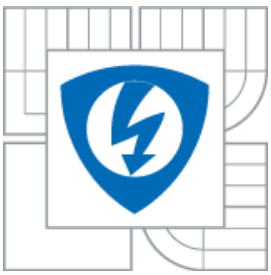
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



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

ZKRÁCENÁ VERZE DIZERTAČNÍ PRÁCE

AUTOR PRÁCE — Ing. DENISA MADĚRÁNKOVÁ
AUTHOR

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

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.

OBSAH

ÚVOD

Bioinformatika je 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). 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. 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 lze považovat za formu biologického digitálního signálu, a tak se zde nachází i možnost aplikace metod zpracování digitálních signálů. Sekvence v symbolickém zápisu musí být před zpracováním převedeny do vhodného numerického formátu. Výběr metody numerické reprezentace silně závisí na typu metody následné analýzy.

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 jsou čtyři číselné vektory, jeden pro každý typ nukleotidu. 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. K identifikaci vícebuněčných živočichů byl jako úsek pro DNA barcoding zvolena část sekvence pro gen cytochrom *c* oxidázu podjednotku 1 (*coxI* nebo *coI*) o přibližné délce 650 nukleotidů.

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. 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 je čeleď. Dále také byly pro testovací data zkonstruovány a porovnány dendrogramy ze standardně používaných evolučních vzdáleností a vzdáleností mezi nukleotidovými denzitními vektory.

1 TEORETICKÝ ÚVOD

Genomické sekvence se standardně zapisují jako sled znaků z abecedy *S{A, C, G, T}*, kde tyto znaky ve stejném pořadí zastupují výskyt bází adenin, cytosin, guanin a tymin. Symbolické sekvence nejsou vhodné pro analýzu metodami digitálního zpracování signálů a z tohoto důvodu se převádí do vhodného numerického formátu pomocí numerické mapy.

1.1 NUMERICKÉ REPREZENTACE

Ideální numerická mapa musí zajistit, že výsledný digitální genomický signál ponese stejnou informaci jako sekvence v symbolickém zápisu. Dále by numerická mapa neměla zavádět další informace, které původní sekvence nenese. 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. 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.

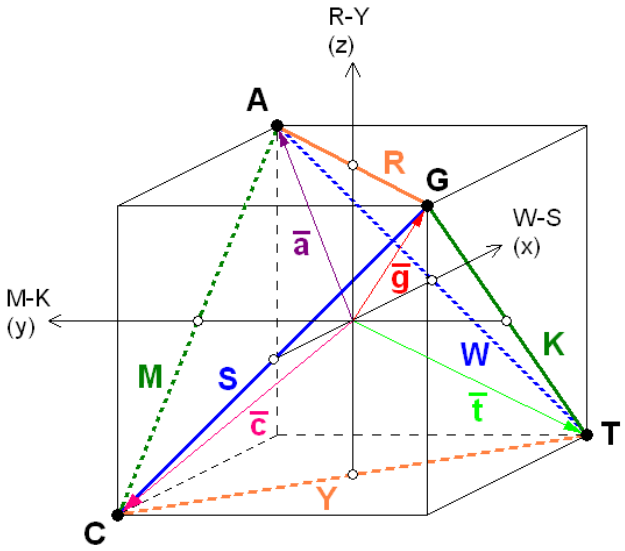
Nukleotidový čtyřstěn

Mezi biologicky významné vlastnosti sekvencí patří biochemické vlastnosti nukleotidů, podle kterých se dělí do skupin [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).

Podle tohoto dělení můžeme vytvořit reprezentaci čtyřstěnem (viz **Obr. 1.1**) [2]. Pomocí tohoto nukleotidového čtyřstěnu je možné odvodit numerické mapy bez ztráty informace. Jednotlivé nukleotidy jsou mapovány čtyřmi vektory symetricky umístěnými v prostoru a směřujícími do vrcholů čtyřstěnu. 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 chemických vlastnostech můžeme zanedbat (resp. která informace je pro nás zajímavá).

Existuje mnoho dalších numerických reprezentací, které však nevycházejí z modelu nukleotidového čtyřstěnu. Nejjednodušší numerická reprezentace je reálnými čísly, např.: A=1, C=2, G=3 a T=4. Tato reprezentace nenese žádné informace o chemických 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.



Obr. 1.1 Nukleotidový čtyřstěn umístěný v pomocné krychli [2].

Často používanou reprezentací je reprezentace binárními indikačními vektory [3], která nezachovává žádnou informaci o chemických vlastnostech DNA sekvence, ale zachovává informaci o periodicitě výskytu nukleotidů. 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 hodnotami 1 a 0. Tuto reprezentaci 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 a DNA sekvence je pak reprezentována třemi numerickými sekvencemi $x_r[n]$, $x_g[n]$ a $x_b[n]$: [4]

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

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

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

Existuje celá řada dalších numerických reprezentací, např.: 2D numerická reprezentace v 1. a 4. kvadrantu [5], 4D vektorová reprezentace [6], [7], jednodušší 2D reprezentace [8], různé grafické reprezentace vhodné k vizualizaci [9], [10], kumulovaná nebo rozbalená fáze [5], [11], či reprezentace EIIP hodnotami (Electron-Ion Interaction Potential) [12].

1.2 PRAKTICKÉ VYUŽITÍ NUMERICKÝCH MAP

Numerických map lze definovat mnoho, ale málokterá nalezla skutečně hodnotné praktické využití. V následujících odstavcích je diskutováno několik oblastí, ve kterých bylo aplikací numerických reprezentací a deterministických metod analýzy dosaženo hodnotných výsledků.

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. V nekódujících oblastech (tzv. introny) není charakteristický vzor zastoupení nukleotidů, a tudíž introny nevykazují stejnou periodicitu jako exony. [13],[14]

Pro predikci kódujících úseků DNA sekvencí existuje mnoho přístupů, např. se využívají HMM či metody založené na homologii sekvencí [15]. Principiálně jednoduchou signálově orientovanou metodou je vyhledání exonů pomocí diskrétní Fourierovy transformace (DFT). 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 reprezentace binárními indikačními vektory u_A, u_C, u_G a u_T. Následně se na všechny indikační vektory aplikuje diskrétní Fourierova transformace:

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

kde n je pořadí vzorku indikačního vektoru, $i = \sqrt{-1}$ a N je délka sekvence [16], [17].

Následně se vypočítá „výkonové“ spektrum:

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

Vysoká hodnota amplitudy spektrálního koeficientu pro frekvenci 1/3 Hz pak značí, že daný úsek sekvence je potenciální exon. 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 určuje poziční rozlišovací schopnost metody. 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.

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.

Tandemové repetice lze nalézt mnohými pravděpodobnostními a statistickými metodami a metodami původně určenými ke zpracování textu [18],[19]. Většina těchto metod potřebuje apriorně znát vzor, tj. nukleotidové složení tandemové repetice nebo alespoň délku vzoru. Nalezení repetic především znesnadňují bodové mutace vzoru.

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 [20],[21].

1.3 MITOCHONDRIÁLNÍ DNA

Mitochondrie, organela eukaryotických buněk, 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). 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. [22]

Genetický obsah mtDNA je u většiny organismů velmi podobný a je značně redukovaný, tj. mtDNA nekóduje všechny proteiny, které mitochondrie potřebuje ke svému fungování. U metazoí (vícebuněční živočichové) kóduje mtDNA 13 mRNA podléhajících translaci do proteinů, 22 tRNA a 2 rRNA [23]. 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 [24]. MtDNA podléhá větší mutační rychlosti než jaderná DNA.

Mitochondriální genom je populárním markerem molekulární diverzity v různých oblastech jako je populační biologie [25],[26], fylogenetika [27], fylogeografie [28] a molekulární ekologie. Popularita plyne z jednoduché extrakce mtDNA z buňky, absence intronů, inzercí a delecí, a především krátkost a jednoduchost tohoto genomu [29]. Využití mitochondriálního genomu k určení molekulární diverzity je založeno na třech předpokladech, které jsou však v současné době podrobovány značné diskuzi [27],[30].

Prvním předpokladem je klonalita, která říká, že mtDNA je předávána pouze po mateřské linii. Součástí předpokladu je také nerekombinační charakter mtDNA (vyjma hub a rostlin) [31][32]. Avšak již bylo pozorováno množství výjimek jak rekombinace, tak zachování otcovské DNA [33],[34]. Omezená platnost klonality musí být brána v potaz při interpretaci genealogických vazeb mezi druhy vzešlých z analýzy mitochondriálního genomu.

Neutrální mutace nebo mutace mírného delečního charakteru jsou druhým předpokladem. Adaptační mutace byly považovány za velmi vzácné [35]. Moderní výzkum odhalil, že tento předpoklad je také značně omezený [36],[37].

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 [38]. Bylo však prokázáno, že mitochondriální DNA mutuje rychleji než jaderná, což je pravděpodobně důsledek většího oxidativního stresu způsobeného funkcí mitochondrií [39],[40].

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 to nepředstavuje závažný problém. V současnosti je populární DNA barcoding používající k identifikaci organismů část mtDNA.

1.4 DNA BARCODING

DNA barcoding je přístup bez jednotné metodiky, který se pomocí krátké sekvence DNA organismu snaží identifikovat jeho příslušnost k jednotlivému živočišnému nebo rostlinnému druhu [41]. Identifikací druhů se rozumí přiřazení neznámého vzorku ke známému a popsanému druhu na základě shody znaků. 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. Práce Paula Heberta z roku 2003 [42] odstartovala praktické realizování tohoto přístupu s využitím mitochondriální DNA pro živočichy a chloroplastové DNA pro rostliny. Pro živočichy byla jako „univerzální“ DNA barcode sekvence vybrána část mitochondriálního genu *coxI* (cytochrom *c* oxidáza

podjednotka I) o přibližné délce 650 bp. Nelze obecně a jednoznačně říci, že úsek mtDNA *coxI* je optimální volbou pro DNA barcoding. 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 [43]). Existuje řada více či méně univerzálních postupů pro získání DNA barcodingových sekvencích pro různé skupiny organismů [44].

Identifikace druhů

K identifikaci druhů pomocí DNA barcode sekvencí se převážně používají distanční metody [45], [46]. 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. U distančních metod 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 [47]. Vnitrodruhová variabilita je však různá pro různé skupiny organismů a nelze mít nastavenu jedinou hodnotu neboť vnitrodruhová a mezidruhová variabilita často překrývají [48],[49]. Mnozí autoři navrhují používat k identifikaci raději znakově orientované metody, které by měly být pro identifikaci druhů spolehlivější [50], [51]. Pro znakovou analýzu DNA barcode sekvencí byl například vyvinut software CAOS (Characteristic Attributes Organization System) [52]. 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. Podobně jako CAOS pracuje i metoda BLOG (Barcoding with LOGic) [53]. Pro identifikaci druhů se vyzkoušely i další přístupy jako je strojové učení [54].

Kromě identifikace druhů Hebert navrhl, že by DNA barcodingové sekvence bylo možné využít k nalezení nových druhů, především těch morfologicky kryptických [55]. Tato možná aplikace rozvířila značnou diskuzi. 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. [50], [56]

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. [57]

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 [58] (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 [59], 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.

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 [60], ryby [61] či obojživelníci [62] nebo rozdílné morfologie sociálního hmyzu [63]. DNA barcoding take může pomoci při odhalování blízce příbuzných a kryptických druhů (např.: [64], [65], [66]), 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 [67], [68], nebo k ověření složení přírodních léčiv [69] a dalším kontrolním testům.

2 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 3. Body 3a. a 3b. cílů jsou zpracovány v kapitolách 4, 5 a 6.

3 NUKLEOTIDOVÉ DENZITNÍ VEKTORY

Metoda nukleotidových denzitních vektorů představuje jednoduchý a efektivní způsob numerické reprezentace symbolické sekvence DNA. 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.

3.1 VÝPOČET DENZITNÍCH VEKTORŮ

Pro výpočet nukleotidových denzitních vektorů (ND vektorů) se nejprve ze symbolické 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 reprezentující nukleotidy o určité biochemické vlastnosti se indikačních vektorech projeví hodnotou 0,5.

ND 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 ND vektory jsou kratší než původní sekvence. Např. pro sekvenci o délce M=100 bp a velikost výpočetního posuvného okna W=5 nukleotidů získáme ND vektory o M-W+1=96 hodnotách. Výpočetní okno o velikosti W tak zkrátí ND 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.

Po úpravě (rozšíření) binárních indikačních vektorů se počítají nukleotidové 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 nukleotidových denzitních vektorů. Velikost posuvného okna je volitelná. Volba velikosti okna je závislá na délce sekvence a na požadované rozlišovací schopnosti výsledného signálu.

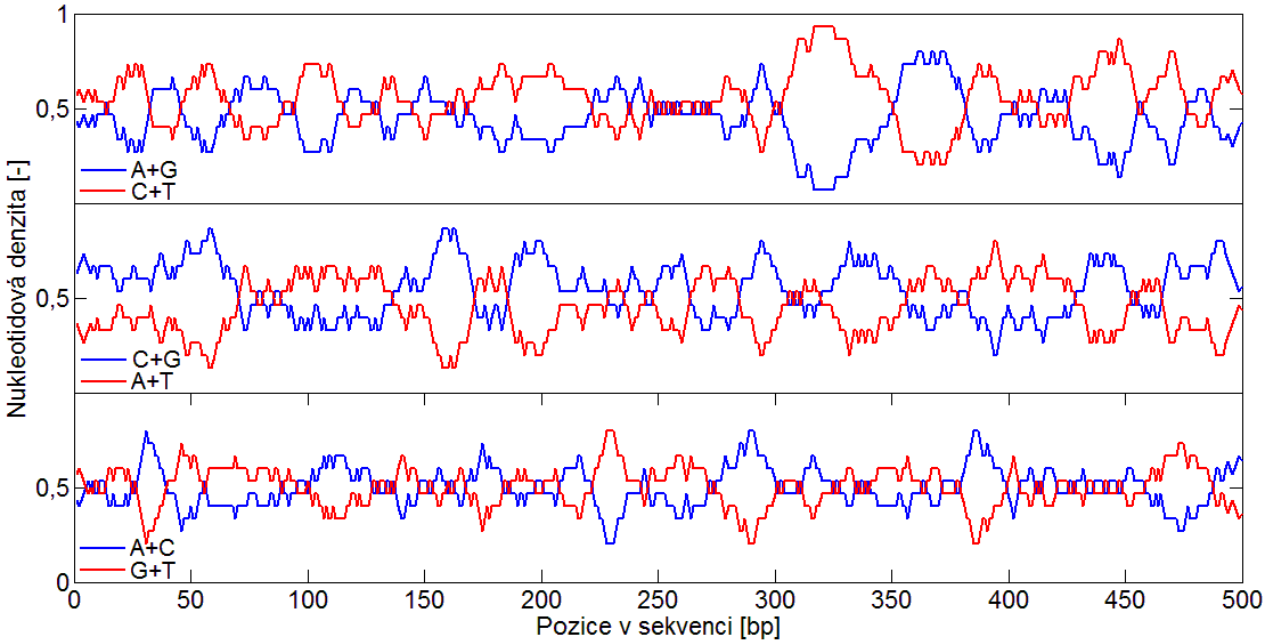
Následující **Tab. 3.1** zobrazuje příklad tvorby binárních indikačních vektorů a nukleotidových denzitních vektorů pro velikost výpočetního okna W=5 nukleotidů s použitím doplnění binárních indikačních vektorů o hodnoty 0,25 na začátky a konce vektorů.

Grafická reprezentace nukleotidových denzitních vektorů je možná několika způsoby. Nejjednodušší je samostatné vykreslení jednotlivých ND vektorů. Tato vizualizace poskytuje

informaci o výskytu jednotlivých nukleotidů v rámci sekvence a můžeme například rychle identifikovat části sekvence s převažujícím výskytem určitého nukleotidu. Další způsob vizualizace tkví v zobrazení sum ND vektorů pro nukleotidy podobných biochemických vlastností, viz **Obr. 3.1**. V tomto typu vykreslení jsou na první pohled dobře patrná komplementarita jednotlivých kategorií vlastností (purin/pyrimidin, silná/slabá vazba, amino/keto skupina).

Tab. 3.1 Příklad tvorby nukleotidových denzitních vektorů pro W=5.

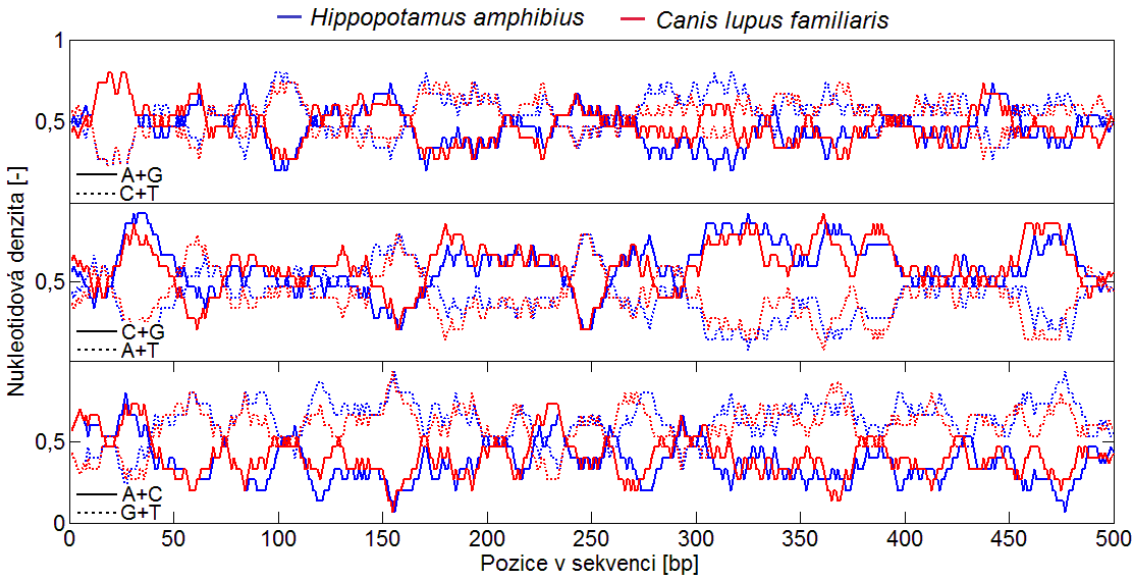
	Sekvence:	**A**	**G**	**T**	**T**	**N**	**A**	**C**	**R**	**A**	**C**		
Binární indikační vektory	b_A	.25	.25	1	0	0	0	.25	1	0	.5	1	0	.25	.25
	b_C	.25	.25	0	0	0	0	.25	0	1	0	0	1	.25	.25
	b_G	.25	.25	0	1	0	0	.25	0	0	.5	0	0	.25	.25
	b_T	.25	.25	0	0	1	1	.25	0	0	0	0	0	.25	.25
Nukleotidové denzitní vektory	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		



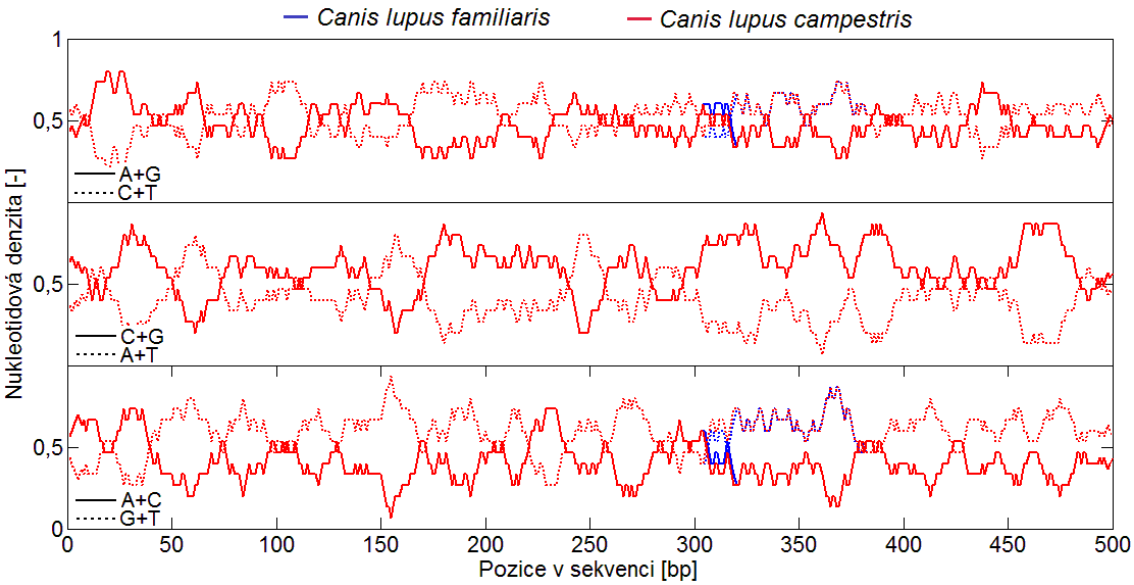
Obr. 3.1 Sumy nukleotidových denzitních vektorů části mitochondriálního genu *coxI* organismu *Canis lupus campestris* dle biochemických vlastností nukleotidů, W=15.

3.2 VLASTNOSTI ND VEKTORŮ A JEJICH PRAKTICKÉ VYUŽITÍ

Teoretická pravděpodobnost, že bychom pro dvě rozdílné nukleotidové sekvence získali stejné nukleotidové denzitní vektory, je velmi malá avšak ne nulová. Nukleotidové denzitní vektory jsou závislé na pozicích jednotlivých nukleotidů a na velikosti posuvného výpočetního okna. Zpětná rekonstrukce symbolické sekvence z ND vektorů 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.



Obr. 3.2 Sumy nukleotidových denzitních vektorů dvou mitochondriálních sekvencí pro 12S rRNA o délce 500 bp organismů *Hippopotamus amphibius* a *Canis lupus familiaris*, $W = 15$.



Obr. 3.3 Sumy nukleotidových denzitních vektorů dvou mitochondriálních sekvencí pro 12S rRNA o délce 500 bp organismů *Canis lupus familiaris* a *Canis lupus campestris*, W=15.

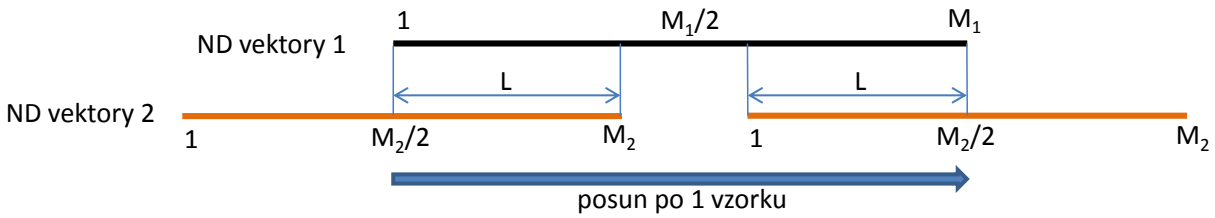
ND vektory mohou najít uplatnění v mnohých oblastech analýzy nukleotidových sekvencí. Jelikož vektory vytváří jakýsi charakteristický vzor pro danou sekvenci, jejich využití ke komparativní genomice se samo nabízí. ND vektory by šlo také využít k rychlému vyhledávání známého motivu v sekvenci nebo homologních sekvencí s použitím metod, které se využívají ve zpracování digitálních signálů. Lokální charakter zastoupení nukleotidů může pomoci při vyhledávání CpG ostrůvků, tj. oblastí s vysokým výskytem cytosinu a guaninu. Konzervovanost některých vzorů v nukleotidových denzitních 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)[70].
Grafické zobrazení ND 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. Naopak sekvence jednoho genu ale jiných organismů

mají značně podobný charakter. Čím příbuznější organismy, tím podobnější ND vektory, viz **Obr. 3.2** pro dva savce *Hippopotamus amphibius* (hroch obojživelný) a *Canis lupus familiaris* (pes domácí). Na **Obr. 3.3** 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.

3.3 IDENTIFIKACE S VYUŽITÍM ND VEKTORŮ

V této dizertační práci jsou ND vektory využity ke komparativní genomice v rámci identifikace do taxonomických skupin na základě porovnávání DNA barcodingových sekvencí. Byly navrženy tři metody založené na ND 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. 3.4** zobrazuje schematicky počáteční a koncové zarovnání obou ND vektorů.



Obr. 3.4 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.

Metoda mediánu ND vektorů

Pro každou sekvenci jedince daného druhu jsou vypočítány ND vektory. ND vektory všech jedinců druhu (nebo vyšší taxonomické skupiny) jsou signálově zarovnány tak, že se nalezne jejich

vzájemná pozice, která minimalizuje Euklidovskou vzdálenost mezi nukleotidovými denzitními vektory, viz rovnice (5.2). Výsledné referenční ND vektory jsou tvořeny mediánovými hodnotami zarovnaných ND vektorů.

Metoda průměru ND vektorů

Pro každou sekvenci jedince daného druhu jsou vypočítány nukleotidové denzitní vektory. ND vektory všech jedinců jsou zarovnány tak, že se nalezne jejich vzájemná pozice, která minimalizuje Euklidovskou vzdálenost dle rovnice (5.2) mezi nimi. Výsledné referenční ND vektory jsou tvořeny průměrnými hodnotami zarovnaných ND vektorů.

Metoda průměru purinových ND vektorů

Referenční ND vektory pro druh se tvoří obdobně jako v předchozí metodě průměru ND 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).

Metoda identifikace

Pro identifikaci neznámé sekvence se postupuje následovně. Pro sekvenci se vypočítají ND vektory a ty se signálově zarovnají postupně se všemi referenčními ND vektory. Signálové zarovnání znamená, že se nalezne vzájemná pozice ND vektorů taková, která dává nejmenší hodnotu délkově normalizované Euklidovské vzdálenost na 1000 nukleotidů, zde označovaná jako END vzdálenost (Euklidovská vzdálenost ND vektorů):

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

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 denzitních 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. Vedle identifikace na základě Euklidovských vzdáleností mezi ND vektory všech nukleotidů byla navrhnuta ještě identifikace porovnáváním ND vektorů referencí tvořených metodou průměru purinových nukleotidů. Tuto vzdálenost označujme jako PEND (Purinová Euklidovská vzdálenost ND vektorů) a počítá se dle vzorce:

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

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.

4 DRUHOVÁ IDENTIFIKACE

Sekvence z databáze BOLD lze získat ve formě souhrnného FASTA souboru zvoleného DNA 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.

4.1 REFERENČNÍ DRUHY

Metody tvorby referenčních ND vektorů a metody identifikace byly otestovány na sekvencích živočišných druhů různých skupin organismů. 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.

Pro vytváření referencí se ze souborů 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.

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 jako medián a průměr ND vektorů, proto byly reference druhů vytvořeny pouze metodou průměru ND vektorů a metodou průměru ND vektorů purinových nukleotidů.

Soubory sekvencí i identifikaci 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. 4.1** 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.

4.2 VÝSLEDKY IDENTIFIKAČNÍ ANALÝZY

Všechny sekvence analyzovaného souboru, které se mají identifikovat, se převedou na ND vektory. Následně se každá sekvence porovnává se všemi referencemi a je přiřazena k tomu referenčnímu druhu, se kterým má nejmenší Euklidovskou vzdálenost.

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.

Tab. 4.1 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

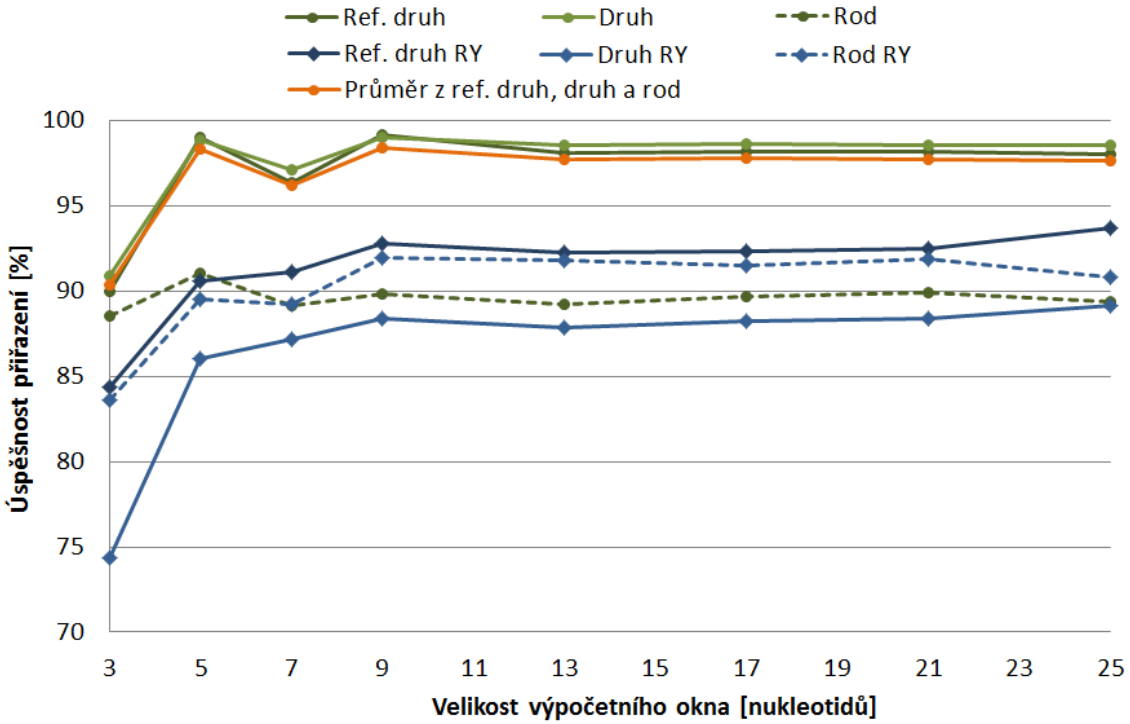
Ref. sek. – počet sekvencí použitých k vytvoření druhových referencí
Rodové sek. – počet sekvencí stejného rodu ale jiného druhu než jsou druhové reference

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.

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. 4.1** 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.

Tab. 4.2 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.



Obr. 4.1 Vážené průměry úspěšností identifikační analýzy v závislosti na velikosti výpočetního okna pro všechny soubory.

Tab. 4.2 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. 4.1** 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 denzitních 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.

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 [59]. 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 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.

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 % [54]. 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 denzitnách 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 % [71].

5 IDENTIFIKACE DO ČELEDÍ

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. To znamená, dostatečné množství nukleotidů je na svých pozicích v rámci čeledě konzervovaných a nedochází k jejich mutacím.

Byly testovány tři varianty vytváření referenčních nukleotidových denzitních vektorů pro čeledě a to použití mediánové, průměrné a průměrné hodnoty jen purinových nukleotidů signálově zarovnaných ND vektorů sekvencí patřící do téže čeledě. 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 je analyzovaná sekvence 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.

5.1 REFERENČNÍ DATABÁZE

Reference čeledí byly vytvořeny ze sekvencí živočišné třídy ptáků. Pokud bude metoda identifikace do vyšších taxonomických jednotek opravdu robustní a výše zmíněný předpoklad o dostatečném informačním obsahu sekvencí platný, měla by být metoda identifikace schopna správně zatřídit sekvence ptáků z referenčních čeledí, i když pocházející z druhů vyskytujících se na jiném kontinentu. Třída ptáků byla vybrána z důvodu dobře propracované taxonomie, se kterou můžeme výsledky porovnávat.

Výběr samotného projektu se sekvencemi pro vytvoření referencí byl náhodný a bez odůvodnění vhodnosti tohoto projektu k vytváření referenční databáze. Reference čeledí vyly vytvořeny obdobným způsobem jako druhové reference z DNA barcoding projektu Birds of Argentina – Phase I (dále jen BARG), které byly jako FASTA soubor staženy z databáze BOLD. Celkem se v souboru nachází 1588 sekvencí. Pro vytváření referencí byly vybrány pouze sekvence delší nebo rovny 600 bp. V **Tab. 5.1** 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.

Referenční ND vektory pro čeledě vypočítané metodami mediánu ND vektorů, průměru ND vektorů a průměru ND vektorů jen purinových nukleotidů se od sebe navzájem liší a liší se také nukleotidové denzity konsenzuální sekvence pro čeleď. Rozdíly vznikly díky variabilitě v rámci čeledí, kterou každá z metod reprezentuje jiným způsobem. Rozdíly mezi referenčními ND vektory úzce souvisí s variabilitou čeledí. Čím větší variabilita čeledí, tím větší rozdíly mezi referenčními ND vektory jednotlivých metod.

Tab. 5.1 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

5.2 VERIFIKACE IDENTIFIKACE DO ČELEDÍ

Verifikace spočívá v identifikaci do čeledí těch stejných sekvencí, z kterých byly reference čeledí vytvořeny. 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 reference (RM), průměrné reference (RP) a průměrné reference jen purinových nukleotidů (RP-RY). Analýza byla provedena pro velikosti výpočetního okna od W=5 do W=19 nukleotidů pro výpočty ND vektorů. Velikost výpočetního okna W=3 nukleotidy nebyla vzhledem k nízké míře úspěšnosti identifikace sekvencí do druhů použita.

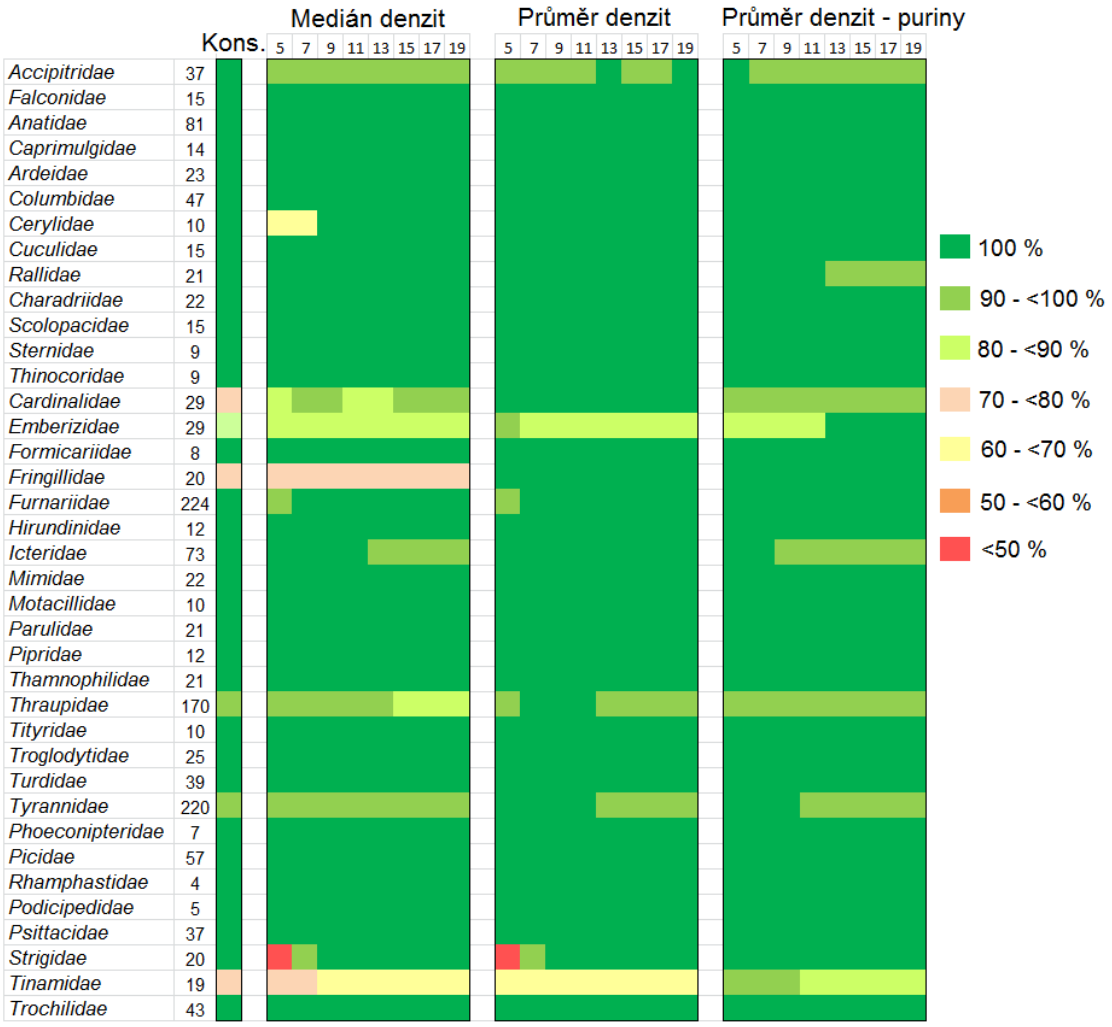
Obr. 5.1 zjednodušeně graficky znázorňuje procentuální úspěšnosti identifikace do čeledí při verifikaci všemi metodami využívající ND vektory v závislosti na délce výpočetního okna a porovnání s metodikou konsenzuální sekvence. Vertikální směr odpovídá pořadí čeledí; horizontální směr odpovídá délce výpočetního okna W=5, 7,…, 19 nukleotidů ve směru zleva doprava. Je vidět, že metoda průměrných denzit jen purinových nukleotidů má celkově nejlepší výsledky. U metody konsenzuální sekvence jsou hodnoty úspěšnosti identifikace do čeledí podobné. Vážený průměr úspěšnosti identifikace čeledí metodou RK byl 96,49 %, nejlepší hodnota pro metodu RM byla 98,08 % a pro metodu RP 99,31 % shodně pro výpočetní okna délek W=9 a W=11 nukleotidů a nejlepšího výsledku dosáhla metoda RP-RY s 99,04 % pro W=5 nukleotidů.

5.3 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. 5.2** 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.

Všechny sekvence z vyjmenovaných projektů byly přiřazovány do referenčních čeledí vytvořených ze sekvencí souboru BARG pomocí metod konsenzuální sekvence, mediánové denzity, průměrné denzity a průměrné denzity 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 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.



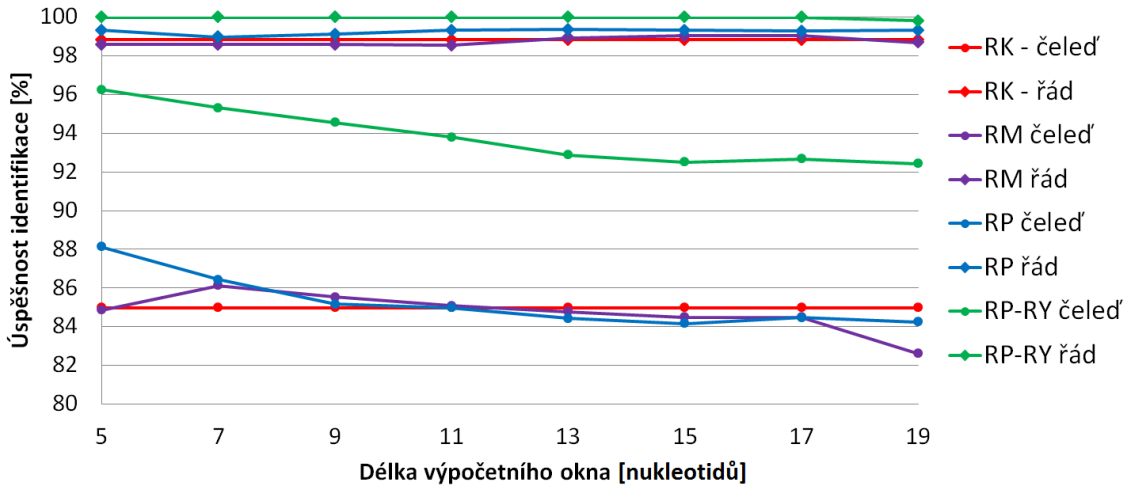
Obr. 5.1 Grafické znázornění procentuální úspěšnosti identifikace do čeledí metodami konsenzus (Kons. = RK), medián denzit (RM), průměr denzit (RP) a průměr denzit jen purinových nukleotidů (RP-RY) při délkách výpočetního okna *W*=5 až *W*=19 nukleotidů (vertikálně) pro soubor BARG.

Tab. 5.2 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

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. **Obr. 5.2** 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 nukleotidových denzit mají velmi podobné výsledky. Metoda průměru purinových nukleotidových denzit 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 metody liší jen málo a pro všechny metody jsou úspěšnosti nad 98 %.



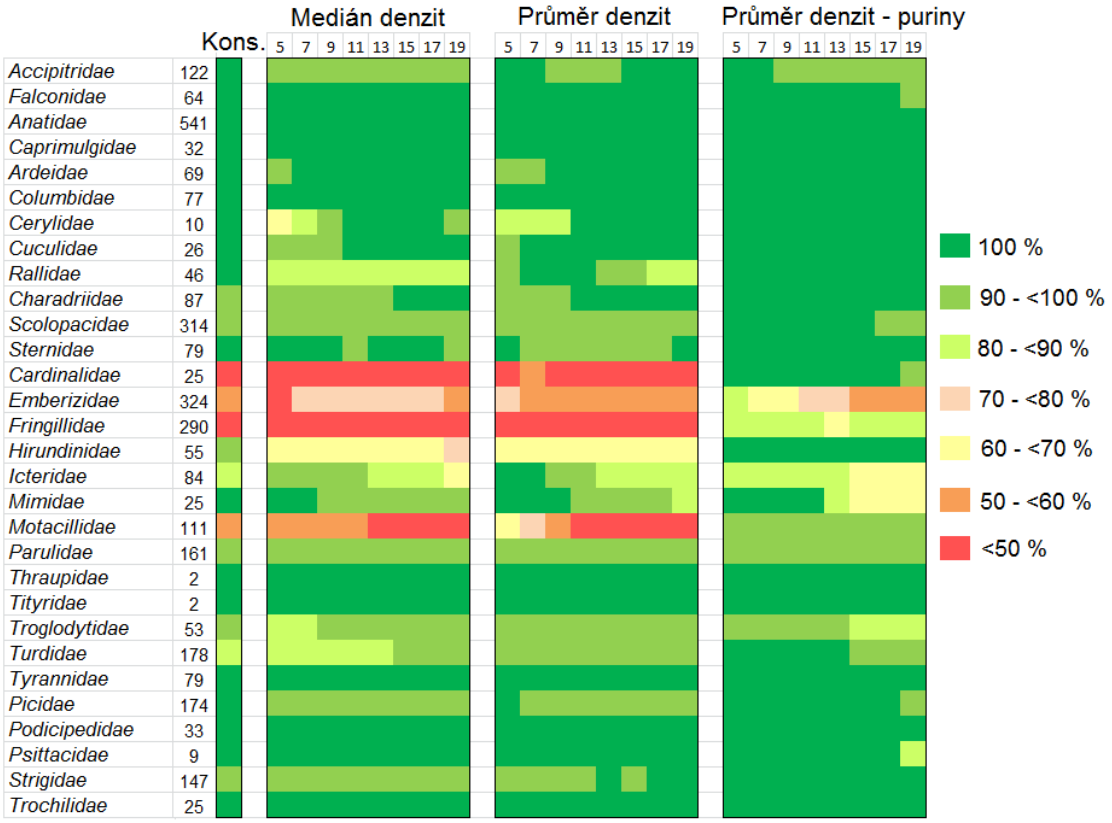
Obr. 5.2 Vážené průměrné úspěšnosti identifikace sekvencí ze všech souborů do čeledí a řádů v závislosti na velikosti výpočetního okna. RK – metoda konsenzuální sekvence, RM – metoda mediánové denzity, RP – metoda průměrné denzity, RP-RY – purinová identifikace.

Na **Obr. 5.3** je graficky znázorněna procentuální úspěšnost identifikace do čeledí pro všechny soubory dohromady (mimo referenční soubor BARG) a na **Obr. 5.4** 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 denzit jen purinových nukleotidů.

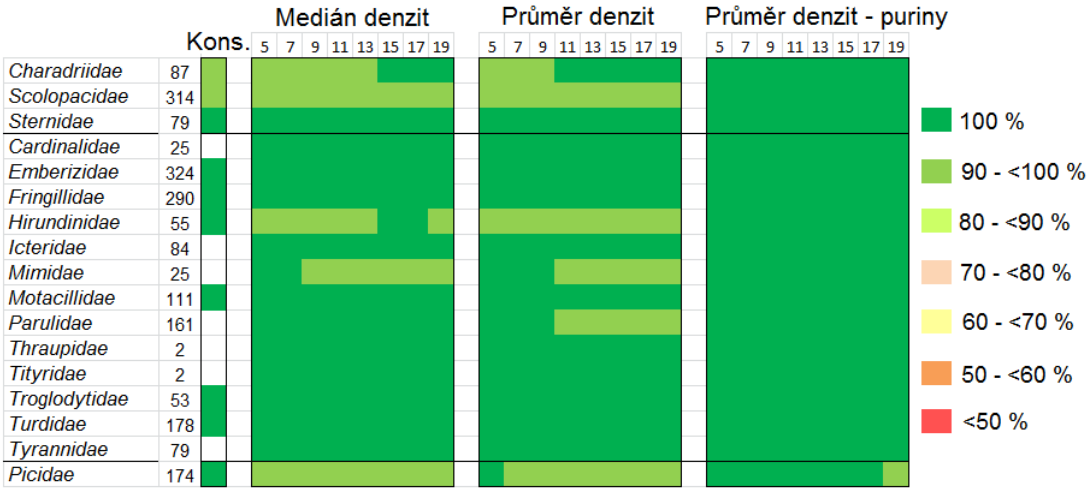
Nejmenší úspěšnosti při identifikaci do čeledí měly sekvence náležící do čeledí Cardinalidae, Emberizidae a Fringillidae z řádu Passeriformes, které byly nesprávně přiřazovány především k čeledím Icteridae, Parulidae nebo Thraupidae ze stejného řádu. Všechny tyto zmíněné čeledě patří do jedné podskupiny 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.

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.



Obr. 5.3 Grafické znázornění procentuální úspěšnosti identifikace do čeledí sekvencí ze všech analyzovaných souborů.



Obr. 5.4 Grafické znázornění procentuální úspěšnosti identifikace do řádů sekvencí ze všech souborů.

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. [72]

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 denzit 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 2221 sekvencí. 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.

6 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. 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.

Byla provedena srovnávací analýza dendrogramů vybraných datasetů sekvencí, které byly použity k identifikační analýze do druhů a čeledí (viz kapitola 4 a 5). 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 JC a K2P korekcí. 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 W=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. 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. Kondenzované dendrogramy 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 = DLP,
- poměr počtu listů kondenzovaného dendrogramu pro rod k počtu rodů v datasetu = RLP,
- poměr počtu listů kondenzovaného dendrogramu pro vyšší taxonomickou skupinu (např. čeleď) k počtu těchto skupin v datasetu = VLP.

Dendrogramy byly zkonstruovány pro projekty CUCAD, GBAP, MNCN, BCBNC a BARG. V **Tab. 6.1** jsou shrnuty kondenzační poměry DLP, RLP a VLP pro jednotlivé soubory a metody výpočtu vzdáleností mezi sekvencemi.

Tab. 6.1 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.

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 [43].

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.

7 ZÁVĚR

Cílem dizertační práce bylo navrhnout numerickou reprezentaci DNA sekvencí, která bude vhodná pro komparativní genomiku se zaměřením na identifikaci druhů. K tomuto účelu byla navržena metoda výpočtu nukleotidových denzitních vektorů, které vyjadřují průměrné zastoupení jednotlivých druhů nukleotidů v definované oblasti DNA sekvence. Dále byla navržena metodika porovnávání ND vektorů pro identifikační účely založená na výpočtu Euklidovské vzdálenosti mezi vektory. Navržená numerická reprezentace a metodika porovnávání byly otestovány na rozsáhlém souboru DNA barcodingových sekvencích získaných z volně dostupné databáze BOLD. Kromě testování identifikace do druhů byla provedena i podobná analýza identifikace do čeledí a řádů. Na závěr byla provedena ještě komparativní analýza dendrogramů konstruovaných ze standardně používaných evolučních vzdáleností a Euklidovských vzdáleností mezi ND vektory.

Pro identifikační analýzu do druhů bylo vytvořeno 751 druhových referencí metodou průměru ND vektorů z celkového počtu 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 %.

Pro identifikaci do čeledí bylo vytvořeno 38 referencí čeledí z 1450 sekvencí 445 různých druhů argentinských ptáků metodami konsenzuální sekvence (RK), mediánu ND vektorů (RM), průměru ND vektorů (RP) a průměru ND vektorů jen purinových nukleotidů (RP-RY). Následně bylo dalších celkem 3244 sekvencí přiřazováno k čeledím, přičemž tyto sekvence pocházely především z ptáků ze Severní Ameriky a Euroasijského kontinentu. Pro tvorbu referencí byly použity velikosti výpočetního okna v intervalu W=5 až W=19 nukleotidů. „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.

Na závěr byly konstruovány dendrogramy pro vybrané soubory sekvencí metodou spojování sousedů ze zavedených Jukesových-Cantorových (JC) a Kimurových 2-parametrických (K2P) evolučních vzdáleností a Euklidovských vzdáleností mezi denzitními vektory všech nukleotidů (END) a Euklidovských vzdáleností mezi sumou denzit purinových nukleotidů (PEND). 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.

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

[3] 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.

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

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

[6] 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.

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

[8] RANDIĆ, M. et al. Novel 2-D graphical representation of DNA sequences and their numerical characterization. *Chemical Physics Letters*. 2003, č. 368. S. 1-6.

[9] 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.

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

[11] ZHANG, Y. et al. On 2D graphical representation of DNA sequence of nondegeneracy. *Chemical Physics Letters*. 2005, č. 411. S. 28-32.

[12] 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.

[13] ESKESEN, S. T. et al. Periodicity of DNA in exons. *BMC Molecular Biology*. 2004, č. 5(12).

[14] SILVERMAN, D. B. a LINSKER, R. A measure of DNA periodicity. *Journal of Theoretical Biology*. 1986, č. 118(3). S. 295-300.

[15] DO, J. H. a CHOI, D.-K. Computational Approaches to Gene Prediction. *The Journal of Microbiology*. 2006, č. 44(2). S. 137-144.

[16] TIWARI, S. et al. Prediction of probable genes by Fourier analysis of genomic sequences. *Computer Applications in the Biosciences*. 1997, č. 13(3). S. 263-270.

[17] AFREIXO, V. et al. Fourier analysis of symbolic data: A brief review. *Digital Signal Processing*. 2004, č. 14. S. 523-530.

[18] 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.

[19] SOKOL, D. et al. Tandem repeats over the edit distance. *Bioinformatics*. 2007, č. 23(2). S. e30-e35.

[20] ANASTASSIOU, D. Frequency-domain analysis of biomolecular sequences. *Bioinformatics*. 2000, č. 16(12). S. 1073-1081.

[21] DIMITROVA, N. et al. 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.

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

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

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

[25] MORITZ, C. et al. Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Annual Review of Ecology Evolution and Systematics*. 1987, č. 18. S. 269-292.

[26] 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.

[27] GALTIER, N. et al. Mitochondrial DNA as a marker of molecular diversity: reappraisal. *Molecular Ecology*. 2009, č. 18. S. 4541-4550.

[28] AVISE, J. C. 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.

[29] GISSI, C. et al. Evolution of the mitochondrial genome of metazoa as exemplified by comparison of congeneric species. *Heredity*. 2008, č. 101. S. 301-320.

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

[31] WHITE, D. J. et al. Revealing the hidden complexities of mtDNA inheritance. *Molecular Ecology*. 2008, č. 17. S. 4925-4942.

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

[33] EYRE-WALKER, A. et al. How clonal are human mitochondria? *Proceedings of the Royal Society B-Biological Sciences*. 1999, č. 266. S. 477-483.

[34] AWADALLA, P. et al. Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science*. 1999, č. 286. S. 2524-2525.

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

[36] GROSSMAN, L. I. et al. Accelerated evolution of the electron transport chain in anthropoid primates. *Trends in Genetics*. 2004, č. 20. S. 578-585.

[37] CASTOE, T. A. 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.

[38] WILSON, A. C. et al. Biochemical evolution. *Annual Review of Biochemistry*. 1977, č. 46. S. 573-639.

[39] BROWN, W. M. et al. 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.

[40] 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.

[41] 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.

[42] BRUNO, W. J. et al. Weighted Neighbor Joining: A Likelihood-Based Approach to Distance-Based Phylogeny Reconstruction. *Molecular Biology and Evolution*. 2000, č. 17(1). S.189-197.

[43] VENCES, M. et al. Deciphering amphibian diversity through DNA barcoding: chances and challenges. *Philosophical Transactions of the Royal Society B-Biological Sciences*. 2005, č. 360. S. 1859-1868.

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

[45] HEBERT, P. D. N. et al. Identification of birds through DNA barcodes. *PLoS Biology*. 2004, č. 2(10). S. 1657-1663.

[46] CYWINSKA, A. et al. Identifzing Canadian mosquito species through DNA barcodes. *Medical and Veterinary Entomology*. 2006, č. 20(4). S. 413-424.

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

[48] 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.

[49] HICKERSON, M. J. et al. DNA Barcoding Will Often Fail to Discover New Animal Species over Broad Parameter Space. *Systematic Biology*. 2006, č. 55(5). S. 729-739.

[50] DESALLE, R. et al. 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.

[51] BERGMANN, T. et al. Character-based DNA barcoding: a superior tool for species classification. *Berliner Und Munchener Tieraryliche Wochenschrift*. 2009, č. 122. S. 446-450.

[52] SARKAR, I. N. et al. CAOS software for use in character-based DNA barcoding. *Molecular Ecology Resources*. 2008, č. 8. S. 1256-1259.

[53] WEITSCHEK, E. et al. 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.

[54] WEITSCHEK, E. et al. Supervised DNA Barcodes species classification: analysis, comparisons and results. *BioData Mining*. 2014, č. 7(4).

[55] HEBERT, P. D. N. et al. 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.

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

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

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

[59] 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.

[60] RUITER, D. E. et al. DNA barcoding facilitates associations and diagnoses for Trichoptera larvae of the Churchill (Manitoba, Canada) area. *BMC Ecology*. 2013, č. 13(5).

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

[62] VENCES, M. et al. DNA Barcoding Amphibians and Reptiles. *Methods in Molecular Biology*. 2012, č. 858. S. 79-107.

[63] SMITH, M. A. et al. 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.

[64] HEBERT, P. D. N. 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.

[65] BARCO, A. 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.

[66] HSU, T.-H. et al. DNA barcoding reveals cryptic diversity in the peanut worm *Sipunculus nudus*. *Molecular Ecology Resources*. 2013, č. 13(4). S. 596-606.

[67] MARALIT, B. A. 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.

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

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

[70] ŠKUTKOVÁ, H. et al. Classification of genomic signals using dynamic time warping. *BMC Bioinformatics*. 2013, č. 14. (Suppl 10):S1.

[71] ZHANG, A. B. et al. Inffering Species Membership Using DNA Sequences with Back-Propagation Neural Networks. *Systematic Biology*. 2008, 57(2), 202-2015.

[72] WILSON, J. J. et al. When species matches are unavailable are DNA barcodes correctly assigned to higher taxa? An assessment using sphingid moths. *BMC Ecology*. 2011, vol. 11(18).

CURRICULUM VITAE

Jméno a Příjmení:	Ing. Denisa Maděránková
Adresa:	Sportovní 282, 664 61 Opatovice
Telefon:	+420 775 025 945
Email:	maderankova@feec.vutbr.cz

Vzdělání

2008–2015 Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, obor Biomedicínská elektronika a biokybernetika, doktorské studium.

2006–2008 Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních, technologií, Ústav biomedicínského inženýrství, obor Biomedicínské a ekologické inženýrství, prezenční magisterské studium, ukončeno státní závěrečnou zkouškou.

2003–2006 Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních, technologií, Ústav mikroelektroniky, obor Mikroelektronika a technologie, prezenční bakalářské studium, ukončeno státní závěrečnou zkouškou.

Profesní zkušenosti

2009–2015 Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství
Pozice: Technicko-hospodářský pracovník
Výuka předmětů: Bioinformatika, Praktika z bioinformatiky, Programování v bioinformatice
Vědecko-výzkumná činnost: návrh a realizace nových metod zpracování genomických a proteomických dat

2007–2008 Ústav analytické chemie AV ČR, v.v.i.
Pozice: Laborant vědeckého oddělení
Náplň práce: Ramanova spektroskopie

2006-2007 BVT Technologies, a.s.
Pozice: Laborant vývoje
Náplň práce: Testování a výstupní kontrola mikroprůtokových zařízení a biosenzorů

Ocenění

2008 Cena děkana FEKT VUT v Brně za diplomovou práci

Projekty

2011 ŠKUTKOVÁ, H.; MADĚRÁNKOVÁ, D.; PROVAZNÍK, I. Inovace výuky v předmětech zaměřených na genomiku a proteomiku, FRVŠ 719/G3, zahájení: 1. 1. 2011, ukončení: 31. 12. 2011.

2009 SEKORA, J.; MADĚRÁNKOVÁ, D.; PROVAZNÍK, I. Inovace výuky v předmětech zaměřených na využití informačních technologií v medicíně na FEKT VUT v Brně, FRVŠ 2597/G1, zahájení: 1. 1. 2009, ukončení: 31. 12. 2009.

Vybrané publikace

2013 MADĚRÁNKOVÁ, D.; PROVAZNÍK, I. Classification of species to higher taxa based on analysis of DNA barcodes - a bird example. In The 10th International Workshop on Computational Systems Biology. Tampere: 2013. s. 75-79. ISBN: 978-952-15-3092- 0.

2011 MADĚRÁNKOVÁ, D.; PROVAZNÍK, I. Motives in Nucleotide Densities of Birds Mitochondrial Gene COX1. In ACM Digital Library: Proceedings of 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies. Barcelona: 2011. s. 1-5. ISBN: 978-1-4503-0913- 4.

Produkty

2014 MADĚRÁNKOVÁ, D.; PROVAZNÍK, I.: R_ NucDen; Balíček funkcí v jazyce R pro výpočet nukleotidových denzitních vektorů a jejich komparativní analýzu. Ústav biomedicínského inženýrství, FEKT VUT v Brně, Technická 12, 61200 Brno. (software)

2013 MADĚRÁNKOVÁ, D.; PROVAZNÍK, I.: *Nástroj pro analýzu nukleotidových denzitních vektorů pro identifikaci organismů*. Ústav biomedicínského inženýrství, FEKT VUT v Brně, Technická 12, 61200 Brno. (software)