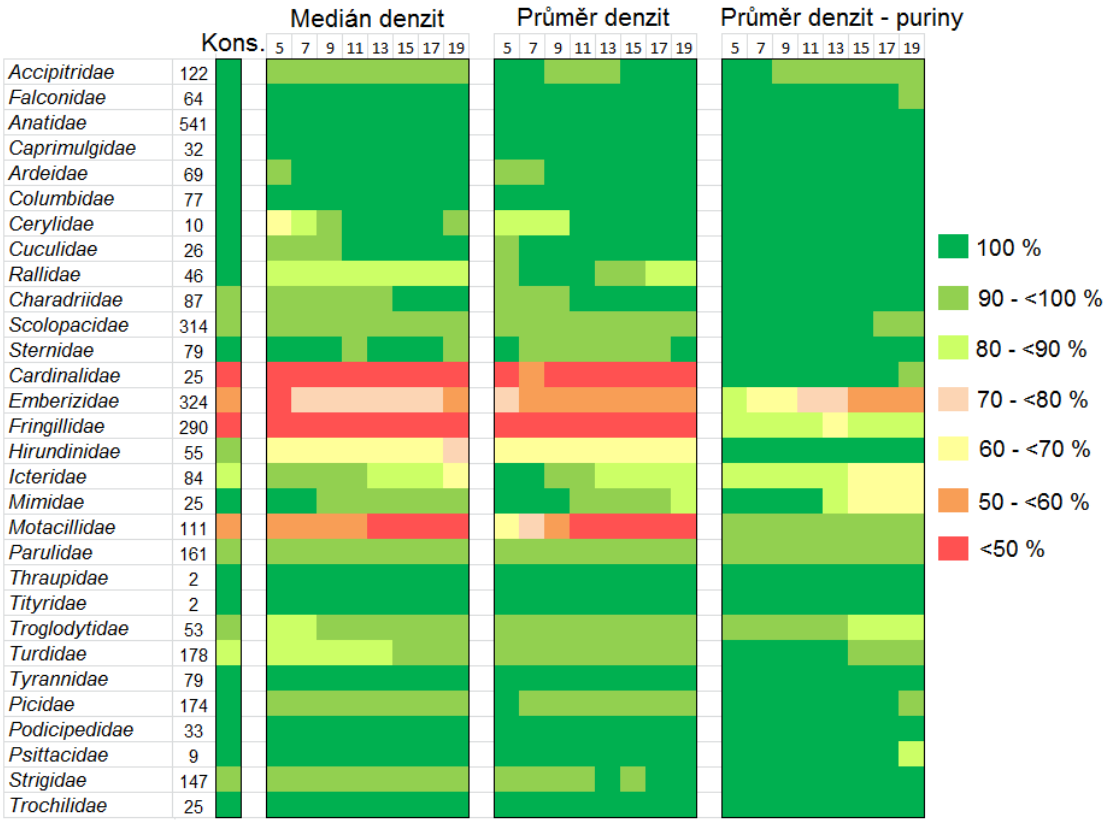
98 %. Metody mediánu a průměru ND vektorů dosáhly maxima pro okna délek *W*=15 a *W*=17 nukleotidů. Nejlepších výsledků pro řády opět dosáhla metoda purinové identifikace RP-RY, která pro okna délek *W*=5 až *W*=17 nukleotidů měla 100% úspěšnost pro všechny soubory. Metoda konsenzuální sekvence RK, která sloužila jako „standardizovaná“ metoda dávala výsledky srovnatelné s metodami mediánu a průměru ND vektorů.



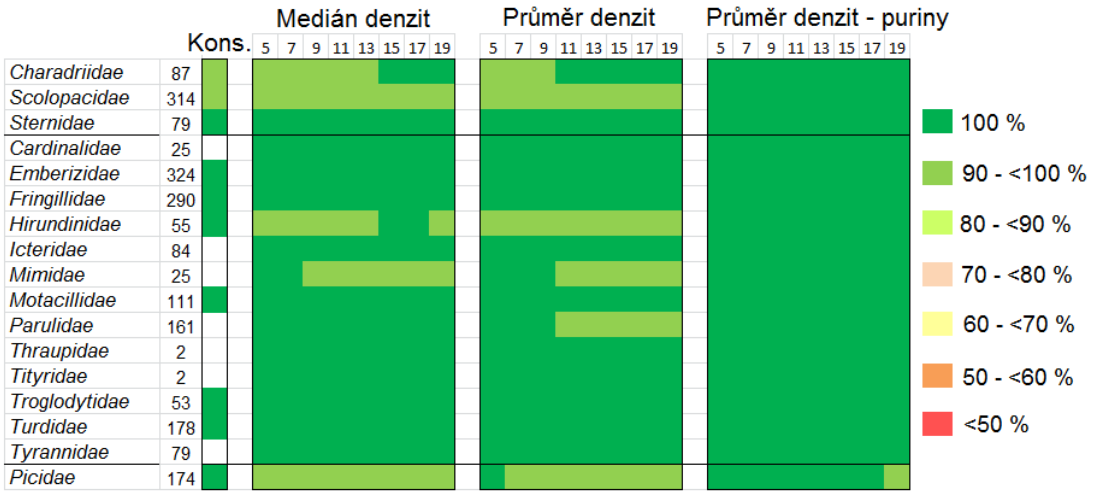
Obr. 7.20 Vážené průměrné úspěšnosti identifikace sekvencí ze všech souborů do čeledí a řádů v závislosti na velikosti výpočetního okna. RK – metoda konsenzuální sekvence, RM – metoda mediánové denzity, RP – metoda průměrné denzity, RP-RY – purinová identifikace.

Nejmenší úspěšnosti při identifikaci do čeledí měly sekvence náležící do čeledí Cardinalidae, Emberizidae a Fringillidae z řádu Passeriformes, které byly nesprávně přiřazovány především k čeledím Icteridae, Parulidae nebo Thraupidae ze stejného řádu. Všechny tyto zmíněné čeledě patří do jedné podskupiny zpěvných ptáků. Čeleď Motacillidae, která má také nízkou úspěšnost identifikace, patří do stejné nadčeledi jako výše zmíněná skupina, a její sekvence byly přiřazovány právě k těmto čeledím. Nejvýraznější zlepšení úspěšnosti přiřazování k čeledím vykázala metoda purinové identifikace právě u těchto problematických blízce příbuzných čeledí.

Na **Obr. 7.21** je graficky znázorněna procentuální úspěšnost identifikace do čeledí pro všechny soubory dohromady (mimo referenční soubor BARG) a na **Obr. 7.22** je znázorněna procentuální úspěšnost identifikace všech sekvencí do řádů. Z těchto obrázků je patrná výrazně větší úspěšnost metody identifikace založené na porovnávání sekvence s referencemi spočítanými jako průměr nukleotidových denzitních vektorů purinových nukleotidů.



Obr. 7.21 Grafické znázornění procentuální úspěšnosti identifikace do čeledí sekvencí ze všech analyzovaných souborů.



Obr. 7.22 Grafické znázornění procentuální úspěšnosti identifikace do řádů sekvencí ze všech souborů.

7.4 DISKUZE

Identifikační analýza ukázala, že referenční nukleotidové denzitní vektory vytvořené jako průměr jen purinových nukleotidů nejlépe ze všech testovaných přístupů charakterizují jednotlivé čeledě i řády. Metoda je samozřejmě úzce provázaná s vlastnosti použitých DNA barcodingových sekvencí. Vhodnost těchto vcelku krátkých

sekvencí k třídění sekvencí podle jejich podobnosti do vyšších taxonomických skupin je diskutabilní. Nepředpokládá se, že by část rychle mutujícího mitochondriální genu mohla nést dostatek informace o příslušnosti druhu do vyšší taxonomické skupiny, které byly vytvořené především na základě morfologické podobnosti. Zde prezentovaná identifikační analýza ukázala, že vhodně zvolená metoda může mezi sekvencemi pro vyšší taxonomické skupiny než je druh či rod nalézt sdílené podobnosti a podle těchto podobností určit příslušnost dalších sekvencí k taxonomickým skupinám. Na základě výsledků lze tvrdit, že ačkoliv byly reference pro jednotlivé čeledě vytvořeny z omezeného počtu druhů ptáků vyskytujících se v části Jihoamerického kontinentu, byla jejich podobnost se sekvencemi ptáků z jiných kontinentů dostatečná k úspěšnému přiřazení k vyšší taxonomické skupině, a tudíž DNA barcodingové sekvence nesou dostatek informace o příslušnosti druhů k vyšším taxonomickým skupinám, jako jsou čeledě a řády. Samozřejmě tento poznatek nelze zobecnit na všechny skupiny organismů.

V impaktovaném periodiku byla za posledních 5 let publikována pouze jedna práce přímo se zabývající identifikací sekvencí do vyšších taxonomických skupin. Wilson a kol. vyhodnocovali úspěšnost identifikace sekvencí lišajovitých motýlů (čeleď Sphingidae) do vyšších skupin jako je rod, tribus a podčeleď v případě, že sekvence stejného druhu se v databázi nenachází. Testováno bylo 118 sekvencí, které byly porovnávány s 1095 sekvencemi, jež tvořily referenční databázi, z které bylo postupně odstraňováno určité procento sekvencí. Identifikace probíhala na principu vytváření fylogenetických stromů metodou NJ z K2P distancí a analyzovaná sekvence byla přiřazena k té taxonomické skupině, se kterou tvořila ve stromu shluk. Přiřazení bylo ovlivněno různými možnostmi topologie shluků. Nejvyšší úspěšnost tohoto přístupu byla 83 % přiřazení do rodu, 74 % do tribusu a 90 % do podčeledi. [174]

V této práci navržená a testovaná metodika založená na nukleotidových denzitních vektorech dosáhla na jiných a rozsáhlejších datech srovnatelných či lepších výsledků. Nejlepších výsledků dosáhla metoda založená na porovnávání s referencemi čeledí vypočítaných jako průměr nukleotidových denzitních vektorů jen purinových nukleotidů. Nejlepší úspěšnosti této metody bylo dosaženo pro výpočetní okno W=5 nukleotidů v průměru pro všechny sekvence a všechny čeledě. Z 30 referenčních čeledí, které měly v testovaných souborech svoje zástupce, jich metoda RP-RY pro výpočetní okno W=5 nukleotidů 100% správně identifikovala 24, což v sumě představuje 2221sekvencí. U dalších 6 čeledí (1023 sekvencí) neklesla úspěšnost pod 80 %. Celkem bylo z 3244 sekvencí správně identifikováno do čeledi 3122 sekvencí. Špatně identifikovaných sekvencí bylo 122 (převážně z čeledě Emberizidae), které byly přiřazovány alespoň k blízce příbuzným čeledím. Identifikace do řádů byla pro okno W=5 nukleotidů bezchybná.

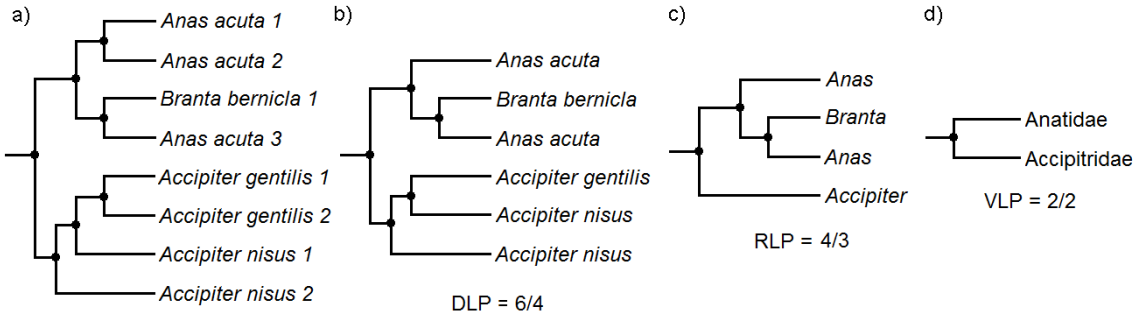
Vynikající úspěšnost této metody oproti ostatním metodám, které porovnávají zastoupení všech nukleotidů v sekvencích, je nejpravděpodobněji způsobená vysokou konzervativností pozic purinových nukleotidů.

8 ANALÝZA DENDROGRAMŮ

V případě databáze BOLD, která je v DNA barcodingu nejobsáhlejší, je identifikace druhů z DNA barcodingových sekvencí založena na vyhledání homologních sekvencí v databázi pomocí algoritmu BLAST. Stejný algoritmus se k účelu vyhledávání nejpodobnějších sekvencí používá i v databázi GenBank, která také obsahuje mimo jiné stejná data jako databáze BOLD. Dalším nástrojem analýzy sekvencí, který databáze BOLD poskytuje (k 31. 7. 2015), je vytvoření dendrogramu ze sekvencí jednotlivých projektů (soubor sekvencí určité lokality nebo skupiny organismů). Dendrogram je konstruován metodou spojování sousedů (neighbor-joining) z *p*-distancí mezi sekvencemi, které mohou být upraveny korekcí Jukesovým-Cantorovým (JC) nebo dvou-parametrickým Kimurovým (K2P) evolučním modelem. Databáze však neumožňuje cílený výběr sekvencí ke konstrukci dendrogramu, lze použít pouze jeden konkrétní projekt sekvencí s výběrem minimální délky sekvencí, které budou použity.

Byla provedena srovnávací analýza dendrogramů vybraných datasetů sekvencí, které byly použity k identifikační analýze do druhů a čeledí (viz kapitola 6 a 7). Konstruovány byly dendrogramy metodou spojování sousedů na základě čtyř metod výpočtu vzdáleností sekvencí. První a druhá metoda odpovídá standardní metodice používané v databázi BOLD, tj. výpočet *p*-distancí s korekcí Jukesovým-Cantorovým a dvou-parametrickým Kimurovým evolučním modelem ze zarovnaných sekvencí. Třetí metoda počítá Euklidovské vzdálenosti mezi nukleotidovými denzitními vektory sekvencí pro všechny čtyři typy nukleotidů (END) a čtvrtá metoda počítá Euklidovské vzdálenosti pro sumu nukleotidových denzitních vektorů purinových nukleotidů (PEND). Pro výpočet nukleotidových denzitních vektorů byly použity výpočetní okna délek 3, 5, 7 a 9 nukleotidů. Pro vykreslení všech čtyř druhů dendrogramů a následné porovnávání byly použity pouze informace o uspořádání uzlů; délka větví porovnávána nebyla.

Datasety sekvencí pro konstrukci dendrogramů obsahovaly všechny sekvence daného DNA barcodingového projektu pro vybranou skupinu organismů. K vyhodnocení stromů byla použita jednoduchá metrika nezávislá na subjektivním hodnocení správnosti/nesprávnosti přiřazení sekvence do shluku. Navržená metodika pouze počítá poměr mezi počtem větví stromu vůči počtu taxonomických jednotek. Vytvořené dendrogramy pro jednotlivé datasety a metodu výpočtu vzdáleností byly kondenzovány pro druh, rod a vyšší taxonomickou skupinu (čeleď, nadčeleď, apod.). Kondenzace dendrogramu znamená, že v případě dvou listů (koncových větví) mající společný uzel a patřící ke stejnému druhu, je jeden z těchto listů (libovolný) a uzel odstraněny. Tento krok se opakuje, až nejsou ve stromě žádné dva listy ve stejném uzlu náležící ke stejnému druhu. Kondenzace pro rod i vyšší taxonomickou skupinu je obdobná. Příklad kondenzace viz **Obr. 8.1**.



Obr. 8.1 Kondenzace dendrogramu: a) originální dendrogram, b) kondenzace pro druh, c) kondenzace pro rod a d) kondenzace pro čeleď.

Kondenzací se odstraní nadbytečné listy pro skupinu sekvencí stejného druhu a ve stromu zůstane místo větve s množstvím listů pouze jeden reprezentativní list. V ideálním případě pak strom po kondenzaci pro druh obsahuje pouze stejný počet listů jako je celkový počet druhů v datasetu. V případě, že některý druh má vysokou vnitrodruhovou variabilitu a v souboru se vyskytují sekvence jiného blízce příbuzného druhu, pak mohou být sekvence těchto druhů uspořádány v jedné větvi a po kondenzaci nemusí být tyto druhy reprezentovány pouze dvěma listy. Obdobně to platí také pro rody a vyšší taxonomickou skupinu. Vliv má samozřejmě i metoda výpočtu vzdáleností sekvencí. Kondenzované dendrogramy tedy reflektují jednak kvalitu sekvencí a také použitou metodu výpočtu vzdáleností sekvencí.

Úspěšnost kondenzace dendrogramu lze jednoduše vyhodnotit poměrem mezi počtem listů a počtem taxonomických jednotek v souboru:

- poměr počtu listů kondenzovaného dendrogramu pro druh k počtu druhů v datasetu (poměr DLP),
- poměr počtu listů kondenzovaného dendrogramu pro rod k počtu rodů v datasetu (poměr RLP),
- poměr počtu listů kondenzovaného dendrogramu pro vyšší taxonomickou skupinu (např. čeleď) k počtu těchto skupin v datasetu (poměr VLP).

V ideálním případě by všechny zmíněné poměry měly mít hodnotu 1. Hodnota poměru v podstatě vyjadřuje do kolika větví je každá z taxonomických jednotek průměrně rozdělena. Pokud nebudeme uvažovat vliv metody konstrukce dendrogramu, pak má na hodnotu výše definovaných poměrů především vliv vnitrodruhová a mezidruhová variabilita a schopnost metody kvantifikace podobností sekvencí tyto variability charakterizovat. A dále u poměrů RLP a VLP hraje roli, jakou měrou reflektují DNA barcodingové sekvence rozdělení druhů do vyšších taxonomických skupin.

Jelikož některé soubory sekvencí, pro které byly dendrogramy konstruovány, obsahovaly i sekvence, jejichž příslušnost k druhu byla identifikační analýzou zpochybněna, či mají sekvence některých druhů značně vysokou anebo překrývající se vnitrodruhovou variabilitu, pak hodnoty poměrů DLP musí být vyšší než 1.

Dendrogramy byly zkonstruovány pro projekty CUCAD, GBAP, MNCN, BCBNC a BARG. V **Tab. 8.1** jsou shrnuty kondenzační poměry DLP, RLP a VLP pro jednotlivé soubory a metody výpočtu distancí mezi sekvencemi.

Tab. 8.1 Poměry počtu větví kondenzovaných dendrogramů pro vybrané DNA barcodingové projekty.

				Parametr				Parametr		
Projekt		**Počet**	**Metoda**	**DLP**	**RLP**	**VLP**	**Metoda**	**DLP**	**RLP**	**VLP**
CUCAD	Sekvencí	716	**JC**	1,26	1,35	2,83	**K2P**	1,26	1,35	2,50
	Druhů	54	**END 3**	1,11	1,12	1,83	**PEND 3**	1,09	1,08	1,67
	Rodů	26	**END 5**	1,11	1,12	1,50	**PEND 5**	1,09	1,08	1,67
	Nadčeledí	6	**END 7**	1,11	1,12	1,83	**PEND 7**	1,09	1,08	1,67
			END 9	1,11	1,19	2,33	**PEND 9**	1,09	1,08	1,67
GBAP	Sekvencí	1515	**JC**	1,31	2,36	76,33	**K2P**	1,28	2,30	76,33
	Druhů	475	**END 3**	1,32	1,96	25,00	**PEND 3**	1,33	1,83	19,67
	Rodů	166	**END 5**	1,24	1,89	24,33	**PEND 5**	1,32	1,89	21,33
	Řádů	3	**END 7**	1,24	1,89	25,33	**PEND 7**	1,29	1,87	23,00
			END 9	1,23	1,91	27,67	**PEND 9**	1,29	1,84	25,00
MNCN	Sekvencí	758	**JC**	3,05	3,09	7,67	**K2P**	1,05	1,03	1,50
	Druhů	42	**END 3**	1,07	1,06	1,50	**PEND 3**	1,23	1,09	1,50
	Rodů	35	**END 5**	1,07	1,06	2,33	**PEND 5**	1,19	1,06	1,83
	Řádů	6	**END 7**	1,07	1,06	2,33	**PEND 7**	1,19	1,06	1,50
			END 9	1,05	1,06	2,33	**PEND 9**	1,24	1,06	1,50
BCBNC	Sekvencí	840	**JC**	1,30	1,38	3,08	**K2P**	1,30	1,42	3,23
	Druhů	96	**END 3**	1,00	1,06	2,08	**PEND 3**	1,91	1,04	1,54
	Rodů	50	**END 5**	1,00	1,06	2,31	**PEND 5**	1,64	1,06	1,62
	Podčeledí	13	**END 7**	1,00	1,08	2,38	**PEND 7**	1,68	1,06	1,77
			END 9	1,00	1,20	2,92	**PEND 9**	1,55	1,06	1,77
BARG	Sekvencí	1588	**JC**	1,05	1,11	2,27	**K2P**	1,04	1,10	2,42
	Druhů	498	**END 3**	1,01	1,10	1,81	**PEND 3**	1,11	1,09	1,50
	Rodů	313	**END 5**	1,02	1,10	1,85	**PEND 5**	1,11	1,12	1,65
	Řádů	26	**END 7**	1,01	1,09	2,15	**PEND 7**	1,12	1,10	1,73
			END 9	1,01	1,10	2,08	**PEND 9**	1,11	1,11	1,88

DLP – poměr kondenzace pro druh
RLP – poměr kondenzace pro rod
VLP – poměr kondenzace pro vyšší taxonomickou jednotku
END *x*, PEND *x* – kde *x* je velikost výpočetního okna *W*

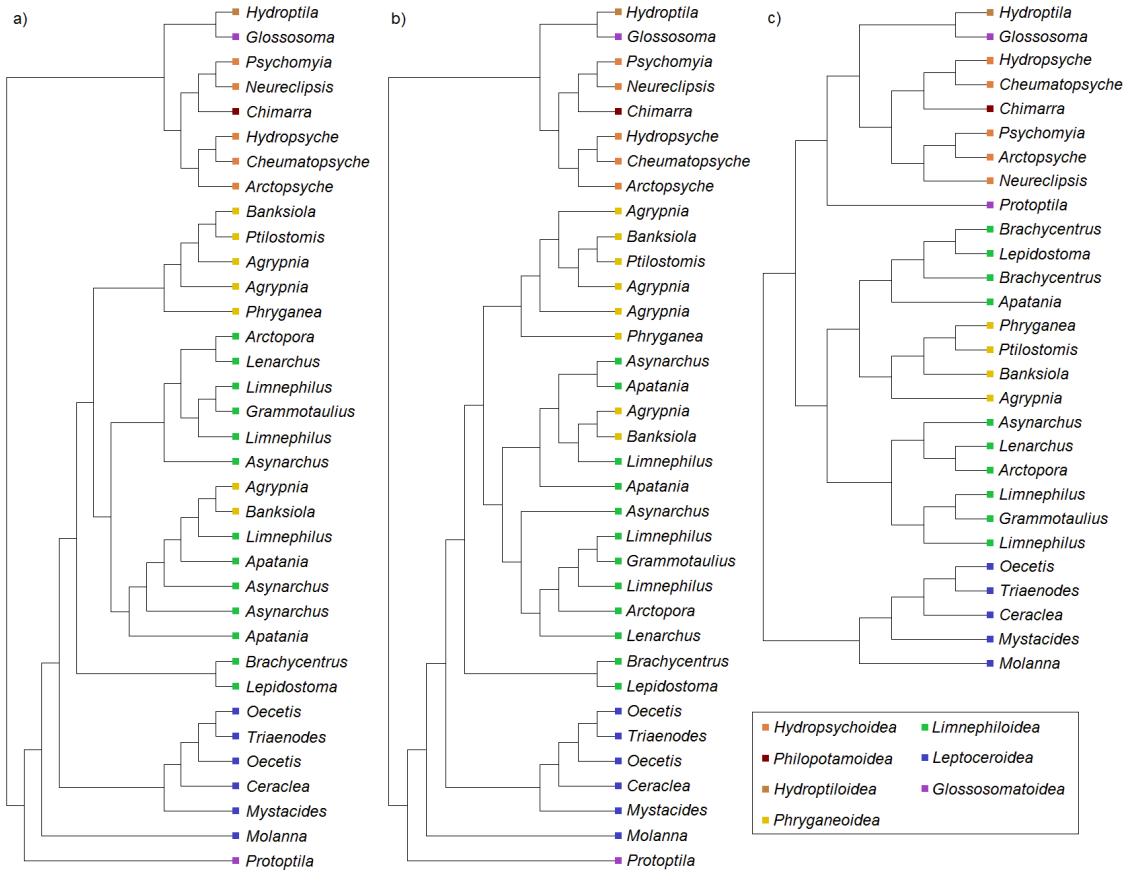
Kondenzace dendrogramů zkonstruovaných z JC evolučních distancí byla nejméně úspěšná ve všech případech, případně shodná s výsledky pro K2P distance. Poměry kondenzace pro druh se často blíží ideální hodnotě 1; nejlepších výsledků bylo dosaženo pro metodu výpočtu vzdáleností END. Vliv velikosti výpočetního okna je velmi malý a není možné jednoznačně určit, která z použitých velikostí je nejlepší. Větší počet listů pro sekvence druhů (poměr DLP) u dendrogramů konstruovaných z hodnot PEND oproti END je způsoben konzervativností obsahu a pozic purinových/pyrimidinových nukleotidů u blízce příbuzných druhů, jež se oddělily od společného předka v nedávné době. V takovém případě byly sekvence těchto druhů v jedné společné větvi a tuto větev nebylo možné redukovat jen na jednoho zástupce pro každý druh. Metoda výpočtu vzdáleností END byla tyto druhy schopna odlišit díky zohledňování obsahu a pozic všech druhů nukleotidů.

Rozdělení do vyšších taxonomických jednotek je limitované informačním obsahem DNA barcodingových sekvencí. Jelikož se jedná o vcelku krátké, rychle mutující sekvence, nemusí být dostatečně odlišné mezi jednotlivými druhy a zároveň dostatečně shodné v rámci vyšší taxonomické skupiny jako jsou čeledě či řády. Identifikační analýza do čeledí pomocí porovnávání referenčních nukleotidových denzitních vektorů pro čeledě s nukleotidovými denzitami analyzovaných sekvencí ukázala, že DNA barcodingové sekvence by mělo být možné k tomu účelu použít s vysokou mírou úspěšnosti (viz kapitola 7).

V dendrogramech zkonstruovaných ze standardně používaných evolučních distancí JC a K2P byly vyšší taxonomické skupiny rozděleny do více samostatných větví než v případě dendrogramů zkonstruovaných ze vzdáleností END a PEND. Je to především případ souboru GBAP se sekvencemi obojživelníků, u nichž se ukázalo jako problémové nejen shlukování podle druhů, ale i rodové rozdělení a především rozdělení do tří řádů bylo značně nekonzistentní. Stromy zkonstruované ze vzdáleností JC a K2P mají několikanásobně vyšší hodnoty kondenzace do řádu než stromy zkonstruované z END a PEND vzdáleností, avšak i pro tyto stromy jsou kondenzační parametry podstatně vyšší než u souborů pro jiné skupiny organismů. Potvrzuje to stanovisko, že část genu *coxI* není pro obojživelníky vhodnou DNA barcodingovou sekvencí [88].

Ve všech případech byly dendrogramy vytvořené na základě standardně používané *p*-distance s korekcí Jukesovým-Cantorovým evolučním modelem „horší“, co se týče počtu větví pro jednotlivé druhy, rody a čeledě (případně rody), než dendrogramy z Euklidovských vzdáleností mezi nukleotidovými denzitními vektory všech nukleotidů nebo jen sumy purinových nukleotidů. Lze říci, že Euklidovská vzdálenost mezi nukleotidovými denzitními vektory dvou sekvencí je minimálně informačně ekvivalentní standardně používaným evolučním distancím JC a K2P, i když jde o deterministickou metriku bez vazby na evoluční modely.

Příklad dendrogramů kondenzovaných pro rody souboru CUCAD viz **Obr. 8.2**, kde je dobře patrný rozdíl mezi počtem listů pro rody u jednotlivých metod kvantifikace podobností mezi sekvencemi.



Obr. 8.2 Kondenzované dendrogramy pro rody projektu CUCAD sestrojené z: a) JC evolučních distancí, b) K2P evolučních distancí a c) END 5 distancí. Barevné značky odkazují na příslušnost rodů k nadčeledím.

Analýza zkonstruovaných dendrogramů ukázala, že Euklidovské vzdálenosti mezi nukleotidovými denzitními vektory jednotlivých sekvencí jsou vhodnou kvantifikací podobnosti/nepodobnosti sekvencí pro účely konstrukce dendrogramů z distančních dat.

9 ZÁVĚR

Cílem dizertační práce bylo navrhnout numerickou reprezentaci DNA sekvencí, která bude vhodná pro komparativní genomiku se zaměřením na identifikaci druhů, dále pro tento účel navrhnout vhodnou metodiku a provést verifikaci na souboru reálných dat. Nukleotidové denzitní vektory byly navrženy jako taková vhodná numerická reprezentace, která vyjadřuje průměrné zastoupení jednotlivých druhů nukleotidů v definované oblasti DNA sekvence. Pro každý typ nukleotidu je v posuvném výpočetním okně vypočítán samostatný číselný vektor průměrného zastoupení tohoto nukleotidu. Výsledkem numerického mapování DNA sekvence jsou pak 4 samostatné vektory, které je možné pro další účely např. sumovat dle biochemických vlastností nukleotidů. Následně byla navržena metodika porovnávání nukleotidových denzitních vektorů pro identifikační účely založená na výpočtu Euklidovské vzdálenosti mezi denzitními vektory. Navržená numerická reprezentace a metodika porovnávání byly otestovány na rozsáhlém souboru DNA barcodingových sekvencích získaných z volně dostupné databáze BOLD. Do testovacích dat byly zahrnuty všechny třídy čelistnatých obratlovců a z bezobratlých pak zástupci chrostíků a komárů. Kromě testování identifikace do druhů byla provedena i podobná analýza identifikace do čeledí a řádů (vyšší taxonomické skupiny), k čemuž byly použity DNA barcodingové sekvence ptáků z různých kontinentů, a na závěr byla provedena ještě komparativní analýza dendrogramů konstruovaných ze standardně používaných evolučních vzdáleností a Euklidovských vzdáleností mezi nukleotidovými denzitními vektory.

Pro identifikaci sekvencí do druhů byla část sekvencí použita pro vytvoření druhových referencí a druhá část dat byla nezávisle přiřazována k vytvořeným druhovým referencím. Druhové reference byly získány tak, že se každá z použitých sekvencí převedla do formátu nukleotidových denzitních vektorů, poté se takto reprezentované sekvence patřící do jednoho druhu signálově zarovnaly, tj. našla se jejich vzájemná pozice s nejmenší Euklidovskou vzdáleností mezi nukleotidovými denzitními vektory, a jako reference pak sloužil průměr ze zarovnaných nukleotidových denzitních vektorů. Druhové reference byly počítány pro různé velikosti výpočetního okna nukleotidových denzitních vektorů v intervalu W=3 až W=25 nukleotidů. Následná identifikační analýza probíhala tak, že „neznámá“ sekvence byla převedena do nukleotidových denzitních vektorů, pak byla signálově zarovnána se všemi referencemi a přiřazena byla k tomu druhu, s jehož referencí měla nejmenší Euklidovskou vzdálenost. Celkem bylo vytvořeno 751 druhových referencí z 6035 sekvencí. Tyto sekvence byly také verifikačně identifikovány s nejlepším váženým průměrem úspěšnosti identifikace 99,17 % pro délku výpočetního okna W=9 nukleotidů. Dalších 7798 sekvencí bylo použitou pouze k identifikaci s nejlepším váženým průměrem úspěšnosti identifikace 99,06 % také pro délku výpočetního okna W=9 nukleotidů.

Soubory sekvencí použité k identifikační analýze obsahovaly kromě druhově referenčních sekvencí také sekvence nepatřící k referenčním druhům, ale patřící do

stejného rodu. Pro tyto rodově referenční sekvence byla také vyhodnocena úspěšnost přiřazení k druhu stejného rodu. Nejlepší vážený průměr úspěšnosti identifikace byl 91,04 % pro délku výpočetního okna *W*=5 nukleotidů; pro okno *W*=9 nukleotidů byla úspěšnost 89,87 %.

V případě, že byla některá sekvence přiřazována k jinému a stále tomu stejnému druhu pro všechny nebo většinu délek výpočetních oken, byly k této sekvenci vyhledány nejpodobnější sekvence algoritmem BLAST v databázi GenBank. Pokud se výsledek BLASTu shodoval s přiřazením identifikační analýzy, pak bylo toto přiřazení bráno jako správné. Tyto případy signalizovaly, že daná sekvence patří k druhu s vysokou vnitrodruhovou variabilitou, která se překrývá s jiným druhem, nebo je také možné, že jedinec organismu byl špatně taxonomicky identifikován. Neodlišitelnost těchto druhů byla potvrzena nejen výsledky vyhledávání homologních sekvencí pomocí algoritmu BLAST, ale také konstrukcí dendrogramů z Jukesových-Cantorových a Kimurových 2-parametrických evolučních vzdáleností, kde nebyly sekvence těchto druhů odděleny v samostatných větvích. Tento fakt znedůvěryhodňuje spolehlivost a jednoznačnost DNA barcodingu pomocí části genu *coxI*, zvláště v oblasti identifikace blízce příbuzných druhů.

Pro identifikaci do čeledí bylo vytvořeno 38 referencí čeledí z 1450 sekvencí 445 různých druhů argentinských ptáků. Následně bylo dalších celkem 3244 sekvencí přiřazováno k čeledím, přičemž tyto sekvence pocházely především z ptáků ze Severní Ameriky a Euroasijského kontinentu. Reference čeledí byly vytvořeny třemi přístupy založenými na nukleotidových denzitních vektorech obdobně jako v případě druhových referencí. Všechny sekvence patřící druhům jedné čeledě byly převedeny na nukleotidové denzitní vektory a byly signálově zarovnány. První metoda brala jako referenci mediánovou hodnotu ze zarovnaných ND vektorů (RM), druhá metoda brala jejich průměrnou hodnotu (RP) a třetí metoda brala průměrnou hodnotu jen purinových nukleotidů (RP-RY). Při identifikaci do čeledí byly analyzované sekvence převáděny do nukleotidových denzitních vektorů a byly signálově zarovnány ke všem referencím čeledí. K čeledím pak byly přiřazovány na základě minimální Euklidovské vzdálenosti mezi ND vektory. Pro tvorbu referencí byly použity velikosti výpočetního okna v intervalu *W*=5 až *W*=19 nukleotidů.

Jelikož neexistuje žádná standardní metodika identifikace sekvencí do vyšších taxonomických skupin, byla za účelem porovnání výsledků navržena metodika co nejvíce využívající standardně používané přístupy. Pro čeledě byly vytvořeny konsenzuální reference tak, že všechny sekvence patřící k jedné čeledi byly vícenásobně zarovnány a ze zarovnání byla odvozena konsenzuální sekvence, tj. sekvence složená z nukleotidů o nejvyšší četnosti na pozici. Analyzované sekvence pak byly přiřazovány k té referenční čeledi, se kterou měla nejmenší K2P vzdálenost (metoda RK).

„Standardní“ metoda RK dosáhla průměrné úspěšnosti identifikace do čeledi 84,99 %, metoda RM dosáhla průměrné nejlepší úspěšnosti 86,12 % pro výpočetní okno *W*=7

nukleotidů, metoda RP 88,13% pro výpočetní okno *W*=7 nukleotidů a celkově nejlepších výsledků dosáhla metoda RP-RY s 96,24 % pro okno *W*=5 nukleotidů. Vyhodnocena byla i úspěšnost identifikace do řádu, které měly více referenčních čeledí, kde byly výsledky pro RK 98,83 %, RM 99,02 % pro výpočetní okno *W*= 15 nukleotidů, RP 99,36 % pro výpočetní okno *W*=13 nukleotidů a RP-RY měla 100 % úspěšnost pro všechny délky výpočetního okna kromě *W*=19 nukleotidů. Identifikace do vyšších taxonomických skupin založená na referencích z průměrných ND vektorů a výpočtu Euklidovské vzdálenosti mezi sumami ND vektorů jen purinových nukleotidů dosáhla relativně vysoké úspěšnosti. Uvedené výsledky jsou v pozitivním slova smyslu překvapivé, neboť reference čeledí byly vytvořeny ze sekvencí argentinských druhů ptáků, ale identifikovány byly sekvence ptáků z jiných kontinentů a převážně i jiných druhů.

Závěrečná analýza dendrogramů zkonstruovaných pro vybrané DNA barcodingové soubory sekvencí měla za úkol ukázat, jestli je možné použít Euklidovské vzdálenosti mezi denzitními vektory i ke konstrukci stromů, a nakolik je topologie takových stromů odlišná od topologie stromů zkonstruovaných ze standardně používaných distancí. Dendrogramy byly konstruovány metodou spojování sousedů ze zavedených Jukesových-Cantorových (JC) a Kimurových 2-parametrických (K2P) evolučních vzdáleností a Euklidovských vzdáleností mezi ND vektory všech nukleotidů (END) a Euklidovských vzdáleností mezi sumou ND vektorů purinových nukleotidů (PEND). Nukleotidové denzitní vektory byly počítány pro okna délek *W*=3 až *W*=9 nukleotidů. K hodnocení stromů se použila nekomplikovaná metrika, která porovnávala počet shluků pro druhy, rody a vyšší taxonomickou skupinu. Ve většině případů obsahovaly dendrogramy z END a PEND vzdáleností nejnižší počet shluků blížící se počtu daných taxonomických jednotek, z čehož vyplývá, že porovnávání nukleotidových denzitních vektorů může sloužit i ke konstrukci dendrogramů.

Tato dizertační práce ukázala, že výpočetně nenáročná numerická reprezentace nukleotidových sekvencí a plně deterministická metodika komparace může dávat značně kvalitní výsledky v porovnání se standardně používanými metodami pro molekulární identifikaci organismů.

Literatura

[1] SNUSTAD, D. P. a SIMMOMS, M. J. *Principles of Genetics*, 5th ed. John Wiley & Sons Inc., 2009. 784 s. ISBN 978-0470903599.

[2] WATSON, J. D. a CRICK, F. H. S. Molecular structure of nucleic acids. *Nature*. 1953, č. 171. S. 737–738.

[3] STANSFIELD, W. D. *Shaums's Outline of Theory and Problems of Genetics*, 2nd ed. The McGraw-Hill Companies, Inc., 2002. 500 s. ISBN 0-07-136206-1.

[4] STANSFIELD, W., COLOMÉ, J. a CANO, R. *Schaum's Outline of Theory and Problems of Molecular and Cell Biology*. The McGraw-Hill Companies, Inc., 1996. 384 s. ISBN 0-07-060898-9.

[5] CLANCY, S. Chemical structure of RNA. *Nature Education*. 2008, č. 7(1):60.

[6] HOLLEY, R. W., APGAR, J., EVERETT, G. A. et al. Structure of a ribonucleic acid. *Science*. 1965, č. 147. S. 1462-1465.

[7] CLANCY, S. DNA transcription. *Nature Education*. 2008, č. 1(1):41.

[8] CLANCY, S. RNA splicing: introns, exons and spliceosome. *Nature Education*. 2008, č. 1(1):31.

[9] CRICK, F. On protein synthesis. *Symposia of the Society for Experimental Biology*. 1958, č. 12. S. 138–163.

[10] CLANCY, S. a BROWN, W. Translation: DNA to mRNA to protein. *Nature Education*. 2008, č. 1(1):101.

[11] FLEGR, J. *Evoluční biologie*. 2nd ed. Academia, 2009. 572 s. ISBN 978-80-200-1767-3.

[12] PRAY, L. DNA replication and causes of mutation. *Nature Education*. 2008, č. 1(1):214.

[13] DRAKE, J. W., CHARLESWORTH, B., CHARLESWORTH, D. et al. Rates of spontaneous mutation. *Genetics*. 1998, č. 148. S. 1667–1686.

[14] LOEWE, L. Genetic mutation. *Nature Education*. 2008, č. 1(1):113.

[15] CLANCY, S. DNA damage & repair: mechanisms for maintaining DNA integrity. *Nature Education*. 2008, č. 1(1):103.

[16] LYNCH, M., BLANCHARD, J., HOULE, D. et al. Perspective: Spontaneous deleterious mutation. *Evolution*. 1999, č. 53. S. 645–663.

[17] SCHEFFLER, I.E. *Mitochondria*. 2nd ed. John Wiley & Sons, Inc., 2008. 484 s. ISBN 978-0471194224.

[18] MCBRIDE, H. M., NEUSPIEL, M., WASIAK, S. Mitochondria: More Than Just a Powerhouse. *Current Biology*. 2006, č. 16. S. R551-R560.

[19] GREEN, D. R. a REED, J. C. Mitochondria and Apoptosis. *Science*. 1998, č. 281. S. 1309-1312.

[20] BOORE, J. L. Animal mitochondrial genomes. *Nucleic Acids Research*. 1999, č. 27(8). S. 1767-1780.

[21] CLAYTON, D. A. Replication and transcription of vertebrate mitochondrial DNA. *Annual Review of Cell and Developmental Biology*. 1991, č. 7. S. 453-478.

[22] MINDELL, D. P., SORENSON, M. D. a DIMCHEFF, D. E. Multiple independent origins of mitochondrial gene order in birds. *Proceedings of the National Academy of Sciences of the United States of America*. 1998, č. 95. S. 10693-10697.

[23] MANDAVILLI, B. S., SANTOS, J. H. a VAN HOUTEN, B. Mitochondrial DNA repair and aging. *Mutation Research*. 2002, č. 509. S. 127-151.

[24] GRISHKO, V. I., RACHEK, L. I., SPITZ, D. R. et al. Contribution of Mitochondrial DNA Repair to Cell Resistance from Oxidative Stress. *Journal of Biological Chemistry*. 2005, č. 280(10). S. 8901-8905.

[25] MASON, P. A., MATHESON, E. C., HALL, A. G. et al. Mismatch repair activity in mammalian mitochondria. *Nucleic Acids Research*. 2003, č. 31(3). S. 1052-1058.

[26] LITONIN, D., SOLOGUB, M., SHI, Y. et al. Human Mitochondrial Transcription Revisited: only TFAM and TFB2M are required for transcription of the mitochondrial genes in vitro. *Journal of Biological Chemistry*. 2010, č. 258(24). S. 18129-18133.

[27] SENGUPTA, S., YANG, X. a HIGGS, P. G. The Mechanism of Codon Reassignments in Mitochondrial Genetic Codes. *Journal of Molecular Evolution*. 2007, č. 64. S. 662-688.

[28] MORITZ, C., DOWLIN, T. E. a BROWN, W. M. Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Annual Review of Ecology Evolution and Systematics*. 1987, č. 18. S. 269-292.

[29] BALLARD, J. W. O. a RAND, D. M. The population biology of mitochondrial DNA and its phylogenetic implications. *Annual Review of Ecology Evolution and Systematics*. 2005, č. 36. S. 621-642.

[30] GALTIER, N., NABHOLZ, B., GLÉMIN, S. a HURST, G. D. D. Mitochondrial DNA as a marker of molecular diversity: reappraisal. *Molecular Ecology*. 2009, č. 18. S. 4541-4550.

[31] AVISE, J. C., ARNOLD, J., BALL, R. M. et al. Intraspecific phylogeography – the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology Evolution and Systematics*. 1987, č. 18. S. 489-522.

[32] GISSI, C., IANNELLI, F. a PESOLE, G. Evolution of the mitochondrial genome of metazoa as exemplified by comparison of congeneric species. *Heredity*. 2008, č. 101. S. 301-320.

[33] BALLARD, J. W. O. a WHITLOCK, M. C. The incomplete natural history of mitochondria. *Molecular Ecology*. 2004, č. 13. S. 729-744.

[34] WHITE, D. J., WOLFF, J. N., PIERSON, M. a GEMMELL, N. J. Revealing the hidden complexities of mtDNA inheritance. *Molecular Ecology*. 2008, č. 17. S. 4925-4942.

[35] XU, J. The inheritance of organelle genes and genomes: patterns and mechanisms. *Genome*. 2005, č. 48. S. 951-958.

[36] BIRKY, C. W. Jr. The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. *Annual Review of Genetics*. 2001, č. 35. S. 125-148

[37] EYRE-WALKER, A., SMITH, N. H. a MAYNARD-SMITH, J. How clonal are human mitochondria? *Proceedings of the Royal Society B-Biological Sciences*. 1999, č. 266. S. 477-483.

[38] AWADALLA, P., EYRE-WALKER, A. a MAYNARD-SMITH, J. Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science*. 1999, č. 286. S. 2524-2525.

[39] PIGANEAU, G., GARDNER, M. a EYRE-WALKER, A. A broad survey of recombination in animal mitochondria. *Molecular Biology and Evolution*. 2004, č. 21. S. 2319-2325.

[40] TSAOUSIS, A. D., MARIN, D. P., LADOUKAKIS, E. D. et al. Widespread recombination in published animal mtDNA sequences. *Molecular Biology and Evolution*. 2005, č. 22. S. 925-933.

[41] KIMURA, M. *The neutral theory of molecular evolution*. Cambridge University Press, New York, 1983. 384 s. ISBN 978-0521317931.

[42] GROSSMAN, L. I., WILDMAN, D. E., SCHMIDT, T. R. a GOODMAN, M. Accelerated evolution of the electron transport chain in anthropoid primates. *Trends in Genetics*. 2004, č. 20. S. 578-585.

[43] CASTOE, T. A., DE KONING, A. P. J., KIM, H.-M. et al. From the cover: Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*. 2009, č. 106. S. 8986-8991.

[44] DOWLING, D. K., FRIBERG, U. a LINDELL, J. Evolutionary implications of non-neutral mitochondrial genetic variation. *Trends in Ecology & Evolution*. 2008, č. 23. S. 546-554.

[45] WILSON, A. C., CARLSON, S. S. a WHITE, T. J. Biochemical evolution. *Annual Review of Biochemistry*. 1977, č. 46. S. 573-639.

[46] BROWN, W. M., GEORGE, M. a WILSON, A. C. Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America*. 1979, č. 76. S. 1967-1971.

[47] CROTEAU, D. L. a BOHL, V. A. Repair of Oxidative Damage to Nuclear and Mitochondrial DNA in Mammalian Cells. *Journal of Biological Chemistry*. 1997, č. 272(41). S. 25409-25412.

[48] WEI, Y. H. a LEE, H. C. Oxidative stress, mitochondrial DNA mutation, and impairment of antioxidant enzymes in aging. *Experimental Biology and Medicine*. 2002, č. 227. S. 671-682.

[49] SWEETLOVE, L. Number of species on Earth tagged at 8.7 million. *Nature News*. 23. 7. 2011

[50] ROZSYPAL, S. et al. *Nový přehled biologe*. Scientia, 2003. s. 797. ISBN 978-80-86960-23-4.

[51] FLEGR, Jaroslav. *Evoluční biologie*. 2nd ed. Academia, 2009. 572 s. ISBN 978-80-200-1767-3.

[52] SCHWARTZ, J. H. Molecular Systematics and Evolution. *In: Encyclopedia of Molecular Cell Biology and Molecular Medicine*. Wiley-VCH Verlag, Weinheim, 2005. 696 s. ISBN 978-3-527-30544-5.

[53] CHAO, K.-M. a YHANG, L. *Sequence Comparison: Theory and Methods*. Springer-Verlag London, 2009. 209 s. ISBN 978-1-84800-320-0.

[54] KARLIN, S. a ALTSCHUL, S. F. Methods for assessing the statistical signifikance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the United States of America*. 1990, č. 87. S. 2264-2268.

[55] DAYHOFF, M. O., SCHWARTZ, R. M. a ORCUTT, B. C. A Model of Evolutionary Change in Proteins. *In: Atlas of Protein Sequence and Structure*. 5(3), Natl Biomedical Research, 1978. s. 345-352. ISBN 978-0912466071.

[56] ALTSCHUL, S. F. Amino Acid Substitution Matrices from an Informaion Theoretic Perspective. *Journal of Molecular Biology*. 1991, č. 219. S. 555-565.

[57] HENIKOFF, S. a HENIKOFF, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*. 1992, č. 89. S. 10915-10919.

[58] GOONESEKERE, N. C.W. a LEE, B. Context-specific amino acid substitution matrices and their use in the detection of protein homologs. *Proteins*. 2008, č. 71. S. 910-919.

[59] EDDY, S. R. Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology*. 2004, č. 22(8). S. 1035-1036.

[60] NEEDLEMAN, S. B. a WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*. 1970, č. 48(3). S. 443-453.

[61] HUANG, X. a CHAO, K.-M. A generalized global alignment algorithm. *Bioinformatics*. 2003, č. 19(2). S. 228-233.

[62] SMITH, T. F. a WATERMAN, M. S. Identification of Common Molecular Subsequences. *J. Mol. Biol*. 1981, č. 147. S. 195-197.

[63] BAFNA, V., LAWLER, E. L. a PEVZNER, P. A. Approximation algorithms for multiple sequence alignment. *Theoretical Computer Science*. 1997, č. 182. S. 233-244.

[64] CARRILLO, H. a LIPMAN, D. The Multiple sequence Alignment Problem in Biology. *Siam Journal on Applied Mathematics*. 1988, č. 48(5). S. 1073-1082.

[65] EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. 2004, č. 32(5). S. 1792-1797.

[66] ALTSCHUL, S. F., GISH, W., MILLER, W. et al. Basic Local Alignment Search Tool. *Journal of Molecular Biology*. 1990, č. 215. S. 403-410.

[67] KORF, I., YANDELL, M. a BEDELL, J. *BLAST*. O´Reilly Media, 2003. s. 362. ISBN 978-0-596-00299-2.

[68] JUKES, T. H. a CANTOR, C. R. Evolution of Protein Molecules. *In: Mammalian Protein Metabolism*. New York: Academic Press, 1969. s. 21–132.

[69] KIMURA, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*. 1980, č. 16(2). S. 111–120.

[70] FELSENSTEIN, J. *Inferring Phylogenies*. Sinauer Associates, 2004. 664 s. ISBN 978-0878931774.

[71] MOUNT, D. W. Maximum Parsimony Method for Phylogenetic Prediction. *CSH Protocols*. 2008.

[72] BONET, M., STEEL, M., WARNOW, T. a YOOSEPH, S. Better Methods for Solving Parsimony and Compatibility. *Journal of Computational Biology*. 1998, č. 5(3). S. 391-407.

[73] CHO, A. Constructing Phylogenetic Trees Using Maximum Likelihood. *Scripps Senior Theses*. 2012, paper 46.

[74] BAR-HEN, A. a KISHINO, H. Comparing the Likelihood Functions of Phylogenetic trees. *Annals of the Institute of Statistical Mathematics*. 2000, č. 52(1). S. 43-56.

[75] BRAUER, M., HOLDER, M. T., DRIES, L. A. et al. Genetic Algorithms and Parallel Processing in Maximum-Likelihood Phylogeny Inference. *Molecular Biology and Evolution*. 2002, č. 19(10). S. 1717-1726.

[76] FAITH, D. P. Distance methods and the Approximation of Most-parsimonious Tress. *Systematic Zoology*. 1985, č. 34(3). S. 312-325.

[77] FARRIS, J. S. Distance data revisited. *Cladistics*. 1985, 1(1), 67-86.

[78] GASCUEL, O., BRYANT, D. a DENISA, F. Strengths and Limitations of the Minimum Evolution Principle. *Systematic Biology*. 2001, č. 50(5). S. 621-627.

[79] BACKELJAU, T., DE BRUYN, L., DE WOLF, H. et al. Multiple UPGMA and Neighbor-joining Trees and the Performance of Some Computer Packages. *Molecular Biology and Evolution*. 1996, č. 13(2). S. 309-313.

[80] SAITOU, N. a NEI, M. The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution*. 1987, č. 4(4). S. 406-425.

[81] ATTENSON, K. The performance of neighbor-joining algorithms of phylogeny reconstruction. *Algorithmica*. 1999, 25, 251-278.

[82] BRUNO, W. J., SOCCI, N. D. a HALPERN, A. L. Weighted Neighbor Joining: A Likelihood-Based Approach to Distance-Based Phylogeny Reconstruction. *Molecular Biology and Evolution*. 2000, č. 17(1). S.189-197.

[83] FIŠER PEČNIKAR, Ž. a BUZAN, E. V. 20 years since the introduction of DNA barcoding: from theory to application. *Journal of Applied Genetics*. 2014, č. 55. S. 43-52.

[84] SACCONE, C., DE GIORGI, C., GISSI, C. et al. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene*. 1999, č. 238. S. 195-209.

[85] SACCONE, C., GISSI, C., REYES, A. et al. Mitochondrial DNA in metazoan: degree of freedom in a frozen event. *Gene*. 2002, č. 286. S. 3-12.

[86] HEBERT, P. D. N. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B-Biological Sciences*. 2003, č. 270. S. 313 – 321.

[87] KLICKA, J., JOHNSON, K. P. a LANYON, S. M. New World Nine-Primaried Oscine Relationships: Constructing a Mitochondrial DNA Framework. *The Auk*. 2000, č. 117(2). S. 321-336.

[88] VENCES, M., THOMAS, M., BONETT, R. M. a VIEITES, D. R. Deciphering amphibian diversity through DNA barcoding: chances and challenges. *Philosophical Transactions of the Royal Society B-Biological Sciences*. 2005, č. 360. S. 1859-1868.

[89] KRESS, W. J., WURDACK, K. J., ZIMMER, E. A. et al. Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*. 2005, č. 102(23). S. 8369-8374.

[90] LI, F.-W., KUO, L.-Y., ROTHFELS, C. J. et al. rbcL and matK Earn Two Thumbs Up as the Core DNA Barcode for Ferns. *PLoS ONE*. 2011, č. 6(10). S. e26597.

[91] LOPEZ, I. a ERICKSON, D. L. *DNA Barcodes Methods and Protocols*. Springer, 2012. 470 s. ISBN 978-1-61779-591-6.

[92] HEBERT, P. D. N., STOECKLE, M. Y., ZEMLAK, T. S. a FRANCIS, CH. M. Identification of birds through DNA barcodes. *PLoS Biology*. 2004, č. 2(10). S. 1657-1663.

[93] CYWINSKA, A., HUNTER, F. F. a HEBERT, P. D. N. Identifzing Canadian mosquito species through DNA barcodes. *Medical and Veterinary Entomology*. 2006, č. 20(4). S. 413-424.

[94] WARD, R. D., HANNER, R. a HEBERT, P. D. N. The campaign to DNA barcode all fishes. *Journal of Fish Biology*. 2009, č. 74. S. 329-356.

[95] COLLINS, R. A. a CRUICKSHANK, R. H. The seven deadly sins of DNA barcoding. *Molecular Ecology Resources*. 2013, č. 13. S. 969-975.

[96] LOHSE, K. Can mtDNA Barcodes Be Used to Delimit Species? A Response to Pons et al. (2006). *Systematic Biology*. 2009, č. 58(4). S. 439-442.

[97] HICKERSON, M. J., MEYER, C. P. a MORITZ, C. DNA Barcoding Will Often Fail to Discover New Animal Species over Broad Parameter Space. *Systematic Biology*. 2006, č. 55(5). S. 729-739.

[98] MEIER, R., ZHANG, G. a ALI, F. The Use of Mean Instead of Smallest Interspecific Distances Exaggerates the Size of the "Barcoding Gap" and Leads to Misidentification. *Systematic Biology*. 2008, č. 57(5). S. 809-813.

[99] MEIER, R., SHIYANG, K., VAIDYA, G. a NG, P. K. L. DNA Barcoding and Taxonomy in Diptera: A Tale of High Intraspecific Variability and Low Identification Success. *Systematic Biology*. 2006, č. 55(5). S. 715-728.

[100] DESALLE, R., EGAN, M. G. a SIDDALL, M. The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society B-Biological Sciences.* 2005, č. 360(1462). S. 1905-1916.

[101] BERGMANN, T., HADRYS, H., BREVES, G. a SCHIERWATER, B. Character-based DNA barcoding: a superior tool for species classification. *Berliner Und Munchener Tieraryliche Wochenschrift*. 2009, č. 122. S. 446-450.

[102] TAYLOR, R. H. a HARRIS, W.E. An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources*. 2012, č. 12. S. 377-388.

[103] BERGMANN, T., RACH, J., DAMM, S. et al. The potential of distance-based thresholds and character-based DNA barcoding for defining problematic taxonomi entities by CO1 and ND1. *Molecular Ecology Resources*. 2013, č. 13. S. 1069-1081.

[104] KIPLING, W. W. a RUBINOFF, D. Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics*. 2004, č. 20. S. 47-55.

[105] MCCRACKEN, K. G. a SORENSEN, M. D. Is homoplasy or lineage sorting the source of incongruent mtDNA and nuclear gene trees in the stiff-tailed ducks (Nomonyx-Oxyura)? *Systematic Biology*. 2005, č. 54. S. 35–55.

[106] DESALLE, R. Species Discovery versus Species Identification in DNA Barcoding Efforts: Response to Rubinoff. *Conservation Biology*. 2006, č. 20(5). S. 1545-1547.

[107] DESALLE, R. Phenetic and DNA taxonomy; a comment on Waugh. *BioEssays*. 2007, č. 29. S. 1289-1290.

[108] WILLEM, J. a FERGUSON, H. On the use of genetic divergence for identifying species. *Biological Journal of the Linnean Society*. 2002, č. 75. S. 509-519.

[109] BLAXTER, M., MANN, J., CHAPMAN, T. et al. Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B-Biological Sciences*. 2005, č. 360. S. 1935-1943.

[110] CASIRAGHI, M., LABRA, M., FERRI, E. et al. DNA barcoding: a six-question tour to improve user´s awareness about the method. *Briefings in Bioinformatics*. 2010, č. 2(4). S. 440-453.

[111] HEBERT, P. D. N., RATNASINGHAM, S. a DEWAARD, J. R. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London Series B-Biological Sciences (Suppl.)*. 2003, č. 270. S. S96-S99.

[112] RUBINOFF, D. Utility of Mitochondrial DNA Barcodes in Species Conservation. *Conservation Biology*. 2006, č. 20(4). S. 1026-1033.

[113] SARKAR, I. N., PLANET, P. J. a DESALLE, R. CAOS software for use in character-based DNA barcoding. *Molecular Ecology Resources*. 2008, č. 8. S. 1256-1259.

[114] WEITSCHEK, E., VAN VELZEN, R., FELICI, G. a BERTOLAZZI, P. BLOG 2.0: a software system for character-based species classification with DNA Barcode sequences. What it does, how to use it. *Molecular Ecology Resources*. 2013, č. 13. S. 1043-1046.

[115] WEITSCHEK, E., FISCON, G. a FELICI, G. Supervised DNA Barcodes species classification: analysis, comparisons and results. *BioData Mining*. 2014, č. 7(4).

[116] GOLDSTEIN, P. Z. a DESALLE, R. Integrating DNA barcode data and taxonomic practice: Determination, discovery, and description. *Bioessays*. 2010, č. 33. S. 135-147.

[117] RATNASINGHAM, S. a HEBERT, P. D. N. BOLD: The Barcode of Life Data System. *Molecular Ecology Notes.* 2007, č. 7. S. 355-364.

[118] KOSKI, L. B. a GOLDING, G.B. The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution*. 2001, č. 52. S. 540-542.

[119] RUITER, D. E., BOYLE, E. E. a ZHOU, X. DNA barcoding facilitates associations and diagnoses for Trichoptera larvae of the Churchill (Manitoba, Canada) area. *BMC Ecology*. 2013, č. 13(5).

[120] KO, H.-L., WANG, Y.-T., CHIU, T.-S. et al. Evaluating the Accuracy of Morphological Identification of Larval Fishes by Applying DNA Barcoding. *PLoS ONE*. 2013, č. 8(1). S. e53451.

[121] VENCES, M., NAGY, Z. T., SONET, G. a VERHEYEN, E. DNA Barcoding Amphibians and Reptiles. *Methods in Molecular Biology*. 2012, č. 858. S. 79-107.

[122] SMITH, M. A., FISHER, B. L. a HEBERT, P. D. N. DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical Transactions of the Royal Society B-Biological Sciences*. 2005, č. 360. S. 1825-1834.

[123] HEBERT, P. D. N., PENTON, E. H., BURNS, J. M. et al. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences of the United States of America*. 2004, č. 101(41). S. 14812-14817.

[124] BARCO, A., EVANS, J., SCHEMBRI, P. J. et al. Testing the applicability of DNA barcoding for Mediterranean species of top-shells (Gastropoda, Trochidae, Gibbula s.l.). *Marine Biology Research*. 2013, č. 9(8). S. 785-793.

[125] HSU, T.-H., NING, Y., GWO, J. C. a ZENG, Z. N. DNA barcoding reveals cryptic diversity in the peanut worm *Sipunculus nudus*. *Molecular Ecology Resources*. 2013, č. 13(4). S. 596-606.

[126] MARALIT, B. A., AGUILA, R. D., VENTOLERO, M. F. H. et al. Detection of mislabeled commercial fishery by-products in the Philippines using DNA barcodes and its implications to food traceability and safety. *Food Control*. 2013, č. 33(1). S. 119-125.

[127] GALIMBERTI, A., DE MATTIA, F., LOSA, A. et al. DNA barcoding as a new tool for food traceability. *Food Research International*. 2013, č. 50. S. 55-63.

[128] WALLACE, L. J., BOILARD, S. M. A. L., EAGLE, S. H. C. et al. DNA barcodes for everyday life: Routine authentication of Natural Health Products. *Food Research International*. 2012, č. 49. S. 446-452.

[129] CRISTEA, P. D. Conversion of nucleotides sequences into genomic signals. *Journal of Cellular and Molecular Medicine*. 2002, č. 6(2). S. 279-303.

[130] CRISTEA, P. D. Representation and analysis of DNA sequences. In: *Genomic Signal Processing and Statistics*. Hindawi Publishing Corporation, 2005, s. 15-64. ISBN 977-5945-07-0.

[131] VOSS, R. Evolution of long-range fractal correlation and 1/f noise in DNA base sequences. *Physical Review Letters*. 1992, č. 68. S. 3805-3808.

[132] ANASTASSIOU, D. Genomic Signal Processing. *IEEE Signal Processing Magazine*. 2001. S. 8-20.

[133] AFREIXO, V., FERREIRA, P. a SANTOS, D. Fourier analysis of symbolic data: A brief review. *Digital Signal Processing*. 2004, č. 14. S. 523-530.

[134] YAU, S., WANG, J., NIKNEJAD, A. et al. DNA sequence representation without degeneracy. *Nucleic Acids Research*. 2003, č. 31(12). S. 3078-3080.

[135] ZHANG, Y., LIAO, B., a DING, K. On 2D graphical representation of DNA sequence of nondegeneracy. *Chemical Physics Letters*. 2005, č. 411. S. 28-32.

[136] RANDIĆ, M. a BALABAN, A. On A Four-Dimensional Representation of DNA Primary Sequences. *Journal of Chemical Information and Computer Sciences*. 2003, č. 43. S. 532-539.

[137] CHI, R. a DING, K. Novel 4D numerical representation of DNA sequences. *Chemical Physics Letters*. 2005, č. 407. S. 63-67.

[138] RANDIĆ, M., VRAČKO, M., LERŠ, N. a PLAVŠIĆ, D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. *Chemical Physics Letters*. 2003, č. 368. S. 1-6.

[139] HAMORI, E. a RUSKIN, J. H Curves, A Novel Method of Representation of Nucleotide Series Especially Suited for Long DNA Sequences. *Journal of Biological Chemistry*. 1983, č. 258(2). S. 1318-1327.

[140] GUO, F., OU, H. a ZHANG, CH. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Research*. 2003, č. 31(6). S. 1780-1789.

[141] NANDY, A. A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. *Current Science*. 1994, č. 66(4). S. 309-314.

[142] LIAO, B. a WANG, T. New 2D Graphical Representation of DNA Sequences. *Journal of Computational Chemistry*. 2004, č. 25(11). S. 1364-1368.

[143] LIAO, B. a WANG, T. 3-D graphical representation of DNA sequences and their numerical characterization. *Journal Of Molecular Structure-Theochem*. 2004, č. 681. S. 209-212.

[144] GUO, X., RANDIĆ, M. a BASAK, S. A novel 2-D graphical representation of DNA sequences of low degeneracy. *Chemical Physics Letters*. 2001, č. 350. S. 106-112.

[145] COSIC, I. Macromolecular Bioactivity: Is It Resonant Interaction Between Macromolecules? – Theory and Applications. *IEEE Transactions on Biomedical Engineering*. 1994, č. 41(12). S. 1101-1114.

[146] ESKESEN, S. T., ESKESEN, F. N., KINGHORN, B. a RUVINSKY, A. Periodicity of DNA in exons. *BMC Molecular Biology*. 2004, č. 5(12).

[147] SILVERMAN, D. B. a LINSKER, R. A measure of DNA periodicity. *Journal of Theoretical Biology*. 1986, č. 118(3). S. 295-300.

[148] GUTIÉRREZ, G., OLIVER, J. L. a MARIN, A. On the origin of the periodicity of three in protein coding DNA sequences. *Journal of Theoretical Biology*. 1994, č. 167(4). S. 413-414.

[149] TRIFONOV, E. N. Elucidating sequence codes: three codes for evolution. *Annals of the New York Academy of Sciences*. 1999, č. 870. S. 330-338.

[150] DO, J. H. a CHOI, D.-K. Computational Approaches to Gene Prediction. *The Journal of Microbiology*. 2006, č. 44(2). S. 137-144.

[151] TIWARI, S., RAMACHANDRAN, S., BHATTACHARYA, A. et al. Prediction of probable genes by Fourier analysis of genomic sequences. *Computer Applications in the Biosciences*. 1997, č. 13(3). S. 263-270.

[152] AFREIXO, V., FERREIRA, P. a SANTOS, D. Fourier analysis of symbolic data: A brief review. *Digital Signal Processing*. 2004, č. 14. S. 523-530.

[153] MABROUK, M. S. A Study of the Potential of EIIP Mapping Method in Exon Prediction Using the Frequency Domain Techniques. *American Journal of Biomedical Engineering*. 2012, č. 2(2). S. 17-22.

[154] FEUK, L., CARSON, A. R. a SCHERER, S. W. Structural variation in the human genome. *Nature Reviews Genetics*. 2006, č. 7. S. 85–97.

[155] CLANCY, S. Copy number variation. *Nature Education*. 2008, č. 1(1):95.

[156] FREEMAN, J. L., PERRY, G. H., FEUK, L. et al. Copy number variation: New insights in genome diversity. *Genome Research*. 2006, č. 16. S. 949-961.

[157] FONDON, J. W. III a GARNER, H. R. Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences of the United States of America*. 2004, č. 101(52). S. 18058-18063.

[158] MYERS, P. Tandem repeats and morphological variation. *Nature Education*. 2007, č. 1(1):1.

[159] KANNAN, S. K. a MYERS, E. W. An Algorithm for Locating Nonoverlapping Regions of Maximum Alignment Score. *Siam Journal on Computing*. 1993, č. 25(3). S. 648–662.

[160] SOKOL, D., BENSON, G. a TOJEIRA, J. Tandem repeats over the edit distance. *Bioinformatics*. 2007, č. 23(2). S. e30-e35.

[161] BENSON, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*. 1999, č. 27(2). S. 573-580.

[162] ANASTASSIOU, D. Frequency-domain analysis of biomolecular sequences. *Bioinformatics*. 2000, č. 16(12). S. 1073-1081.

[163] DIMITROVA, N., CHEUNG, Y. H. a ZHANG, M. Analysis and Visualization of DNA Spectrograms: Open Possibilities for the Genome Research. *In: Proceedings of the 14th annual ACM international conference on Multimedia*. New York, 2006, s. 1017-1024. ISBN 1-59593-447-2.

[164] SUSSILLO, D., KUNDAJE, A. a ANASTASSIOU, D. Spectrogram Analysis of Genomes. *Eurasip Journal on Applied Signal Processing*. 2004, č. 1. S. 29-42.

[165] ŠKUTKOVÁ, H., VÍTEK, M., BABULA, P. et al. Classification of genomic signals using dynamic time warping. *BMC Bioinformatics.* 2013, č. 14. (Suppl 10):S1.

[166] ZHOU, X., JACOBUS, L. M., DEWALT, R. E. et al. Ephemeroptera, Plecoptera, and Trichoptera fauna of Churchill (Manitoba, Canada): insights into biodiversity patterns from DNA barcoding. *Journal of the North American Benthological Society*. 2010, č. 29(3). S. 814-837.

[167] COSTA, F. O., LANDI, M., MARTINS, R. et al. A Ranking System for Reference Libraries of DNA Barcodes: Application to Marine Fish Species from Portugal. *PLoS ONE*. 2012, č. 7(4). S. e35858.

[168] KOTAKI, M., KURABAYASHI, A., MATSUI, M. et al. Molecular Phylogeny of the Diversified Frogs of Genus Fejervarya (Anura: Dicroglossidae). *Zoological Science*. 2010, č. 27. S. 386-395.

[169] ZHANG, A. B., SIKES, D. S., MUSTER, C. a LI, S. Q. Inffering Species Membership Using DNA Sequences with Back-Propagation Neural Networks. *Systematic Biology*. 2008, č. 57(2). S. 202-2015.

[170] PERRINS, Ch. *Firefly Encyklopedia of Birds*. Firefly Books, 2003. 640 s. ISBN-13: 978-1552977774.

[171] *IOC World Bird List* [online]. [Cit. 30. 7. 2015]. Dostupné z: http://www.worldbirdnames.org/

[172] SUH, A., PAUS, M., KIEFMANN, M. et al. Mosozoic retroposons reveal parrots as the closest living relatives of passerine birds. *Nature Communications*. 2011, č. 2:443.

[173] HACKETT, S. J., KIMBALL, R. T., REDDY, S. et al. A phylogenomic study of birds reveal their evolutionary history. *Science*. 2008, č. 320(5884). S. 1763-1768.

[174] WILSON, J. J., ROUGERIE, R., SCHONFELD, J. et al. When species matches are unavailable are DNA barcodes correctly assigned to higher taxa? An assessment using sphingid moths. *BMC Ecology*. 2011, č. 11(18).

Seznam symbolů a zkratek

AK – aminokyselina
ATP – adenosintrifosfát
BLAST – Basic Local Alignment Search Tool
BLOG – Barcoding with LOGic
BLOSUM – BLOcks amino-acid SUbstitution Matrix
BOLD – The Barcode of Life Data Systems
bp – páry bází
CAOS – Characteristic Attributes Organization System
CNVs – Copy Number Variations
coxI – cytochrom *c* oxidáza podjednotka I
DFT – diskrétní Fourierova transformace
DNA – deoxyribonukleová kyselina
EIIP – Electron-Ion Interaction Potential
FASTA – typ datového souboru genomických nebo proteomických sekvencí
Gbp – giga páry bází
GK – genetický kód
HMM – skryté Markovovy modely
HSP – těžké vlákno mitochondriální DNA, kóduje více mRNA
IUPAC – International Union of Pure and Applied Chemistry
JC – Jukesův-Cantorův evoluční model
K2P – Kimurův dvouparametrický evoluční model
LSP – lehké vlákno mitochondriální DNA, kóduje méně mRNA
MOTU – Molecular Operational Taxonomic Unit
mRNA – mediátorová ribonukleová kyselina
mtDNA – mitochondriální deoxyribonukleová kyselina
NCBI – National Center for Biotechnology Information
ND – nukleotidové denzitní vektory
ORF – otevřený čtecí rámec
PAM – Point Accepted Mutation
RNA – ribonukleová kyselina
rRNA – ribozomová ribonukleová kyselina
SNP – jednonukleotidový polymorfismus
tRNA – transferová ribonukleová kyselina
UPGMA – Unweighted Pair Group Method with Arithmetic Mean
VNTRs – Variable Number Tandem Repeats

Seznam obrázků

Seznam tabulek