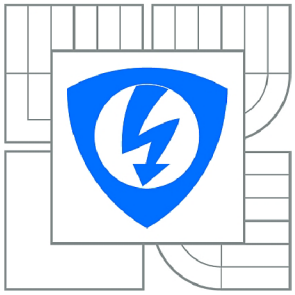
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGINEERING

VLIV REDUKCE NUKLEOTIDOVÝCH DENZITNÍCH VEKTORŮ NA MOLEKULÁRNÍ IDENTIFIKACI ORGANISMŮ

EFFECT OF NUCLEOTIDE DENSITY VECTOR'S REDUCTION ON MOLECULAR IDENTIFICATION OF ORGANISMS

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

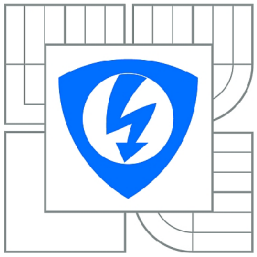
AUTOR PRÁCE — MARKÉTA KOSKOVÁ

AUTHOR

VEDOUCÍ PRÁCE — Ing. DENISA MADĚRÁNKOVÁ

SUPERVISOR

BRNO 2014



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor
Biomedicínská technika a bioinformatika

Studentka: Markéta Kosková ***ID:*** 147520
Ročník: 3 ***Akademický rok:*** 2013/2014

NÁZEV TÉMATU:

Vliv redukce nukleotidových denzitních vektorů na molekulární identifikaci organismů

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte literární rešerši na téma molekulární identifikaci organismů. 2) V libovolném programovém prostředí naprogramujte výpočet nukleotidových denzitních vektorů ze sekvencí DNA s různými možnostmi ošetření okrajových částí sekvencí. 3) Z volně dostupných databází vyberte sekvence jednoho genu pro soubor blízce příbuzných druhů i druhů z různých skupin organismů. 4) Zanalyzujte vliv ošetření okrajů denzitních vektorů na souboru sekvencí. 5) Proveďte identifikační analýzu pomocí porovnávání nukleotidových denzit referenčního souboru sekvencí s nezávislým souborem sekvencí stejných druhů. Vyhodnoťte úspěšnosti identifikace při použití samostatných denzitních vektorů pro jednotlivé nukleotidy a sumy purinových/pyrimidinových nukleotidů.

DOPORUČENÁ LITERATURA:

[1] SCHEFFLER, Immo E. Mitochondria. 2nd ed., Wiley-Blackwell, 2007, 472 s. ISBN 978-0-470-04073-7.
[2] BLAXTER, Mark. The promise of a DNA taxonomy. Phil. Trans. R. Soc. Lond. B., 2004, vol. 359, pp. 669-679.

Termín zadání: 10.2.2014 ***Termín odevzdání:*** 30.5.2014

Vedoucí práce: Ing. Denisa Maděránková
Konzultanti bakalářské práce:

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

ABSTRAKT

Tato práce se zabývá vlivem redukce nukleotidových denzitních vektorů na molekulární identifikaci organismů. Na začátku teoretické části je popsán pojem druh, jeho vývoj a s ním související mutace. V další časti je uvedena taxonomie, především ta molekulární, a DNA barcoding spolu s popisem mitochondriální DNA a jeho využitím k identifikaci druhů. Na konci teoretické části se nachází informace o nukleotidových denzitních vektorech, na které navazuje praktická část. V ní je pomocí Matlabu provedena analýza ošetření okrajů denzitních vektorů na 188 sekvencích. S nejlepší hodnotou je následně provedena identifikační analýza organismů a její vyhodnocení při použití denzitních vektorů pro jednotlivé nukleotidy a jejich sumy. To pro 4 soubory referenčních a 5 souborů analyzovaných sekvencí stejných druhů.

KLÍČOVÁ SLOVA

Druh, molekulární taxonomie, DNA barcoding, nukleotidové denzitní vektory.

ABSTRACT

This work deals with the effect of nucleotide density vector's reduction on molecular identification of organisms. At the beginning of the theoretical part, the work explains the concept of species, its development and its mutations. The next section provides basics of taxonomy, particularly the molecular taxonomy and DNA barcoding, with a description of mitochondrial DNA and its usability in the identification of species. The end of the theoretical part provides information about the nucleotide density vectors which the practical part is focused on. Analysis of auxiliary values on nucleotide density vectors was accomplished in Matlab by evaluating 188 real DNA barcode sequences. The identification analysis of organism was performed with the best auxiliary value for 4 datasets of the reference barcode sequences and 5 datasets of analyzed sequences of same organisms. Afterwards, the evaluation of the analysis was done by using separate nucleotide density vectors of individual nucleotides and their amounts.

KEYWORDS

Species, molecular taxonomy, DNA barcoding, nucleotide density vectors.

Kosková, M. *Vliv redukce nukleotidových denzitních vektorů na molekulární identifikaci organismů*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií. Ústav biomedicínského inženýrství, 2014. 64 s. Bakalářská práce. Vedoucí práce: Ing. Denisa Maděránková.

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma Vliv redukce nukleotidových denzitních vektorů na molekulární identifikaci organismů jsem vypracovala samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autorka uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědoma následků porušení ustanovení § 11 a následujících zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne ....

(podpis autora)

PODĚKOVÁNÍ

Děkuji své vedoucí bakalářské práce Ing. Denise Maděránkové za odborné vedení, trpělivost, ochotu a cenné rady, které mi v průběhu zpracování bakalářské práce věnovala.

V Brně dne ....

(podpis autora)

OBSAH

SEZNAM OBRÁZKŮ

ZDROJE OBRÁZKŮ

Obr. 6: http://www.gate2biotech.cz/jak-mutace-v-mitochondrialni-dna-zpusobuje-hluchotu/

Obr. 7: http://www.gate2biotech.cz/jak-mutace-v-mitochondrialni-dna-zpusobuje-hluchotu/

Obr. 8: http://www.sabarcodes.co.za/1/post/2013/04/the-barcode-of-life-dna barcoding-initiative.html

SEZNAM TABULEK

ÚVOD

Pokud chceme jedince organismů přiřazovat k jednotlivým druhům, ať už na základě molekulární nebo morfologicky orientované klasické taxonomie, je v první řadě potřeba kategorii druh definovat. Přestože to zní jednoduše, opak je pravdou. Zatím nikdo nevymyslel univerzální definici druhu, na které by se všichni shodli.

Johny Ray (1628-1705) byl pravděpodobně první autor, který se snažil stanovit vědeckou definici druhu: „*Druhy jsou skupiny rostlin, které v mezích své proměnlivosti plodí shodné potomstvo* [1].“ Mezi další formulované definice patří evoluční pojetí G. G. Simpsona (1902-1984) : „*Evoluční druh je rodina (posloupnost populací předků a potomků), jež se vyvíjela izolovaně od jiných rodin a která se vyznačuje osobitou evoluční rolí a tendencemi* [1].“ Van Valen L. M. (1976) se snažil obohatit dosavadní teorie o ekologický prvek a sestavil definici: „*Druh je rodina (nebo skupina blízce spřízněných rodin), jež ve svém areálu obsadila adaptivní zónu alespoň minimálně odlišnou od zón, které obsadily jiné rodiny, a jež se vyvíjela odděleně ode všech rodin mimo její areál* [1].“ Snahy o odhalení fylogenetické historie skupin organismů přinesly teorii J. Cracrafta (1983): „*Druh je nejmenší diagnostikovatelný shluk jednotlivých organismů, uvnitř něhož panuje rodičovská posloupnost mezi předky a potomky* [1].“ Mayr (1942) vytvořil biologický koncept druhu: „*Biologické druhy jsou skupiny aktuálně nebo potenciálně se křížících přirozených populací, jež jsou reprodukčně izolováni od jiných takových skupin* [1].“ Samozřejmě to není výčet všech vytvořených definic pojmu druh, nýbrž jen ukázka toho, jak některé z nich znějí. [1]

V minulosti probíhala identifikace organismů pouze na základě jejich morfologických znaků vědci na to vyškolenými. S rozvojem techniky se však začal využívat jiný způsob a to molekulární identifikace, která vychází ze znaků uložených v sekvenci DNA každého jedince. Pod těmito znaky si můžeme představit pořadí nukleotidů uložených v DNA. Toho, jak jsou nukleotidy za sebou zařazeny, využíváme i u vytváření nukleotidových denzitních vektorů, což je jedna z metod numerické reprezentace řetězce DNA. Jelikož je tato reprezentace pro každý druh odlišná, mohla by být použita pro druhovou identifikaci. V praktické části budeme různými hodnotami prodlužovat okrajové části sekvence, aby nedošlo ke zkreslení sekvence, a zkoumat vliv takto ošetřených okrajů. S nejlepší okrajovou hodnotou bude následně provedena identifikační analýza organismů právě na základě porovnávání nukleotidových denzit mezi referenčním a analyzovaným souborem, která bude zakončena vyhodnocením úspěšnosti tohoto způsobu identifikace organismů.

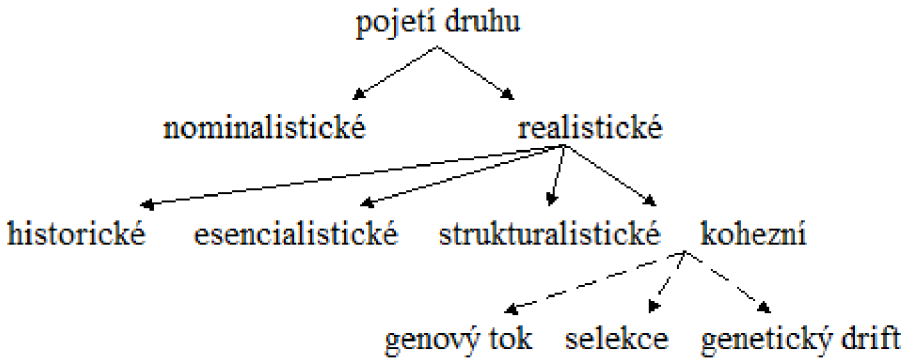
1 POJEM DRUH

Druh je přirozená vymezená skupina jedinců, kteří jsou si podobní v určitých specifických znacích a platí, že jejich podobnost uvnitř jednoho druhu je mnohem větší než podobnost jedinců pocházejících z druhů jiných. [2]

Druhy lze vymezit na základě jejich vlastností. Je potřeba brát ohled na biologické, morfologické a ekologické vlastnosti a i jejich dědičnost. Není to ovšem tak jednoduché, protože existují druhy, které jsou na základě některých vlastností nerozlišitelné a naopak také druhy, uvnitř kterých jsou jedinci značně odlišní. Kromě vlastností je dalším důležitým aspektem i prostředí, ve kterém se organismy vyskytují a vlastnosti tohoto prostředí, jež mají na vývoj jedinců také vliv. Na vznik a trvání druhu působí prostředí vyvoláním různých odchylek. Z nich jsou vybírány ty, které jsou pro jedince nejvýhodnější a organismy s těmito vlastnostmi jsou od ostatních izolováni, tudíž nemůže dojít ke křížení jedinců s těmito výhodnými vlastnostmi s jedinci, kteří mají vlastnosti jiné, méně výhodné. Nesmí být opomenuto ani prostorové a časové vymezení druhu. Každý druh je vyznačen jedinci, kteří žijí v určitém prostředí a každý má své určité trvání v čase. U různých druhů je prostor a čas odlišný. [3]

V důsledku těchto aspektů je možné formulovat následovnou obecnou evoluční definici druhu. Biologický druh je nejmenší v prostoru a čase vymezený soubor jedinců určitých vlastností, který tvoří kvalitu odlišnou od kvality příbuzných druhů, včetně druhu, z něhož v průběhu své fylogeneze vznikl. Je od ostatních druhů reproduktivně izolován a nachází se ve stavu dočasné vývojové rovnováhy s podmínkami svého životního prostředí [3]. Toto je ovšem jedna z mnoha definic, které byly vytvořeny, a neexistuje jediná univerzální definice, s kterou by všichni souhlasili. Dnešní biologové se totiž shodnou v tom, že druhy existují, ale v otázce proč existují a jak je vymezit se už bohužel liší. Vyskytují se možná či teoretická pojetí druhů a ty můžeme roztřídit na základě otázek, na které zastánci jednotlivých pojetí odpovídají odlišně. [2]

První otázkou je, zda druhy v přírodě existují objektivně (realistické pojetí druhu), nebo je vzájemně vymezuje až člověk (nominalistické pojetí druhu). Pokud se přikloníme k první možnosti, že v přírodě existují druhy nezávisle na člověku, musíme položit druhou otázku a to, zda existence různých druhů zákonitě vyplývá z vlastností živých systémů či z vlastností prostředí, ve kterém žijí, nebo zda je jejich existence pouze důsledkem historické náhody. Nakonec se musíme zeptat, jestli jsou příčiny existence různých druhů vnitřní (výsledkem vlastností vnitřních prvků živých systémů) nebo zda jsou dány zvnějšku (vlastnostmi jejich životního prostředí). Rozdělení pojetí druhu je zobrazeno na Obr. 1. [2]



Obr. 1: Rozdělení pojetí druhu (převzato z [1]).

1.1 Nominalistické pojetí druhu

Nominalistické pojetí druhu předpokládá, že druhy uměle vymezuje člověk – taxonom. Objektivně druhy v přírodě podle nominalistů totiž neexistují. Za reálně existující považují pouze jedince, jejichž přirozenou vlastností je individuální variabilita. Z variability mezi jedinci vyplývá, že někteří si budou podobni více a jiní méně, a že tedy hranice mezi druhy budou „rozmazané“, tudíž není snadné některé druhy přesně vymezit. Ostrá hranice mezi druhy je podle tohoto pojetí umělá konvenční záležitost (stejně jako u vyšších taxonů). Nominalistické pojetí existuje pouze jedno a v současné době stojí stranou. [2]

1.2 Realistické pojetí druhu

Realistické pojetí druhu předpokládá, že druhy a hranice mezi nimi nevymezuje člověk, nýbrž že existují v přírodě objektivně. Na rozdíl od nominalistického pojetí se vyskytuje více druhů toho pojetí - historické, esencialistické, strukturalistické a kohezní. Všechny vycházejí z předpokladu, že druhy existují nezávisle. Bez vymezení člověkem. Liší se ale v příčině odlišnosti jednotlivých druhů. [2]

1.2.1 Historické pojetí druhu

Historické pojetí druhu předpokládá, že existence odlišných druhů může být způsobena historickou náhodou. Toto vysvětlení může spočívat v tom, že ve vývoji každého druhu se střídají krátká evoluční období plasticity s dlouhými obdobími evoluční stáze, kdy se druh nemění. Vlastnosti druhů pak odrážejí podmínky, které nastaly v době evoluční plasticity druhu. Jiné historické vysvětlení nabízí teorii mnohorozměrnosti fenotypu

(fenotyp je soubor všech pozorovatelných vlastností a znaků živého organismu [4]) organismů. V mnohorozměrném prostoru je velmi nepravděpodobné, že se dva jedinci náhodně potkají. Vlivem mutací se rozšiřují fenotypová spektra druhů, ale jejich prolnutí u dvou druhů je téměř nemožné a druhy si tedy zachovají svou odlišnost. [2]

1.2.2 Esencialistické pojetí druhu

Esencialistické pojetí předpokládá, že jedincům stejného druhu je společná určitá vnitřní kvalita (esence), kterou se odlišují od příslušníků jiných druhů. Podle zastánců tohoto pojetí jsou počet druhů organismů i hranice mezi druhy předem jednoznačně určeny. Vnitrodruhovou variabilitu vysvětlují esencialisté jako nestejnou míru či kvalitu esence. K vysvětlení druhovosti esencialistické pojetí není příliš vhodné. Druhy jsou produkty neopakovatelného sledu událostí, na jejich formování se podílely zákonité procesy a náhoda. [2]

1.2.3 Strukturalistické pojetí druhu

Strukturalistické pojetí druhu předpokládá, že existence druhů a jejich vzájemná odlišnost vyplývají z vlastností jejich strukturních prvků. Kvůli tomu je možné esencialismus považovat za extrémní formu strukturalismu. Strukturalisté vychází z myšlenky, že fenotyp organismů je určován především deterministickými procesy vyplývajícími z vlastností jejich stavebních prvků a také mechanismy ontogeneze (průběh vývoje jedince [4]). Druhotná role je přisouzena vnějšímu prostředí. [2]

1.2.4 Kohezní pojetí druhu

Díky náhodným mutacím je v rámci druhu neustále generována variabilita. Kvůli tomu by se genotypové (genotyp je soubor všech genetických vlastností organismu [4]) i fenotypové spektrum každého druhu mělo neustále rozšiřovat a mělo by docházet ke stírání hranic mezi druhy. Opak je ale pravdou. Mechanismy druhové koheze jsou různé, ale společné jim je to, že jsou odpovědné za udržování podobnosti mezi příslušníky stejného druhu a nepřímo za rozdíly mezi jedinci patřícími do různých druhů. Tyto mechanismy působí proti rozšiřování fenotypového i genotypového spektra druhů a zabraňují jejich splývání a prolínání. Mezi mechanismy druhové koheze patří pohlavní rozmnožování a u některých organismů i působení přírodního výběru. [2]

2 VÝVOJ DRUHU (SPECIACE)

Jedním ze základních rysů života na Zemi je obrovská biodiverzita, která se projevuje existencí velkého počtu odlišných druhů (diverzita) a růzností těchto druhů (disparita) [2]. Speciace je proces vzniku nového druhu, který může trvat několik stovek tisíc nebo i milionů let. U nepohlavně se rozmnožujících druhů je to z velké míry záležitostí konvence - druh se postupně mění následkem mutací, až nakonec od určitého okamžiku začne být považován za nový druh. U druhů, které se rozmnožují pohlavně (v přírodě převažují), je pro vznik nového druhu rozhodující vytvoření reprodukčně - izolační bariéry, která představuje složitý a zajímavý evoluční problém. Díky této bariéře nemůže dojít ke křížení příslušníků starého a nového druhu. [2, 5]

Speciací existuje velké množství lišících se svým mechanismem. Můžeme je rozdělit z hlediska typu na štěpné a fyletické, z hlediska délky trvání na okamžité a postupné nebo podle kontaktu s mateřským druhem na alopatrické a sympatrické. Dále se dělí na speciace s ohledem na pohlavní a nepohlavní rozmnožování. [2]

2.1 Štěpná a fyletická speciace

Fyletická speciace znamená, že se jeden druh mění jako celek v druh jiný, tj. mění se fenotypové vlastnosti jeho příslušníků. Díky tomuto typu speciace přibývá celkový počet druhů. Biodiverzita se ovšem nemění, protože dojde pouze k přeměně jednoho druhu v jiný. [2]

Štěpná speciace se od fyletické liší tím, že v tomto případě se jeden druh mateřský rozpadne na dva nebo více druhů dceřiných, které se dále anageneticky (co do svých fenotypových vlastností) vyvíjejí samostatně, nezávisle na sobě. [2]

2.2 Okamžitá a postupná speciace

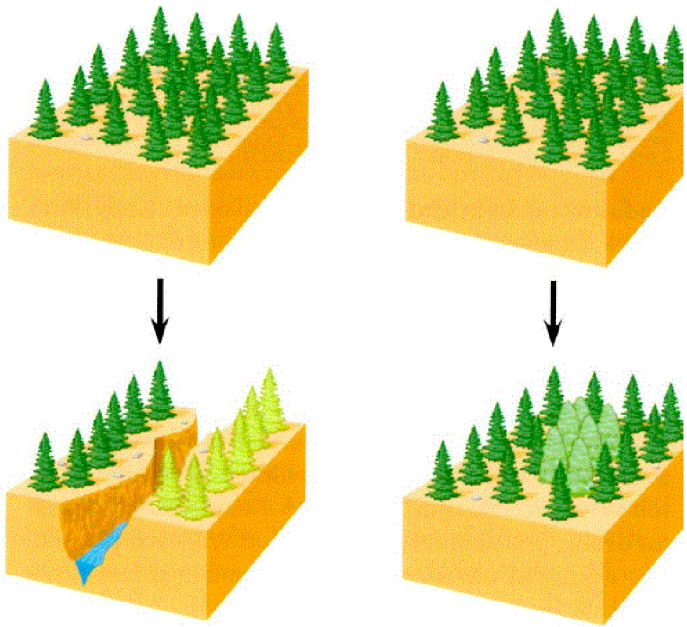
Vzhledem k tomu, že většina speciací trvá dlouho, jsou řazeny mezi speciace postupné. Mezi dvěma populacemi se pomalu hromadí genetické rozdíly, které vedou k fenotypovému odlišení jedinců obou populací a zároveň u pohlavně se rozmnožujících druhů dochází ke vzniku reprodukčně izolační bariéry. U každého druhu trvá vytvoření reprodukčně izolační bariéry různě dlouhou dobu a závisí na různých aspektech (nárůst genetických rozdílu pouze genetickým driftem - trvá dlouho, na speciaci se podílí i přirozený výběr - rychlejší). Příkladem postupné speciace je speciace alopatrická (viz

kapitola 2.3). [2]

Jsou známy i speciace, které proběhnou téměř okamžitě. Mezi tyto speciace se řadí speciace polyploidizační, kdy se kvůli poruše buněčného dělení vytvoří u diploidního druhu tetraploidní jedinec (nejčastěji u rostlin). Starý diploidní a nový tetraploidní druh mohou koexistovat na stejném území kvůli rozdílným ekologickým nárokům. [2]

2.3 Alopatrická a sympatrická speciace

Alopatrická speciace je nejjednodušší cesta, jakou může nový druh vzniknout. Jedná se o postupný vývoj druhu bez kontaktu s mateřským druhem. Jestliže se část populace zeměpisně izoluje od původní populace a zůstane po dostatečně dlouhou dobu reprodukčně izolovaná, nahromadí se u ní genetické změny, které povedou k fenotypovému a následně ekologickému rozrůznění populací. Vnější reprodukční bariéry způsobené geografickou izolací budou časem doplněny i vnitřními bariérami, čímž je zabráněno vzájemnému křížení, a to i v případě, že jsou populace uměle přeneseny na stejné území. Pokud se tyto dvě oddělené populace dostanou do kontaktu dříve, než se dostatečně diferencují, mohou opět splynout. V opačném případě budou žít sympatricky vedle sebe nebo jeden druh může vytlačit druhý. Pro lepší pochopení alopatrického vývoje druhu viz Obr. 2 – levá část. [2, 6]



Obr. 2: Alopatrická speciace a sympatrická speciace - geografické vztahy nových druhů vzhledem k jejich rodičovským druhům (převzato z [6]).

Při sympatrické speciaci se nový druh formuje na stejném území jako druh mateřský čili bez geografické izolace. Oproti alopatrické speciaci mezi vzájemně diferencujícími druhy tedy nevznikají vnější reprodukční bariéry, ale pro izolaci genofondu (soubor všech genů dané populace) je vyžadován vznik vnitřních reprodukčních bariér. U živočichů může být sympatrická speciace výsledkem určité části populace, která se stává reprodukčně izolovanou z důvodu přechodu do prostředí, k potravnímu zdroji nebo jinému zdroji, nevyužívanému rodičovskou populací nebo může být výsledkem vzniku preference partnera (například kvůli zbarvení) [6]. Sympatrický vývoj druhu je znázorněn v pravé části Obr. 2. [2, 6]

2.4 Speciace s ohledem na pohlavní rozmnožování

Speciace u druhů s pohlavním rozmnožováním je složitější než u druhů s rozmnožováním nepohlavním. Při nepohlavním rozmnožování stačí, aby u některého jedince došlo k mutaci. Důsledkem této mutace se změní fenotypové vlastnosti příslušné linie jedinců a může vzniknout nový druh, pokud ovšem v přírodě existuje volná ekologická nika (to je prostor, který v ekosystému zaujímá určitý druh [4]), kterou může nový druh zaplnit. [2]

Jedná - li se o pohlavní rozmnožování, jsou mechanismy speciace složitější. Pouze fenotypové rozdíly mezi mateřským a dceřiným druhem nestačí. K odštěpení nového druhu je zapotřebí vznik reprodukčně - izolační bariéry mezi novým a původním druhem. Tato bariéra zabrání toku genů mezi druhy a stírání fenotypových rozdílů mezi jejich zástupci. [2]

Vnější reprodukčně - izolační bariéry existují v prostředí nezávisle na biologických vlastnostech organismů. Vyskytují se díky rozmanitosti okolního prostředí, které má ostře vymezené hranice a tvoří tak významné bariéry, které zamezují pohybu jedince z jedné strany bariéry na druhou nebo tento pohyb zpomalují. To omezuje ekologické nebo genetické interakce jedinců, což může vest k fenotypové i genetické diferenciaci členů a k následné speciaci. Příkladem geografická, behaviorální, časová, gametická nebo mechanická izolace druhů. Opakem jsou vnitřní reprodukčně - izolační bariéry, které existují, aby nedošlo ke křížení jedinců různých druhů žijících na stejném území, což by postupně mazalo hranice mezi druhy. Tyto bariéry jsou přímo nebo nepřímo určeny genotypem organismů a vznikají či zanikají důsledkem genetických procesů (nejčastěji v důsledku mutací). [2, 6]

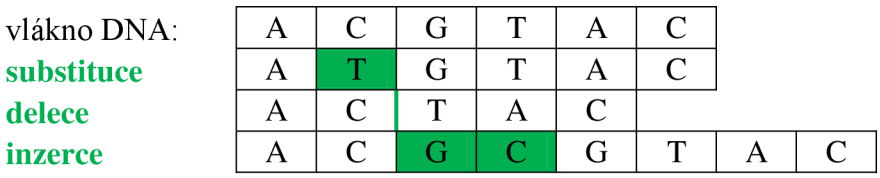
3 MUTACE

Biologové 20. století zjistili, kde se bere variabilita, ze které jsou přirozeným výběrem vybráni životaschopnější jedinci, a proč tato variabilita v důsledku křížení (u pohlavně se rozmnožujících druhů) nevymizí. Jediným zdrojem evolučních novinek a variability na úrovni druhu jsou mutace, což jsou změny ve struktuře genetického materiálu (u většiny organismů v DNA), při nichž se mění smysl genetické informace, aniž by byla porušena pravidla zápisu genetické informace. Pokud by změny tato pravidla porušovaly, jednalo by se o poškození DNA. V buňce existují enzymatické systémy, které mohou poškozená místa rozpoznat a opravit. Někdy ovšem oprava možná není a reparační proces je zdrojem vzniku mutací. [5, 7]

Mutace vznikají také důsledkem poruch při replikaci (kopírování) nebo při rozdělení chromozomů do dceřiných buněk v průběhu buněčného dělení. Mutace jsou nezbytné pro biologickou evoluci, protože bez vzniku nových mutací by se evoluce zastavila a organismy by se nemohly dále vyvíjet. Jsou označovány jako složitý a vysoce adaptivní proces. Existuje více typů mutací s různou fyzickou povahou, mechanismem vzniku nebo významem pro biologickou zdatnost [7]. Známe mutace pozitivní, negativní, selekčně neutrální, spontánní a indukované. Zde si ale popíšeme pouze mutace fyzické povahy, které rozlišujeme na bodové, na úrovni celých úseků DNA, chromosomové a genomové mutace. [7]

3.1 Bodové mutace

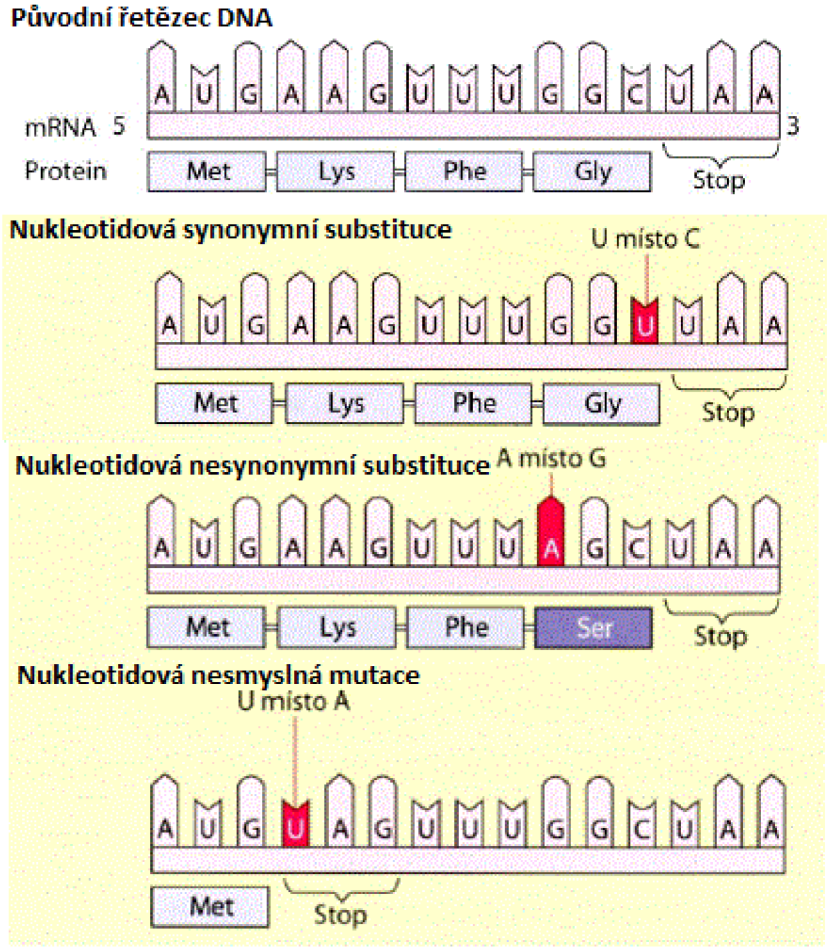
Bodové mutace jsou chemické změny v jednom páru bází genu a probíhají na úrovni vlákna DNA. Pokud k této mutaci dojde v gametě (pohlavní buňka) nebo buňce, která dává vznik gametám, může být přenesená na potomstvo. [6]



Obr. 3: Zobrazení bodových mutací (substituce, inzerce, delece).

Nejčastější formou bodové mutace je substituce (viz Obr. 3), kdy se jedná o záměnu jednoho nukleotidu za druhý. Pokud jde o výměnu nukleotidu s purinovou bází

(adenin - A , guanin - G) za nukleotid s pyrimidinovou bází (cytozin - C, thymin - T) nebo naopak, jedná se o transverzi. Jestliže se mění nukleotid s nukleotidem obsahujícím stejný typ báze, jde o tranzici. Pokud dojde k substituci na úseku DNA, která kóduje protein, je možné tyto mutace rozdělit podle jejich vlivu na strukturu proteinu. [6, 8, 9]



Obr. 4: Typy bodových mutací (převzato z [6]).

Genetický kód je degenerovaný, což znamená, že stejná aminokyselina je kódována celou řadou různých kodonů (nukleotidových tripletů), a proto se nemusí záměna nukleotidu v kodonu na struktuře proteinu nijak projevit. Popsaný typ mutace je mutace synonymní (viz Obr. 4), neměnící smysl. A jde tedy o změnu kodonu (triplet bází) pro určitou aminokyselinu v jiný kodon stejného smyslu. Naopak mutace se změnou smyslu, nesynonymní (viz Obr. 4), jsou takové mutace, kvůli kterým je jedna aminokyselina nahrazena jinou. Pokud se jedná o aminokyselinu s podobnými fyzikálně – chemickými vlastnostmi, hovoříme o záměně konzervativní, kterých je nejvíce. Jejich důsledkem se terciální struktura proteinu a jeho biologická funkce výrazně nemění.

Posledním typem záměnových mutací jsou nesmyslné mutace (viz Obr. 4). Při nich vzniká z kodonů pro aminokyseliny některý ze tří kodonů terminačních (UAA, UAG, UGA), tudíž v tomto místě dojde k předčasnému ukončení syntézy proteinového řetězce v průběhu translace. Vzniká tak nefunkční protein. [7, 8, 9]

Dále mezi bodové mutace patří delece (viz Obr. 3), při které dochází ke ztrátě jednoho nebo více nukleotidů z nukleotidové sekvence. V důsledku toho dochází ke zkracování řetězce. Patří sem i inzerce (viz Obr. 3), při které naopak vložení jednoho nebo více nukleotidů do nukleotidové sekvence řetězec prodlouží. Oba typy mutací tedy mění počty nukleotidů v sekvenci, a tudíž i její délku. Důsledkem toho dochází k posunu čtecího rámce. Proto v podstatě nezměněná sekvence je překládána jako sekvence jiná (čtecí rámec bere jiné trojce nukleotidů). Kvůli tomu mají tyto mutace mnohem častěji katastrofální dopad na výsledný protein než substituce. [6, 7, 8, 9]

Frekvence jednotlivých typů mutací se liší v závislosti na typu organismu (bakterie, eukaryota) a na genomu, ve kterém se mutace objevuje (jádro, mitochondrie, plastid), ale také na nukleotidech, které se vyskytují poblíž dané pozice [7].

3.2 Ostatní typy mutací

Mutací na úrovni celých úseků DNA je opět několik typů. Delece, inzerce a duplikace způsobují změnu délky vlákna DNA a to tak, že je určitý úsek vlákna ztracen, vložen nebo zmnožen. Translokace jsou charakterizovány přemístěním určitého úseku DNA na jiné místo v genomu a při inverzi je určitý úsek DNA z chromosomu vyjmut a vložen na stejné místo v opačném pořadí. Translokace a inverze můžou mít velký význam při speciaci. Mohou sloužit jako jeden z mechanismů vytváření mezidruhové bariéry, protože pokud se kříží jedinci, kteří mají větší počty přestaveb DNA, mají sníženou fertilitu (plodnost). [7]

Chromozomové a genomové mutace nemění samotné geny, nýbrž mění strukturu nebo počet chromozomů. Chromozomové mutace jsou strukturní změny chromozomů a jejich následky závisí na tom, zda je i po strukturní přestavbě zachované normální množství genetické informace [9]. Pokud ne, tak dochází k fenotypovým projevům, které se odvíjí od toho, která část genetické informace chybí, přebývá nebo je poškozená [9].

Důsledkem genomových mutací je změna samotného genomu. Vznikají v důsledku poruch v průběhu buněčného dělení. Vlivem těchto mutací mohou vznikat organismy, u kterých je zvýšen nebo snížen počet určitých chromozomů (aneuploidie) nebo celé chromosomové sady (polyploidie). [4, 7]

4 TAXONOMIE

Taxonomie je věda o taxonech, o jejich poznávání, vymezování, třídění, pojmenovávání a klasifikaci [10]. Je to složenina dvou řeckých slov *tax* (srovnat, uspořádat) a *nomia* (zvyk, zákon) [11]. Počátky taxonomie sahají do 18. století k Linneově (Carl von Linné byl švédský vědec a botanik) knize *Systema naturae*, což přeloženo znamená „Systém přírody" a představuje třídění všeho živého v přírodě [6].

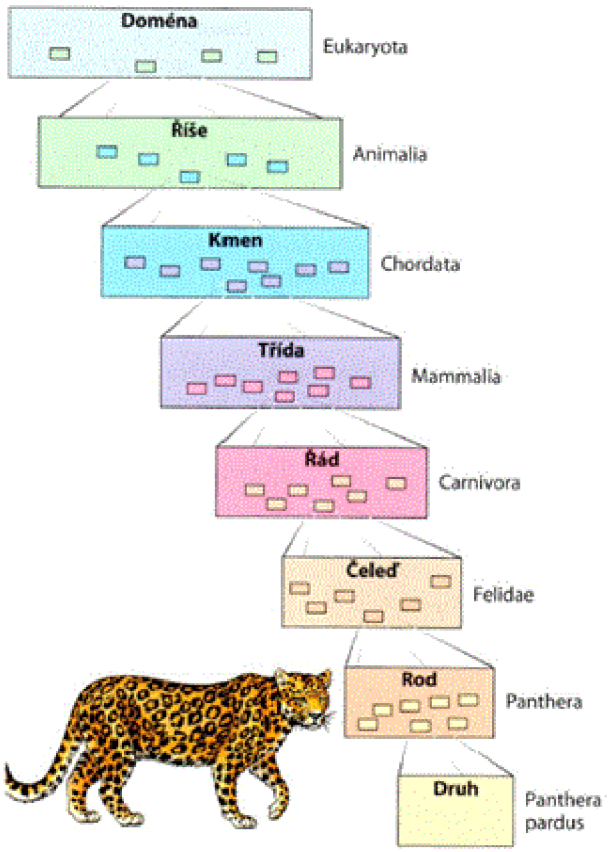
Taxon je jakýkoli přirozený a rozlišitelný soubor žijících i vymřelých příbuzných skupin organismů, který je tak vyhraněný, že ho přijímáme jako jednotku klasifikace [10]. Jedinci, kterými je tvořen, jsou si příbuzní a svými znaky se odlišují od ostatních taxonů. Jednotlivé taxony jsou hierarchicky zařazeny do klasifikačních kategorií různé úrovně. Příbuzné rody jsou řazeny do jedné čeledi, čeledi do řádů, řády do tříd, třídy do kmenů, kmeny do říší a říše do domén. Každá taxonomická úroveň je tedy komplexnější než úroveň předchozí. Například všechny druhy koček jsou savci, ale všechny savce nepředstavují pouze kočky. [6, 10]

Každý druh má dvouslovné latinské jméno čili binomen [6]. První část binomického označení je rod, do kterého druh patří. Druhou částí je jméno druhové a odpovídá jednomu druhu příslušného rodu. Tento princip pojmenování druhu se nazývá binominální nomenklatura. Dvouslovné jméno pro každý druh a hierarchické zařazení druhů do systému organismů jsou dvě zásadní vlastnosti Linneova systému. Pro ujasnění je na Obr. 5 zařazení *Panthera pardus* (leopard) do taxonomického systému. [6, 10]

Znak je jakákoli vlastnost organismu, kterou se liší od jiných organismů nebo taxonů [11]. Každý znak může mít jeden nebo více stavů (hodnot, projevů) [11]. Taxonomickým znakem můžeme nazvat jakoukoli vlastnost organismu, která má nejméně dva stavy a je použitelná pro klasifikaci [11]. Stavy znaků mohou být odlišné kvantitativně (např. přítomnost/nepřítomnost nebo tvar), meristicky (počet prvků) nebo spojitě kvantitativně (např. délka těla) [10]. Zdroji znaků mohou být údaje strukturální, ontogenetické a morfometrické (rozměry a jejich poměry), cytotaxonomické (stavba, počet a chování chromozomů), fyziologické, biochemické, genetické a další [10]. Pro identifikaci organismu za pomoci DNA barcodingu (viz kapitola 5) budou ale nejdůležitější znaky molekulárně - biologické. To jsou údaje o molekulární struktuře genomu a jeho složení [10].

Strukturální znaky jsou nejčastěji používané. Hlavním důvodem je to, že se jedná o snadno dostupné informace oproti jiným znakům. Zvláštní skupinou znaků jsou znaky makromolekulární. Jsou poskytované studiem nukleových kyselin (DNA, RNA)

a jejich bezprostředních produktů, což jsou proteiny [10]. Tyto znaky mohou být získány metodami přímými (např. mapováním genů, sekvencováním nukleotidů v DNA a RNA, aminokyselin v proteinech, studiem elektroforetických reakcí polymorfních bílkovin) nebo nepřímými (např. zjišťováním imunologické vzdálenosti různých organismů, stupně mezidruhové hybridizace DNA) [10].



Obr. 5: Zařazení leoparda do taxonomického systému (převzato z [6]).

4.1 Klasická taxonomie

Klasická taxonomie je odvětví taxonomie, které se zabývá jinými než molekulárními znaky organismů. Organismy zařazuje do jednotlivých kategorií pomocí jejich morfologických a anatomických znaků, zkoumáním fyziologie organismů (tj. růst, rozmnožování, metabolismus, reakce na zevní podněty), dále také sledováním jejich chování nebo jejich ekologie. Poslední jmenované se zabývá vnitrodruhovými i mezidruhovými vztahy mezi organismy a jejich vztahem k vnějšímu prostředí.

Zjednodušeně lze říci, že sledováním toho, jak jedinec vypadá, jak se chová a jaké je jeho přirozené prostředí, je možné organismy sdružit do skupin, taxonů. Naší

problematiky se ale mnohem více týká taxonomie molekulární.

4.2 Molekulární taxonomie

Při kategorizaci organismů je užitečné srovnávat nejen jejich anatomické vlastnosti, ale i jejich makromolekuly. Molekulární taxonomie, jak už název napovídá, využívá molekulární znaky a pomocí nich zařazuje organismy do hierarchického systému. Pod těmito znaky si můžeme představit znaky uložené právě v sekvencích makromolekul, mezi které patří DNA, RNA a proteiny [12]. Molekulárního taxonoma nejvíce zajímá primární struktura DNA, jelikož na této úrovni vznikají evoluční novinky ve formě mutací a informace zde uložená se dědí na potomstvo podle jasných pravidel. Primární sekvence DNA v sobě také nese informaci o historii daného lokusu DNA (pozice, kterou na chromozomu zaujímá jeden nebo více genů) nebo proteinu [12]. Z daného lokusu můžeme vyčíst identitu jedince a jeho příbuznost s ostatními jedinci v populaci, druhovou příslušnost nebo dokonce příbuzenský vztah tohoto druhu nebo příslušné taxonomické skupiny s jinými [12]. [6, 12]

Molekulární znaky, jak bylo zmíněno výše, jsou znaky uložené v molekulách DNA, RNA nebo proteinů. Například každá pozice nukleotidu na vlákně DNA představuje znak ve formě jednoho ze čtyř typů DNA bází: A (adenin), G (guanin), C (cytozin) nebo T (thymin). Tyto znaky mají svoje výhody i nevýhody. [12]

Mezi výhody patří to, že tyto znaky pocházejí z úrovně, kde vznikají evoluční novinky neboli mutace, obvykle víme, jak se dědí, mnoho z nich nezávisí na prostředí a jsou velmi často selekčně neutrální. Poslední jmenované znamená, že nejsou ovlivňovány přírodním výběrem. Těchto znaků je obrovské množství a jsou použitelné na všech úrovních taxonomie - od porovnávání jedinců v rámci populace až po rekonstrukci hluboké fylogeneze. K dalším výhodám můžeme zařadit to, že se znaky dají jednoznačně popsat, protože nabývají několika diskrétních stavů (čtyři nukleotidy, dvacet aminokyselin), jsou lépe vážitelné (neuděláme chybu, když jim přisoudíme stejnou váhu) a lépe se kvantifikuje stupeň nejistoty. Vybrané znaky je možné s opatrností použít jako molekulární hodiny pro odhad stáří. [12]

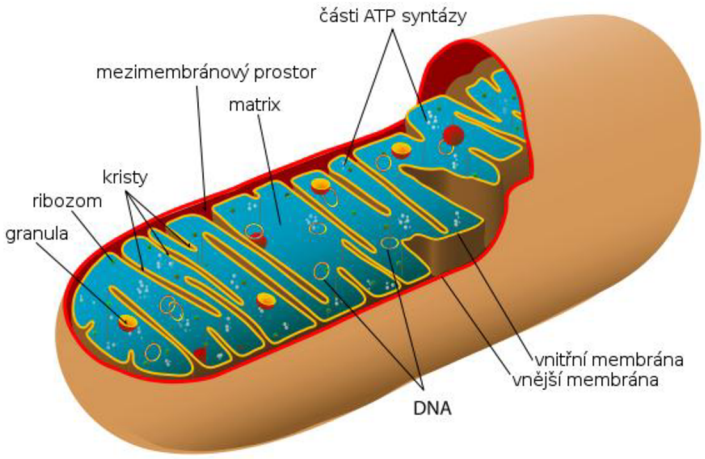
Nevýhodou je, že se většina molekulárních znaků vůbec neprojeví ve fenotypu, jejich získávání je finančně náročnější než získávání znaků morfologických a také že při jejich získávání je někdy nutné organismus nebo jeho část nenávratně zničit. [12]

V molekulární taxonomii se pro identifikaci druhů používají dva základní principy a to distanční metody a znakově orientované metody. Obě budou popsány v textu dále ve spojení s DNA barcodingem.

5 DNA BARCODING

5.1 Mitochondriální DNA

Mitochondrie je organela (viz Obr. 6), která se nachází téměř ve všech eukaryotických buňkách, včetně buněk rostlin, živočichů, hub a prvoků. Každá buňka obsahuje různé množství mitochondrií a to v závislosti na buněčné úrovni metabolické aktivity. Některé buňky mají jednu velkou mitochondrii, častěji se však vyskytují stovky, či dokonce tisíce mitochondrií. Tato organela je obklopena dvěma membránami. Vnější je hladká, vnitřní je zvrásněná a obsahuje záhyby. Ty se nazývají kristy a zvětšují povrch membrány asi pětkrát. Uvnitř mitochondrie se nachází mitochondriální matrix. Hlavní funkcí této organely je tvorba ATP (adenosintrifosfát), což je molekula, která poskytuje chemickou energii pro všechny procesy v buňce. [6]



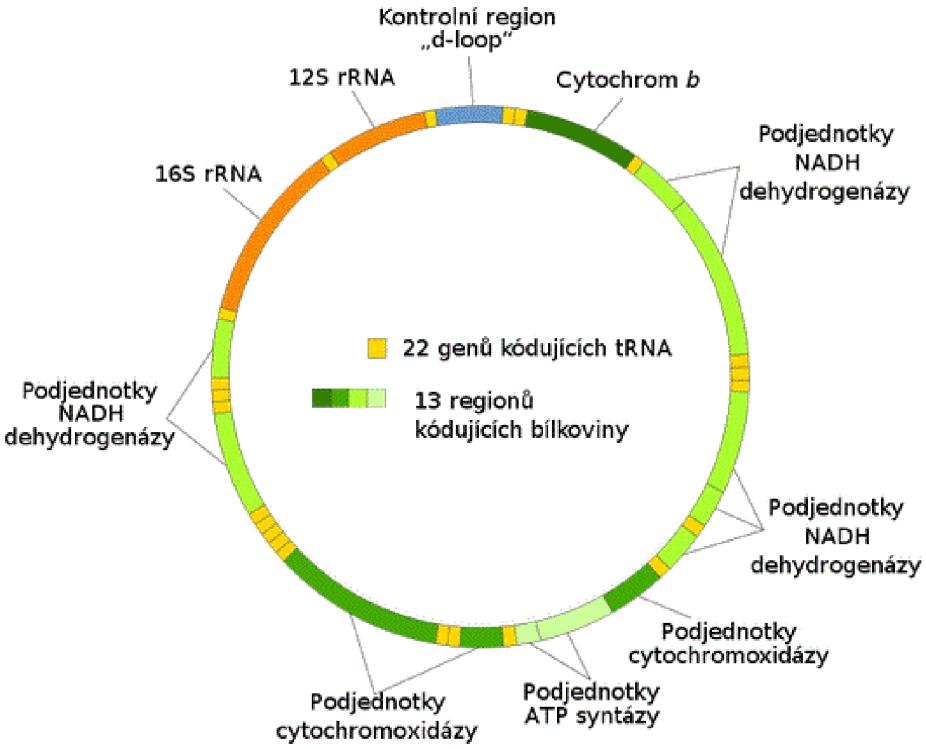
Obr. 6: Mitochondrie (převzato viz Zdroje obrázků).

Dělení mitochondrií probíhá obdobným způsobem jako u bakterií – dojde k růstu a následnému příčnému dělení. To dokazuje, že jsou bakteriálního původu. Navíc je známo, že mitochondrie vznikly před více než miliardou let a to tak, že aerobní bakterie byla pohlcena anaerobní eukaryotní buňkou. Od té doby se většina původních genů přenesla do buněčného jádra a v samotné mitochondrii zůstalo poměrně málo genů. [13]

Mitochondrie obsahují vlastní haploidní genom, kterým je mitochondriální DNA (mtDNA). Tato DNA je dvouvláknová kruhová sekvence DNA, ale bylo objeveno i množství lineárních sekvencí DNA, a nachází se v mitochondrii ve více kopiích. Je relativně malá a snadno izolovatelná. Komplementární vlákna DNA mohou být rozlišena na základě skladby bází, které způsobují různou hustotu řetězců. Jedno vlákno

je označováno jako těžké, jelikož obsahuje větší množství purinových bází (A, G) a druhé jako lehké kvůli převaze pyrimidinových bází (C, T). [14]

Zjištění, že existuje i DNA, která leží v jiném prostoru, než je eukaryotické jádro, byl skutečně velký objev. To otevřelo zcela nové obzory ve studiu biogeneze (vznik a vývoj živých organismů nebo jejich částí z jiných) mitochondrií. Během posledních tří desetiletí bylo geneticky i sekvenčně charakterizováno několik stovek mitochondriálních genomů vícebuněčných živočichů, rostlin a hub. Sekvence nukleotidů lidské mtDNA byla první doložená kompletní sekvence mitochondriálního genomu a byla publikována v roce 1981. Následně bylo zjištěno, že ostatní mtDNA metazoí (vícebuněční živočichové) obsahují stejné geny s velmi malým množstvím výjimek. Pořadí genů ovšem vždy stejné není. [14]



Obr. 7: Lidská mitochondriální DNA (převzato viz Zdroje obrázků).

Velikost mtDNA se značně liší mezi živočichy, rostlinami a houbami. Mitochondriální DNA u většiny zvířat je dlouhá 16 až 20 tisíc bází, zatímco rostlinná je 10× až 100× delší a variabilnější. To však není úměrné počtu genů, které mtDNA kóduje. [14]

Jak bylo řečeno, genetický obsah mtDNA je u většiny organismů velice podobný. U metazoí kóduje mtDNA 13 mRNA (mediánových RNA), které podléhají translaci, 22 tRNA (transferových RNA) a 2 rRNA (ribozomální RNA). Odhaduje se, že

velikost mitochondriálního proteomu (soubor všech proteinů) je 800 až 1500 proteinů a liší se v závislosti na typu tkáně. Z nich pouze 13 je kódováno mitochondriálním genomem a každý je nezbytný pro oxidativní fosforylaci. Tyto mitochondriálně kódované proteiny jsou součástí 4 komplexů. *ND1-6* a *ND4L* jsou z komplexu I, *cyt b* z komplexu III, 3 podjednotky *cytochrom oxidázy* z komplexu IV a 2 *ATP syntázy* z komplexu V. Rozložení genů lidské mtDNA můžeme vidět na Obr. 7. Mitochondriální DNA obsahuje kromě kódujících úseků i úseky nekódující, nazývající se *D-loop*. [14]

Mitochondriální DNA je velmi oblíbeným molekulárním znakem pro fylogenetiku, populační studie, fytogeografii a molekulární ekologii [15].

5.2 Princip DNA barcodingu

DNA barcoding je metoda, která se pomocí krátké sekvence DNA snaží identifikovat druhy organismů. *Barcode* po přeložení do češtiny znamená čárový kód, který se používá k identifikaci zboží v supermarketech. Podobně může být nahlíženo na sekvence DNA jako na genetické čárové kódy, které jsou uložené v každé buňce. Myšlenka je taková, že stejně jako kombinace tenkých a tlustších čar na kódu zboží by mohla fungovat kombinace písmen nukleotidů v řetězci DNA. Ty potom jednoznačně vedou k jednomu rostlinnému nebo živočišnému druhu. Vizualizace sekvence písmen pomocí barevných proužků připomíná právě čárový kód, jak je vidět na Obr. 8. Tato grafická reprezentace má význam pouze jako dodatečná vizuální informace pro člověka. [16]

I když se o DNA barcodingu vědělo už dříve, poprvé se o něm píše v článku *Biological identification through DNA barcodes*, který v roce 2003 publikoval kanadský profesor Paul Hebert. Ten šel dále než předešlé studie. Snažil se najít jednoduchý gen, který by mohl sloužit k identifikaci druhů napříč celým zvířecím královstvím. Navíc slíbil, že výsledky DNA barcodingu budou lepší než ty, kterých může být dosaženo pomocí morfologických studií, jelikož je možné za pomoci DNA barcodingu identifikovat i kryptické taxony nebo různá životní stádia a pohlaví druhů, což je klasickými morfologickými metodami obtížné nebo nemožné. Tento převrat taxonomie získal své zastánce i odpůrce. [16, 17, 18]

Praktická realizace tohoto přístupu je uskutečněna za pomoci mitochondriální DNA pro živočichy a chloroplastové DNA pro rostliny. Pro tyto účely má mtDNA živočichů výhody zejména v tom, že je snadno extrahovatelná z buňky, kódující úseky (exony) nejsou odděleny úseky nekódujícími (introny), jelikož chybí. Nedochází k rekombinaci a inzerce a delece jsou vzácné (nedochází k posunu čtecího rámce). [16]

Neexistuje přesvědčivý důvod, aby se analýza soustředila na specifický mitochondriální gen, ale pro živočichy byla vybrána část genu *cox1* (cytochrom c oxidáza podjednotka I), která je dlouhá 648 párů bází a její umístění v mitochondriálním genomu je možné pozorovat na Obr. 7. Tento úsek vybral Paul Hebert, jelikož se mu ukázal jako velmi účinný při určování ptáků, motýlů, ryb a dalších zvířecích skupin. Pro použití *cox1* hovoří zejména počet kopií v buňce [17]. Jaderný genom je v buňce většinou uložen ve dvou kopiích, mitochondriální ve sto až deseti tisících kopií. Tím je mitochondriální genom jasně upřednostňován, protože i u velmi degenerovaných vzorků DNA je naděje na získání sekvenovatelné kopie daného úseku. Další výhodou použití *cox1* je jeho velká mezidruhová rozmanitost. I mezi velmi blízce příbuznými druhy se sekvence liší několikanásobně více než geny jaderné [19]. Nemůžeme opomenout ani poslední výhodu a to tu, že *cox1* je dostatečně krátký, aby mohl být rychle sekvenován a dost dlouhý na to, aby mohl sloužit k identifikaci mezi druhy [20].

Obr. 8: Pořadí nukleotidů v sekvenci mtDNA *Orcinus orca* (kosatka dravá) a tomu odpovídající čárový kód, kde jsou jednotlivé nukleotidy zobrazeny barevně, A – zeleně, C – modře, G – černě, T – červeně (převzato viz Zdroje obrázků).

Pravdou ale je, že tento gen není univerzální pro všechny, jak si Paul Hebert myslel. Například pro obojživelníky *cox1* opravdu nejlepší volba není kvůli tomu, že velká vnitrodruhová variabilita této skupiny organismů znemožňuje jejich jednoznačnou identifikaci [17]. Taktéž pro rostliny tento mitochondriální gen není vhodný, protože se vyvíjely příliš pomalu, ale dvě genové oblasti v chloroplastu, kterými jsou *matK* a *rbcL*, byly schváleny jako barcodové oblasti pro suchozemské rostliny [20].

Každopádně pro gen použitý pro DNA barcoding je důležité, aby jeho vnitrodruhová variabilita byla nízká. Malá vnitrodruhová a velká mezidruhová rozmanitost ukazují na zřetelné genetické hranice mezi druhy a díky tomu umožňují přesnou identifikaci jedince [19].

5.3 Identifikace druhů

Jednou z hlavních otázek týkající se začlenění molekulárních informací do taxonomických aspektů biologie je, k čemu využít čárové kódy. Existují dva samostatné úkoly, ke kterým se barcody používají. Prvním z nich je použití údajů z DNA k rozlišování mezi druhy (identifikace druhů) a druhý je použití sekvence DNA pro objevování nových druhů. [18]

Hned úvodem je důležité říci, že DNA barcoding se používá pouze k identifikaci druhu v porovnání s již známým druhem. Objevování nových druhů není vhodné, jelikož na základě jednoho genu nemůžeme určit příslušnost organismu k novému druhu. Je to dáno tím, že jeden gen, přibližně 650 párů bází, je příliš krátký úsek a nemá proto dostatečnou vypovídací hodnotu. Je možné ale odhalit druh, o kterém si můžeme myslet, že je nový, a následně pomocí hlubší analýzy myšlenku ověřit.

K identifikaci druhů se v molekulární taxonomii používají distanční a znakové metody. Profesor Hebert navrhl použití genetické vzdálenosti jako standardní metodu pro analýzu barcodových dat a většina barcodových studií ho následovala. U distančních metod se jedná o výpočet p-distancí s korekcí některým z evolučních modelů a následné vytvoření fylogenetického stromu. K vytvoření fylogenetického stromu se nejčastěji používá metoda *Neighbour-joining*, ale je možné využít i metodu *UPGMA*. Mechanismus identifikace organismů distanční metodou je následovný. Nejprve pomocí morfologických znaků určíme skupiny organismů, které osekvenujeme. Následně vzniklé barcodové sekvence zpracujeme distanční metodou. Vzhledem k tomu, že víme, o který druh se jedná, poznáme, zda tato metoda určila organismus správně či nikoli. Pakliže metoda fungovala, získané sekvence se uveřejní v databázi barcodových sekvencí. [16, 18]

U této metody je také nutné určit práh distanční hodnoty. Práh určuje, zda se ještě jedná o stejný druh nebo už o druh jiný. Hodnota prahu musí odrážet vnitrodruhovou variabilitu, která je ale pro různé skupiny organismů odlišná, a tudíž nastává komplikace - nelze mít nastavenu pouze jednu hodnotu prahu pro všechny jedince. [17]

Tento problém řeší použití znakových metod, které jsou podle některých lepší, jak z teoretických tak praktických důvodů, než běžně používané metody distanční.

Místo analýzy založené na vzdálenostech, kde jsou porovnávány sekvence jako celé jednotky a stromy jsou konstruovány na základě celkové podobnosti, znakové metody hledají diagnostické diskrétní znaky nebo kombinace těchto znaků v sekvenci DNA (pod těmi jsou samozřejmě myšleny nukleotidy). Pozice těchto znaků jsou v řetězci DNA jedinečné a to umožní snadnou identifikaci druhu. Je možné i zvýšit počet těchto diagnostických znaků a tím nastavit úroveň bezpečnosti pro definování taxonu. Nevýhodou této metody je, že pokud je počet charakteristických znaků nedostačující, tak diagnóza selže. Nebude tedy poskytovat „nejbližšího souseda“, kterým ovšem nemusí být vždy nejbližší příbuzný, jak tomu je u distanční metody. Mezi znakové metody zařazujeme například metodu *Maximum Parsimony* (maximální úspornost) nebo *Maximum Likelihood* (maximální pravděpodobnost). [18, 17, 21]

Charakteristické znaky jsou stavy znaků, které jsou přítomny pouze v jedné vývojově společné skupině. Jsou rozděleny do čtyř hlavních skupin. *Pure attributes* jsou sdíleny všemi členy jedné vývojově společné skupiny a nikdy se nenachází u jiných skupin. *Private attributes* jsou přítomny pouze u některých členů této skupiny a u ostatních skupin chybí. Oba typy těchto charakteristických znaků mohou být buď *simple*, což znamená, že jsou omezeny pouze na pozice jednoho znaku nebo mohou být *compound*, které hledají pozice více znaků. [18]

Pro realizaci analýzy DNA barcode sekvencí znakovou metodou byl vyvinut *Characteristic Attributes Organization System* (CAOS) software. Samotný algoritmus systému charakteristických znaků (CAOS) je založen na tom, že členové dané taxonomické skupiny sdílí znaky, které se nevyskytují v jiných skupinách. Algoritmus CAOS tak identifikuje znakovou metodou každou vývojově společnou skupinu v každém uzlu fylogenetického stromu, který je nejprve sestaven z daného datového souboru. Výsledná diagnostika pak může být použita pro následnou klasifikaci nových dat do taxonomického uskupení zastoupeného tímto stromem. Strom je používán pouze jako prostředek k identifikaci diagnostických znaků – nemusí nutně představovat domnělé fylogenetické vztahy. [18]

Při zařazování organismů do druhu ale narazíme na problém. Pojem druh nemá jednotnou definici, jak bylo zmíněno výše (kapitola 1). Pro molekulární přístupy byla vytvořena *Molecular Operational Taxonomic Unit* (MOTU), která má podobný rozsah jako tradiční druhy. DNA barcoding nám tedy nepřiřazuje organismy ke druhům, ať už jakkoli definovaným, ale k MOTU. MOTU byla navržena kvůli tomu, že identifikace druhů na základě genetických odlišností je nemožná, protože genetická variabilita mezi jednotlivými organismy přímo nesouvisí s taxony. Je to tím, že druh se neustále vyvíjí a vzniklé mutace mají za následek hromadění rozdílů. [22]

5.4 Komponenty DNA barcodingu

Prvním krokem DNA barcodingu je získání vzorku druhu, který chceme identifikovat. Vzorky můžeme získat z přírodních historických muzeí, herbářů, zoologických zahrad, zmrazených tkáňových sbírek, semenných bank a z dalších skladišť biologického materiálu nebo jiných libovolných zdrojů. Získaný vzorek tkáně je následně dopraven do laboratoře, kde z něj laboranti získají jeho DNA, ze které izolují *barcode*. Nejlépe vybaveným laboratořím to netrvá déle než pár hodin. Tato data jsou poté umístěna v databázi pro pozdější analýzu. [23]

Databáze jsou velmi podstatným prvkem DNA barcodingu. Jedním z nejdůležitějších komponent barcodové iniciativy je vytvoření veřejné referenční knihovny identifikátorů druhů, které by mohly být použity pro přiřazení neznámých vzorků tkání ke známým druhům. V současnosti se této role ujímají dvě hlavní barcodové databáze. První z nich je The International Nucleotide Sequence Database Collaborative, která je partnerskou databází nejpoužívanější databáze primárních sekvencí GenBank v USA, The Nucleotide Sequence Database of the European Molecular Biology Lab v Evropě a DNA Data Bank of Japan. Tyto databáze se dohodly na datových standardech CBOL pro barcodové záznamy. Druhou hlavní databází je Barcode of Life Database (BOLD), která byla vytvořena a je udržována University of Guelph v Ontariu. Tato databáze nabízí vědcům způsob, jak shromažďovat, spravovat a analyzovat data DNA barcodingu. [23]

Posledním složkou je analýza dat. Vzorky jsou identifikovány podle nejbližší nalezené shody s referenčním záznamem v databázi. CBOL´s Data Analysis Working Group vytvořila the Barcode of Life Data Portal, který nabízí výzkumníkům nové a více flexibilní přístupy k uskladnění, spravování, analýze a zobrazení jejich barcodových dat. [23]

6 DENZITA NUKLEOTIDŮ

Denzita nukleotidů je jednoduchá a efektivní metoda numerické reprezentace symbolické sekvence DNA [15]. Jedná se o vytvoření vektorů, které vyjadřují průměrné zastoupení nukleotidu v definovaném úseku sekvence [15]. Tyto vektory jsou čtyři, což odpovídá počtu nukleotidů. Je tedy zřejmé, že máme jeden vektor pro každý nukleotid.

Při výpočtu nukleotidových denzitních vektorů je potřeba nejprve ze symbolické sekvence vytvořit indikační vektory $u_A[n]$, $u_C[n]$, $u_G[n]$, $u_T[n]$ [30]. To jsou vektory, které obsahují na pozici n hodnotu 1, pokud se daný nukleotid na tomto místě v sekvenci vyskytuje, nebo hodnotu 0, za podmínky, že se tam nukleotid nevyskytuje. Použitím posuvného okna o délce W na indikační vektory jsou za pomoci (1) [15] vypočítány jednotlivé denzitní vektory po každý nukleotid.

$$d_X[n] = \frac{\sum_{i=1}^{N+W-1} u_X[i]}{W}, n = 1 \dots N \qquad (1)$$

Kromě W se jako další neznámé v tomto vzorci vyskytují N, které chápeme jako délku sekvence a X, které představuje jeden ze čtyř typů nukleotidů. Výpočet funguje tak, že se posuvné okno pohybuje po celé délce indikačních vektorů a vrací průměr z hodnot v okně. Tím získáme čtyři denzitní nukleotidové vektory $d_A[n]$, $d_C[n]$, $d_G[n]$, $d_T[n]$. Vzhledem k tomu, že n je nastaven jako prostřední prvek posuvného okna, musí být velikost W liché číslo. Můžeme si ho libovolně zvolit, logicky v závislosti na délce sekvence a na požadovaném rozlišení. [15]

Během počítání ovšem dojde k problému. Pokud se při výpočtu okno posune vždy o jednu hodnotu, nově vzniklý denzitní vektor bude vždy na okrajích zkrácen o *W-1* hodnot. Tomu můžeme předejít tím, že indikační vektory prodloužíme pomocnými hodnotami, které by ideálně neměly způsobit zkreslení okrajových částí denzitních vektorů. Analýza vlivu volby pomocných vektorů je popsána v kapitole 7. Na začátek i konec indikačních vektorů tedy dodáme takové množství pomocných hodnot jako je *W/2* zaokrouhleno k nižšímu celému číslu. [15]

Jako příklad pro lepší pochopení tvoření denzitních vektorů je na Obr. 9 uvedena smyšlená sekvence nukleotidů *ATAGGTC*, ke které jsou dopočítány se žlutým podkreslením indikační vektory. Ty jsou prodlouženy pomocnými hodnotami 1, jež jsou podkresleny modře. Indikační vektory byly zpracovány za pomoci (1), kde *W=5*, díky čemuž byly získány nukleotidové denzitní vektory uvedené také na obrázku se zeleným podkreslením.

sekvence nukleotidů:			A	T	A	G	G	T	C		
u_A	1	1	1	0	1	0	0	0	0	1	1
u_C	1	1	0	0	0	0	0	0	1	1	1
u_G	1	1	0	0	0	1	1	0	0	1	1
u_T	1	1	0	1	0	0	0	1	0	1	1

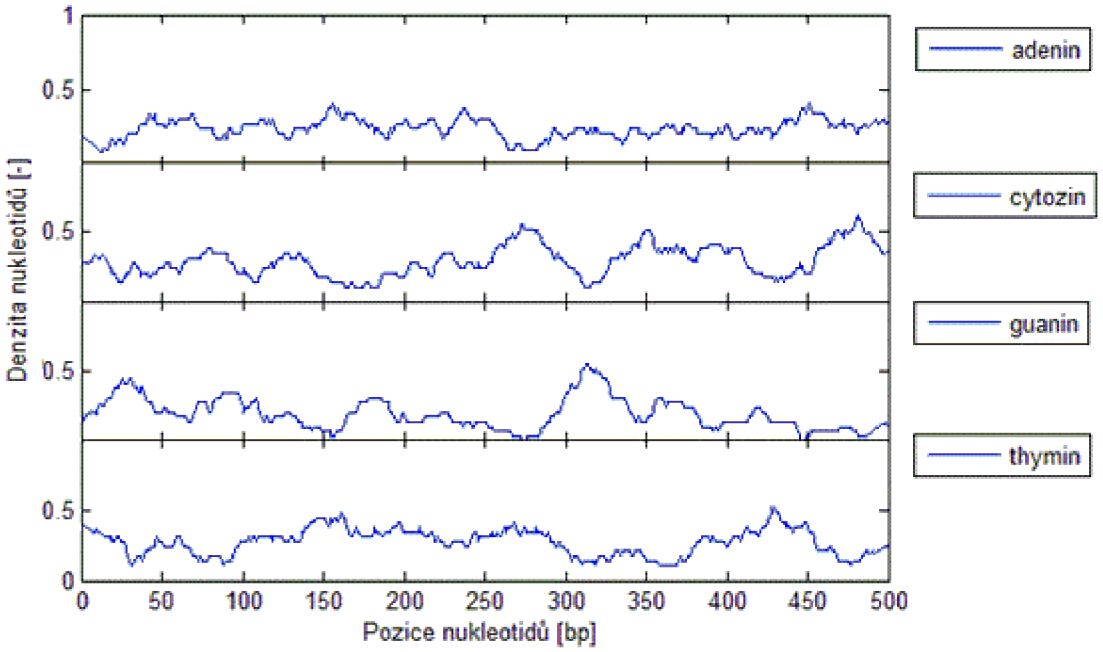
d_A	0,8	0,6	0,4	0,2	0,2	0,2	0,4
d_C	0,4	0,2	0	0	0,2	0,4	0,6
d_G	0,4	0,4	0,4	0,4	0,4	0,6	0,6
d_T	0,6	0,4	0,2	0,4	0,2	0,4	0,6

Obr. 9: Příklad výpočtu indikačních vektorů (nahoře) a denzitních vektorů s oknem 5 a pomocnými hodnotami 1 na okrajích (dole).

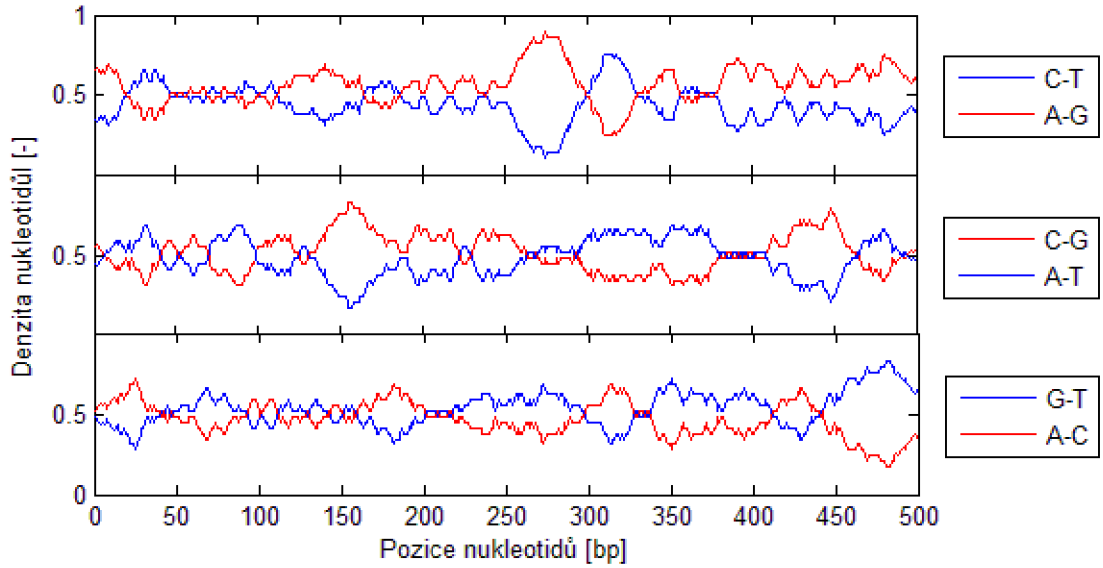
Grafickou reprezentaci denzit nukleotidů lze provést dvojím způsobem. První a nejjednodušší možností je vykreslení samostatných denzitních vektorů pro každý z nukleotidů zvlášť, jak je možné vidět na Obr. 10. Jedná se o první sekvenci z FASTA souboru FCFP dlouhou 500 párů bází. Signál na tomto obrázku vznikl při délce okna 29 nukleotidů a ošetření okrajů nejlepší variantou a to hodnotou 0,25.

Jako druhou možnost lze vytvořit sumy denzitních vektorů podle podobnosti biochemických vlastností bází v nukleotidech. Tyto vlastnosti se v sekvencích DNA rozlišují na tři hlavní typy. Podle molekulární struktury rozlišujeme báze purinové (A, G) a pyrimidinové (C, T). Další možné rozdělení je podle síly vazby mezi komplementárními bázemi - na silné, tvořené třemi vodíkovými můstky (C, G) a na slabé, které jsou tvořeny pouze dvěma můstky (A, T). Poslední variantou jak báze rozdělit je podle obsaženého radikálu. Ve struktuře bází se může objevit *amino* skupina NH_3 (A, C) nebo *keto* skupina $C{=}O$ (G, T).

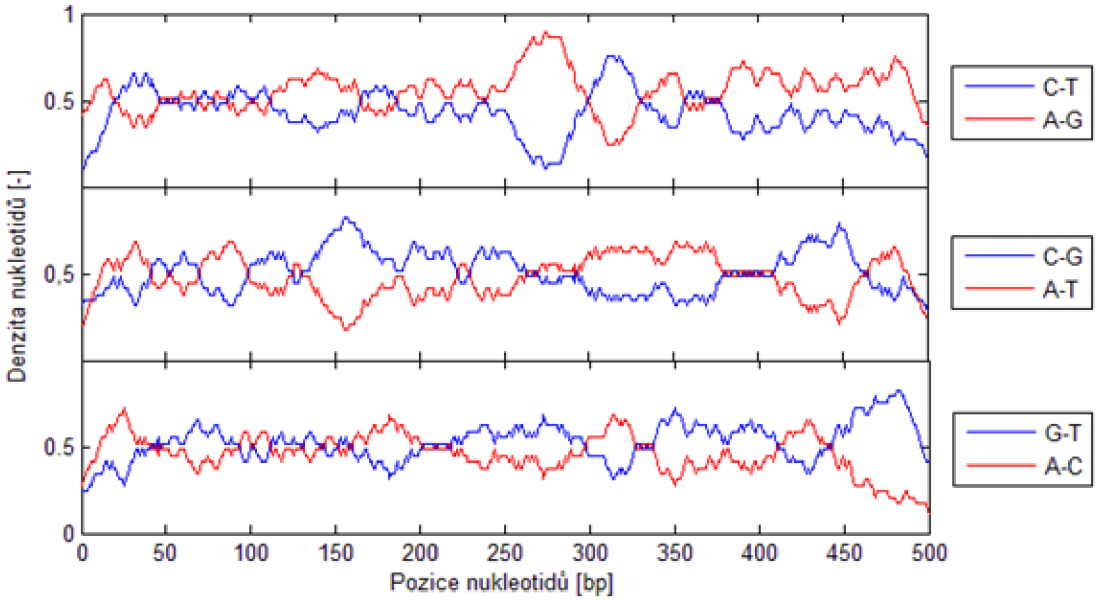
Sumy denzitních vektorů podle biochemických vlastností nukleotidů byly vytvořeny opět z první sekvence FASTA souboru FCFP dlouhé 500 párů bází. Vždy bylo využito okno dlouhé 29 nukleotidů a měnilo se akorát okrajové ošetření. Na Obr. 11 jsou okraje ošetřeny hodnotou 0,25, díky čemuž dochází k nejmenšímu zkreslení koncových hodnot. Na Obr. 12 a 13 jsou okraje ošetřeny hodnotami 0 a 1 a je patrné, že koncové hodnoty jsou mnohem více zkreslené.



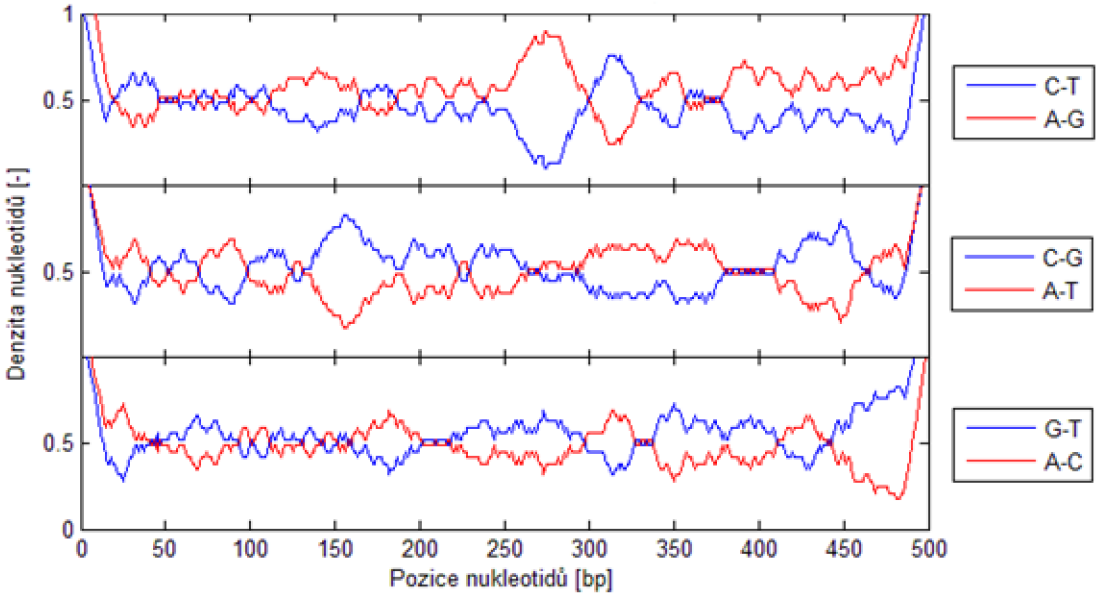
Obr. 10: Průběh nukleotidové denzity při délce okna 29 nukleotidů a ošetření okrajů hodnotou 0,25. Jedná se o sekvenci druhu *Anthias anthias* (první sekvence) z FASTA souboru FCFP dlouhou 500 bp.



Obr. 11: Průběhy sum nukleotidových denzit podle biochemických vlastností při délce okna 29 nukleotidů a ošetření okrajů hodnotou 0,25. Jedná se o sekvenci druhu *Anthias anthias* (první sekvence) z FASTA souboru FCFP dlouhou 500 bp.



Obr. 12: Průběh nukleotidové denzity podle biochemických vlastností při délce okna 29 nukleotidů a ošetření okrajů hodnotou 0. Jedná se o sekvenci druhu *Anthias anthias* (první sekvence) z FASTA souboru FCFP dlouhou 500 bp.



Obr. 13: Průběh nukleotidové denzity podle biochemických vlastností při délce okna 29 nukleotidů a ošetření okrajů hodnotou 1. Jedná se o sekvenci druhu *Anthias anthias* (první sekvence) z FASTA souboru FCFP dlouhou 500 bp.

7 PRAKTICKÁ ČÁST

7.1 Analýza vlivu pomocných hodnot

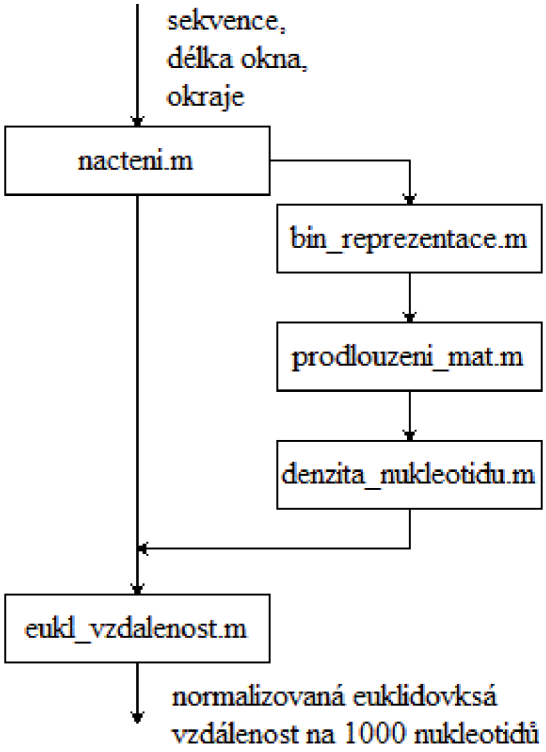
Okraje sekvencí mohou být ošetřeny různými způsoby. Jak bylo již výše zmíněno, je to důležité k tomu, aby došlo k co nejmenšímu zkreslení při výpočtu denzitních vektorů. Pro analýzu vlivu pomocných hodnot na zkreslení okrajových částí denzitních vektorů byly zvoleny tři hodnoty – 0, 1 a 0,25. Vybraná číslice 0 značí nepřítomnost některého z nukleotidů na přidaných pozicích a číslo 1 naopak jejich přítomnost. Třetí možností je hodnota 0,25, která se volí z důvodu, že odpovídá 25% pravděpodobnosti výskytu nukleotidu na tomtéž místě jako dříve uvedené 0 a 1.

Tab. č. 1: Vytvořené funkce a jejich vstupní a výstupní parametry.

Funkce	Vstup	Volá si	Výstup
bin_reprezentace.m	sekvence	-	indikační vektory pro jednotlivé nukleotidy
prodlouzeni_mat.m	sekvence, okraje, délka okna	*bin_reprezentace.m*	indikační vektory pro jednotlivé nukleotidy s ošetřenými okraji
denzita_nukleotidu.m	sekvence, okraje, délka okna	*prodlouzeni_mat.m*	denzitní vektory pro jednotlivé nukleotidy
nacteni.m	sekvence, okraje, délka okna	*denzita_nukleotidu.m*	euklidovská vzdálenost mezi referenční a testovanou sekvencí
eukl_vzdalenost.m	denzitní vektory referenční a testované sekvence	-	euklidovská vzdálenost mezi referenční a testovanou sekvencí

Při samotné analýze se vychází z FASTA souboru s názvem FCFP (Fishes of Portugal – West Coast 1), který obsahuje 188 DNA sekvencí ryb z portugalského pobřeží a který byl získán z databáze BOLD. Pro výpočet je použito pět funkcí, které jsou popsány v Tab. č. 1 a jejich blokové schéma lze vidět na Obr. 14. FASTA soubor spolu s volitelnou délkou okna a vybranou okrajovou hodnotou vstupují do funkce *nacteni.m*. Tato funkce si volá pomocnou funkci *denzita_nukleotidu.m*, aby vypočítala denzitní vektory pro referenci a analyzovanou sekvenci – obě zkrácené na 500 párů

bází, a dále funkci *eukl_vzdalenost.m*, která vypočítané vektory porovná podle (2). Tato funkce počítá eukleidovskou vzdálenost nukleotidových denzit pro analyzovanou a referenční sekvenci a její výsledná hodnota je výstupem funkce *nacteni.m*.



Obr. 14: Blokové schéma vytvořených funkcí.

Aby bylo jasné, co je referenční a analyzované sekvence (viz Obr. 15), bude následující odstavec věnován tomuto objasnění. Referenční sekvence je v tomto případě sekvence dlouhá 500 párů bází. Z ní je vytvořen vektor denzit pro jednotlivé nukleotidy a zkrácen opět na 500 párů bází a potom ještě z obou konců o délku okna. Analyzovaná sekvence se vytvoří tak, že se nejprve na obou koncích zkrátí o délku okna, tyto okraje se ošetří hodnotami 0, 1 nebo 0,25 a teprve potom se z ní vytvoří denzitní vektory. Denzitní vektory obou sekvencí jsou stejně dlouhé.

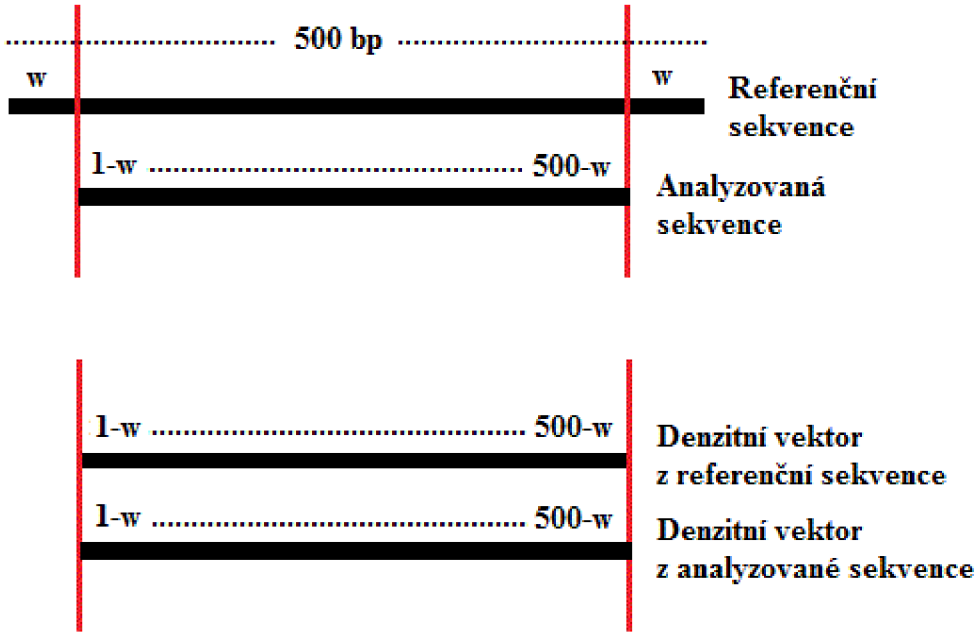
V programu se následně počítá rozdíl mezi referenčními a analyzovanými denzitními vektory jako eukleidovská vzdálenost:

$$E_{i,j} = \frac{1}{N_Z}\sqrt{(d_{A,i} - d_{A,j})^2 + (d_{C,i} - d_{C,j})^2 + (d_{G,i} - d_{G,j})^2 + (d_{T,i} - d_{T,j})^2} \qquad (2)$$

kde $d_{A,i}$ je denzitní vektor referenční sekvence a $d_{A,j}$ je denzitní vektor vytvořený pro

analyzovanou sekvenci – obě pro adenin, jak napovídá dolní index A. Zbytek neznámých je možné si vyložit analogicky. Poslední neznámou je N_Z, což je číslo, kterým se výsledek normalizuje.

Eukleidovská vzdálenost je normalizována na 1000 párů bází, aby produkované hodnoty mohly být dobře porovnatelné pro všechny délky oken (lichá čísla v rozmezí 5 – 29). Čím nižší vyjde hodnota, tím blíže je analyzovaná sekvence k referenci a tím lepšího výsledku bylo dosaženo.



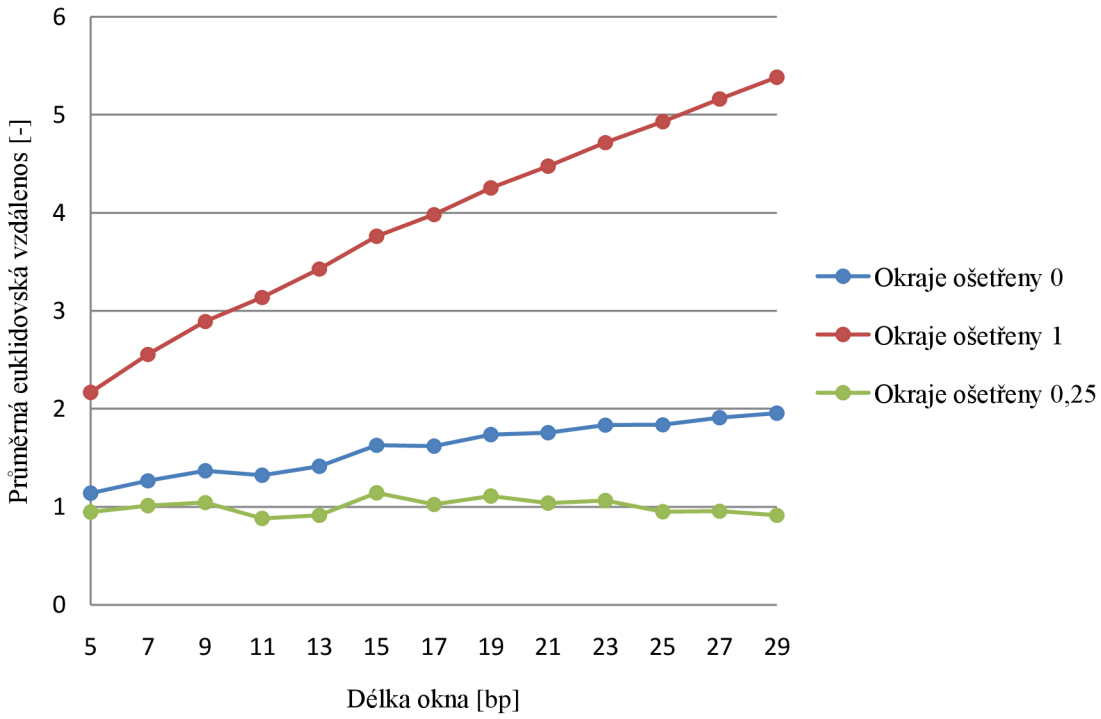
Obr. 15: Nukleotidové denzitní vektory analyzované a referenční sekvence.

Při analýze se tedy každá sekvence převedla na denzitní vektor a pro každou délku okna ošetřila všemi třemi hodnotami objasněnými na začátku této kapitoly. Z výsledných hodnot funkce *nacteni.m* byla následně pro každou délku okna vypočítaná eukleidovská vzdálenost, její průměrná hodnota pro všechny sekvence, směrodatná odchylka a vytvořen graf.

Jak ukazuje Obr. 16, nejlepších výsledků bylo dosaženo pro okrajové hodnoty 0,25 a nejhorších při zvolení čísla 1. Celkové vyhodnocení výsledků pro jednotlivé ošetření okrajů a délku okna je možné vidět v Tab. č. 2, kde jsou uvedené průměry eukleidovské vzdálenosti a směrodatné odchylky pro jednotlivé délky okna.

Tab. č. 2: Průměrné euklidovské vzdálenosti a jejich směrodatné odchylky pro 3 typy ošetření okrajů a délky okna 5 – 29 bp.

Délka výpočetního okna [bp]	Ošetření okrajů 0		Ošetření okrajů 0,25		Ošetření okrajů 1	
	Ē	σ	Ē	σ	Ē	σ
5	1,1379	0,1179	0,9452	0,1207	2,1676	0,0650
7	1,2671	0,0929	1,0112	0,0964	2,5568	0,0736
9	1,3654	0,1406	1,0443	0,1626	2,8932	0,0970
11	1,3197	0,1144	0,8803	0,1494	3,1386	0,0991
13	1,4124	0,1031	0,9143	0,1256	3,4291	0,1067
15	1,6298	0,1185	1,1410	0,1110	3,7597	0,1074
17	1,6188	0,1401	1,0252	0,1566	3,9838	0,1176
19	1,7357	0,1508	1,1076	0,1680	4,2532	0,1246
21	1,7559	0,1262	1,0359	0,1503	4,4763	0,1378
23	1,8351	0,1239	1,0633	0,1356	4,7191	0,1453
25	1,8384	0,1219	0,9491	0,1477	4,9291	0,1555
27	1,9095	0,1225	0,9565	0,1666	5,1620	0,1651
29	1,9553	0,1384	0,9136	0,1841	5,3836	0,1711



Obr. 16: Průměrné eukleidovské vzdálenosti mezi nukleotidovými denzitami pro různé hodnoty ošetření okrajů a délky výpočetního okna.

7.2 Identifikace organismů pomocí nukleotidových denzit

Nukleotidové denzitní vektory lze využít jako prostředek k identifikaci organismů. Princip je jednoduchý a spočívá v tom, že se nejprve vytvoří databáze referenčních denzitních vektorů a s ní se následně porovnávají vektory denzit analyzovaných organismů. Na základě euklidovské vzdálenosti mezi analyzovanou a referenční sekvencí je pak reference s nejmenší euklidovskou vzdáleností přiřazena k analyzovanému druhu, čímž je druh identifikován.

Pro tuto identifikační analýzu se k vytváření denzit používají pomocné hodnoty 0,25, jelikož vyšly jako nejvhodnější – nejméně zkreslily okraje signálu u zkušebního souboru FCFP, viz kapitola 7.1.

K verifikaci metody identifikace založené na porovnávání nukleotidových denzitních vektorů byly zvoleny FASTA soubory sekvencí části mitochondriálního genu *cox1* z třídy ryb, hmyzu (řád chrostíci), ptáků a savců (podřád netopýři).

7.2.1 Vytvoření referenčního souboru

Prvním krokem pro identifikační analýzu je vytvoření referenčního souboru nukleotidových denzitních vektorů. Tento referenční soubor byl vytvořen ze čtyř FASTA souborů získaných z databáze BOLD. Mezi zvolené soubory patří soubor FCFP (ryby z portugalského pobřeží), soubor CUCAD (chrostíci z Churchillu), soubor MNCN (neotropičtí ptáci) a soubor BCBNC (netopýři z Guyany). Množství sekvencí, druhů a rodů v jednotlivých souborech ukazuje Tab. č. 3.

Tab. č. 3: Počty sekvencí, druhů a rodů v referenčních souborech.

Soubor	Počet sekvencí	Počet druhů	Počet rodů
FCFP	188	48	36
CUCAD	716	54	26
MNCN	758	43	36
BCBNC	840	94	50

Konečný referenční soubor vznikl spojením všech výše zmíněných FASTA souborů. Následně byly všechny sekvence ve spojeném souboru převedeny na denzitní vektory. Některé druhy jsou v souboru zastoupeny několika sekvencemi různých jedinců. V tomto případě byla referenční denzita vytvořena zprůměrováním vícenásobně zarovnaných denzitních vektorů jednotlivých sekvencí.

Je důležité si uvědomit, že vektory denzit se mění na základně délek posuvného okna. V závislosti na tomto poznatku tedy bylo zapotřebí vytvořit i patřičný počet referenčních souborů. Vzhledem k tomu, že pro testování byla zvolena lichá okna o délkách 3 – 19 bp, referenčních souborů muselo vzniknout devět. Každý pro každou velikost výpočetního okna.

Realizace referencí proběhla v prostředí Matlab za pomoci několika funkcí. Nejprve se jeden z výše zmíněných FASTA souborů pomocí funkce *ZAROVNANI_BUNEK.m* převedl na *cell* buňky, které v prvním sloupci měly vypsány všechny názvy druhů (každý druh pouze jednou) a v dalších sloupcích jim odpovídající sekvence, které se zarovnaly, pokud k některému druhu patřilo více sekvencí. Ukázku výstupu této funkce zobrazuje Tab. č. 4.

Tab. č. 4: Příklad výstupu funkce *ZAROVNAN_BUNEK.m* pro prvních sedm sekvencí ze souboru FCFP.

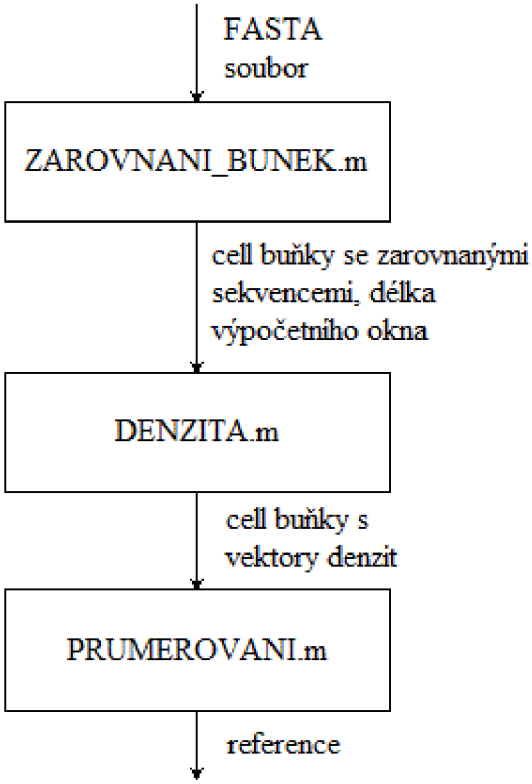
Název druhu	**Sekvence**			
Anthias anthias	CCTCTACC…			
Argentina sphyraena	CCTTTACC…	CCTTTACC…	CCTTTACC…	CCTTTACC…
Arnoglossus imperialis	CCTCTATC…	CCTCTATC…		

Tab. č. 5: Příklad výstupu funkce *DENZITA.m* pro prvních sedm sekvencí ze souboru FCFP a pro délku výpočetního okna 19 bp.

Název druhu	**Denzitní vektory**			
Anthias anthias	0,22 0,21 … 0,38 0,37 … 0,12 0,16 … 0,28 0,26 …			
Argentina sphyraena	0,17 0,16 … 0,33 0,32 … 0,17 0,21 … 0,33 0,32 …	0,17 0,16 … 0,33 0,32 … 0,17 0,21 … 0,33 0,32 …	0,17 0,16 … 0,33 0,32 … 0,17 0,21 … 0,33 0,32 …	0,17 0,16 … 0,33 0,32 … 0,17 0,21 … 0,33 0,32 …
Arnoglossus imperialis	0,17 0,16 … 0,38 0,37 … 0,12 0,16 … 0,33 0,32 …	0,17 0,16 … 0,38 0,37 … 0,12 0,16 … 0,33 0,32 …		

Soubor takto vzniklých *cell* buněk spolu s délkou výpočetního okna *W* byly vstupem pro funkci *DENZITA.m*, která převedla sekvence na denzitní vektory (viz Tab. č. 5

Tab. č. 5), což se stalo vstupem do konečné funkce *PRUMEROVANI.m*. Poslední zmíněná funkce pouze zprůměruje denzity druhů, které obsahovaly více sekvencí, čímž vytvoří referenční denzitní vektory. Tímto procesem, který je také blokově znázorněn na Obr. 17, prošly všechny čtyři FASTA soubory, spojily se a tím vytvořily konečný referenční soubor. Celé se to zopakovalo devětkrát – pro devět výpočetních oken, jak bylo zmíněno již výše.



Obr. 17: Blokové schéma vytvoření referenční denzity pro druh z několika sekvencí.

7.2.2 Přiřazení referenčního druhu k analyzovanému

Samotná analýza proběhla s pěti FASTA soubory, které byly získány opět z databáze BOLD a které obsahují převážně stejné druhy organismů, jako byly organismy referenční. Mezi analyzované soubory byly zařazeny soubory FCFPS a FCFPW (ryby z portugalského jižního a západního pobřeží), dále soubor DSTRI (chrostíci z Churchillu), soubor BRAS (neotropičtí ptáci) a nakonec soubor BCDR (netopýři z Guyany). Vzhledem k tomu, že identifikační analýza byla výpočetně náročná, poslední tři soubory byly zkráceny (o 469, 492 a 3867 sekvencí). Toto zkrácení

proběhlo tak, že se od každého druhu ponechalo pouze maximálně pět sekvencí a navíc byly odstraněny ty sekvence, které nemají referenční druh a ani nepatří ke stejnému rodu. Množství sekvencí, druhů a rodů v jednotlivých souborech po zkrácení ukazuje Tab. č. 6.

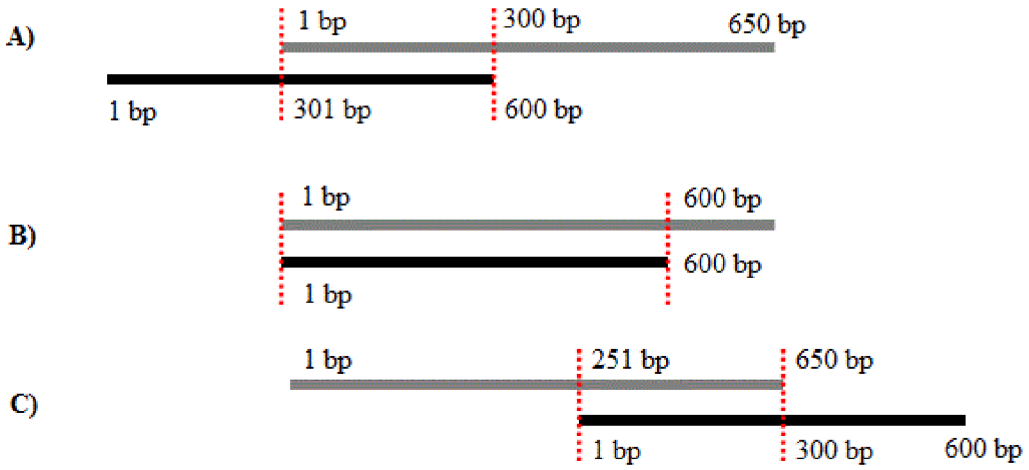
Tab. č. 6: Počty sekvencí, druhů a rodů v souborech určených k analýze.

Soubor	Počet sekvencí	Počet druhů	Počet rodů	Počet sekvencí s referencí pro druh	Počet sekvencí s referencí pro rod
FCFPS	200	55	47	137	156
FCFPW	207	49	42	134	156
DSTRI	71	20	9	40	71
BRAS	145	73	29	85	145
BCDR	267	61	36	254	267

Přiřazení referencí k analyzovaným sekvencím se uskutečnilo opět v programu Matlab prostřednictvím skriptu *ANALYZA.m*. Pro zkrácené soubory byl vytvořen skript ANALYZA2.m – tyto soubory nejsou ve FASTA formátu, nýbrž mají příponu *.mat.* Tento skript si postupně načítá analyzované soubory, všechny délky výpočetního okna, reference (právě podle délky posuvného okna) a výběr nukleotidů, podle kterých se následně počítají euklidovské vzdálenosti. Možných výběrů je osm. Mezi první čtyři patří samostatné nukleotidy (A, C, G, T), dalšími třemi jsou nukleotidy podle biochemických vlastností (AC – obsahují amino skupinu, AG - pyrimidiny, CG – 3 vodíkové můstky mezi komplementárními vlákny) a jako poslední byly zvoleny všechny nukleotidy dohromady. Výstupem celého skriptu je 360 xls souborů - tabulek, kde jsou v prvním sloupci vypsány všechny analyzované druhy, ve druhém minimální euklidovské vzdálenosti a ve třetím přiřazené reference (někdy jich je přiřazeno více).

Jelikož sekvence analyzovaného a referenčního druhu mají různou délku, samotný proces přiřazení začíná signálovým zarovnáním. Hledá se vzájemná poloha analyzované a referenční sekvence, při které mezi nimi nastane nejmenší euklidovská vzdálenost (podle všech čtyř nukleotidů za pomoci funkce *PRIRAZENI.m*). Pátrání po optimálním zarovnání zobrazuje Obr. 18. Jak je vidět, jde o to, že se kratší sekvence posunuje po delší (vždy o jeden nukleotid) a na překrývajících se úsecích (na obrázku vyznačeny svislými čárami) se počítá euklidovská vzdálenost mezi jejich denzitními vektory získanými díky funkci *denzita_nukleotidu.m*. Minimální délka společného úseku je vždy polovina kratší sekvence, jak také ukazuje obrázek. Jakmile kratší sekvence dojde na konec, vznikne vektor všech euklidovských vzdáleností, ze kterých je vybrána ta nejmenší. To je hledané optimální zarovnání.

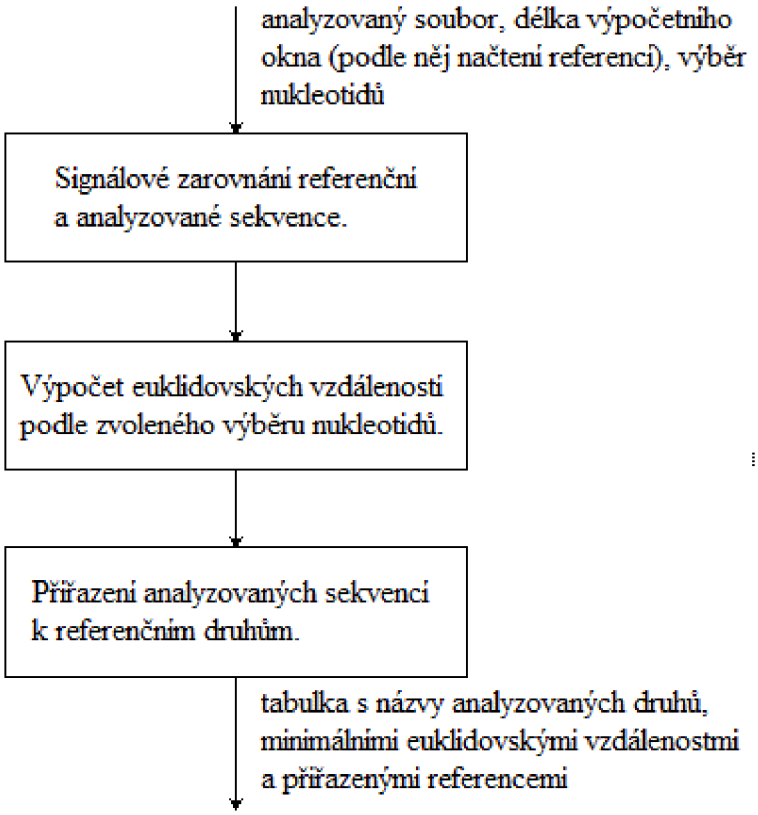
Identifikační analýza tedy probíhá tak, že se vezme analyzovaná sekvence, s ní se postupně výše zmíněným způsobem zarovnají všechny reference a zaznamenají se úseky, které byly zvoleny pro každé zarovnání jako nejvhodnější. Na těchto úsecích se pak počítají znovu euklidovské vzdálenosti, ale ty již slouží k identifikaci. Právě zmíněné euklidovské vzdálenosti jsou počítány pomocí již vzpomenuté funkce *PRIRAZENI.m*, do které vstupují výběr nukleotidů, podle kterých má být euklidovská vzdálenost vypočtena, a denzitní vektory referenční i analyzované sekvence.



Obr. 18: Posun kratší sekvence po delší při signálovém zarovnání z výchozího stavu (A) ke konečnému stavu (C).

K analyzovanému druhu je přiřazen ten referenční druh, u kterého vyšla euklidovská vzdálenost ze všech referenčních sekvencí nejmenší. Název analyzovaného druhu, tato minimální vzdálenost a přiřazený referenční druh (nebo více referencí) jsou pak zapsány do prvního řádku *cell* buněk. Zjednodušené blokové schéma přiřazení popisuje Obr. 19.

Tento celý proces proběhne pro všechny analyzované sekvence, čímž se vyplní všechny řádky v *cell* buňkách. Vznikne výsledná tabulka, která je exportována do Microsoft Office Excel. Příklad části výsledné tabulky pro soubor FCFPS, pro délku výpočetního okna 19 bp a pro výpočet euklidovských vzdáleností podle adeninu ukazuje Tab. č. 7.



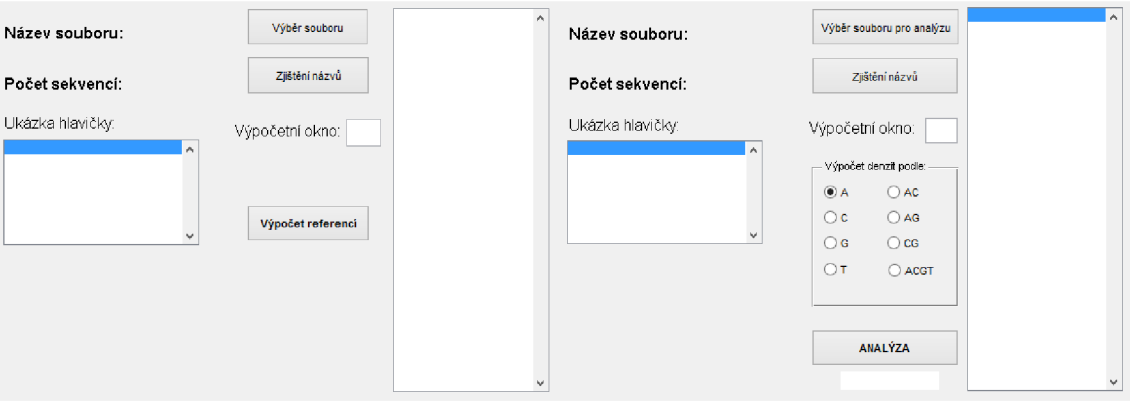
Obr. 19: Blokové schéma skriptu *ANALYZA.m.*

Tab. č. 7: Příklad výstupu skriptu *ANALYZA.m* pro prvních 7 sekvencí souboru FCFPS.fas (délka výpočetního okna 19 bp a výpočet euklidovských vzdálenosti podle A).

Druh analyzované sekvence	**Minimální euklidovská vzdálenost**	**Přiřazený referenční druh**
Actinopterygii	0,0024	*Serranus hepatus*
Argentina sphyraena	0,0005	*Argentina sphyraena*
Argentina sphyraena	0,0013	*Argentina sphyraena*
Belone belone	0,0024	*Conger conger*
Benthodesmus simonyi	0,0024	*Scomber scombrus*
Benthodesmus simonyi	0,0024	*Galbula albirostris*
Boops boops	0	*Boops boops*

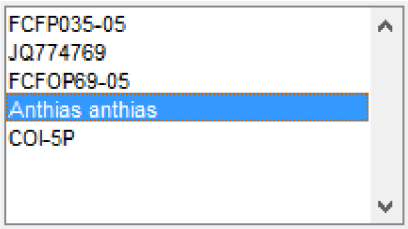
7.2.3 GUI aplikace pro identifikační analýzu

Aby byla identifikační analýza uživatelsky příjemnější, byla vytvořena v programu Matlab uživatelská aplikace *Identifikacni_analyza.m*, která tvorbu referencí a následnou identifikaci druhů zjednodušuje. Čelní panel této aplikace je možné vidět na Obr. 20.



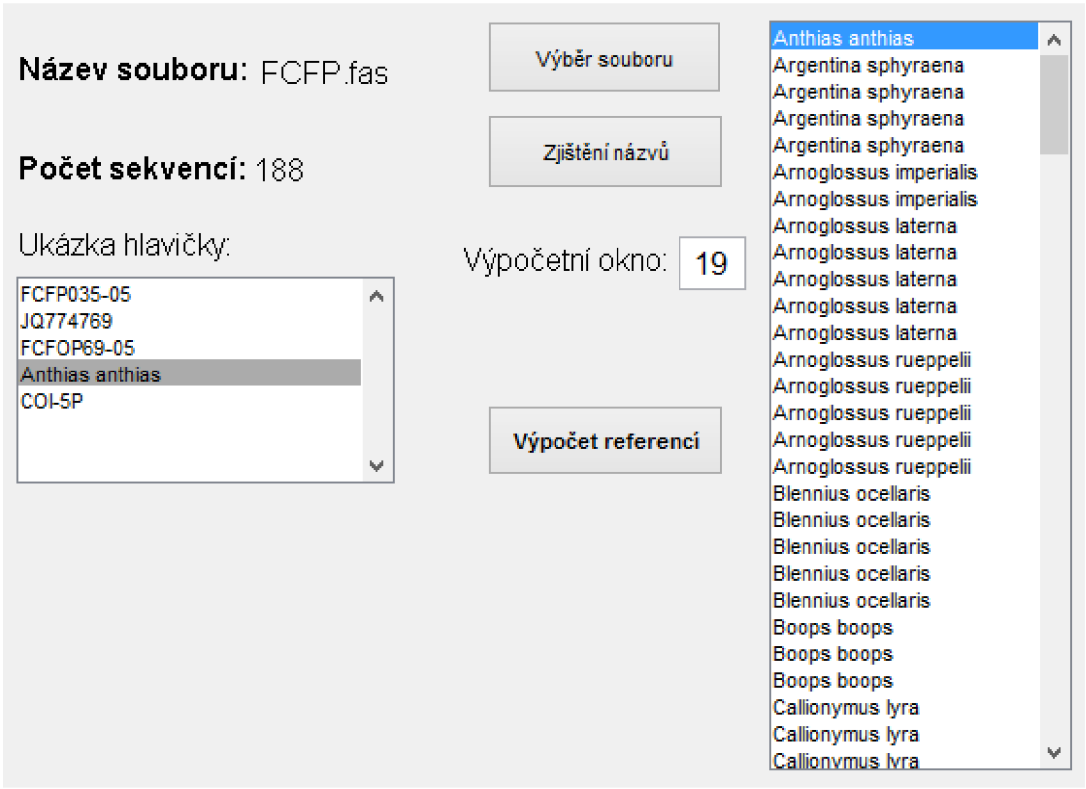
Obr. 20: Čelní panel aplikace *Identifikacni_analyza.m.*

Levá část panelu slouží k vytvoření referencí. Po kliknutí na tlačítko *Výběr souboru* se otevře klasické okno systému Windows, ve kterém uživatel zvolí FASTA soubor, který má posloužit k tvorbě referencí. Jakmile je soubor vybrán, zapíše se jeho název, počet sekvencí a ukázka hlavičky prvního druhu na patřičná místa na panelu. Pokud chce mít uživatel vypsány názvy všech druhů vyskytujících se ve zvoleném souboru, stačí označit řádek s názvem druhu v *Ukázce hlavičky* a kliknout na tlačítko *Zjištění názvů*. Názvy jsou pak vypsány do listboxu díky funkci *JMENA_GUI.m*. Ukázku hlavičky pro první druh souboru FCFP *Anthias anthias* znázorňuje Obr. 21.

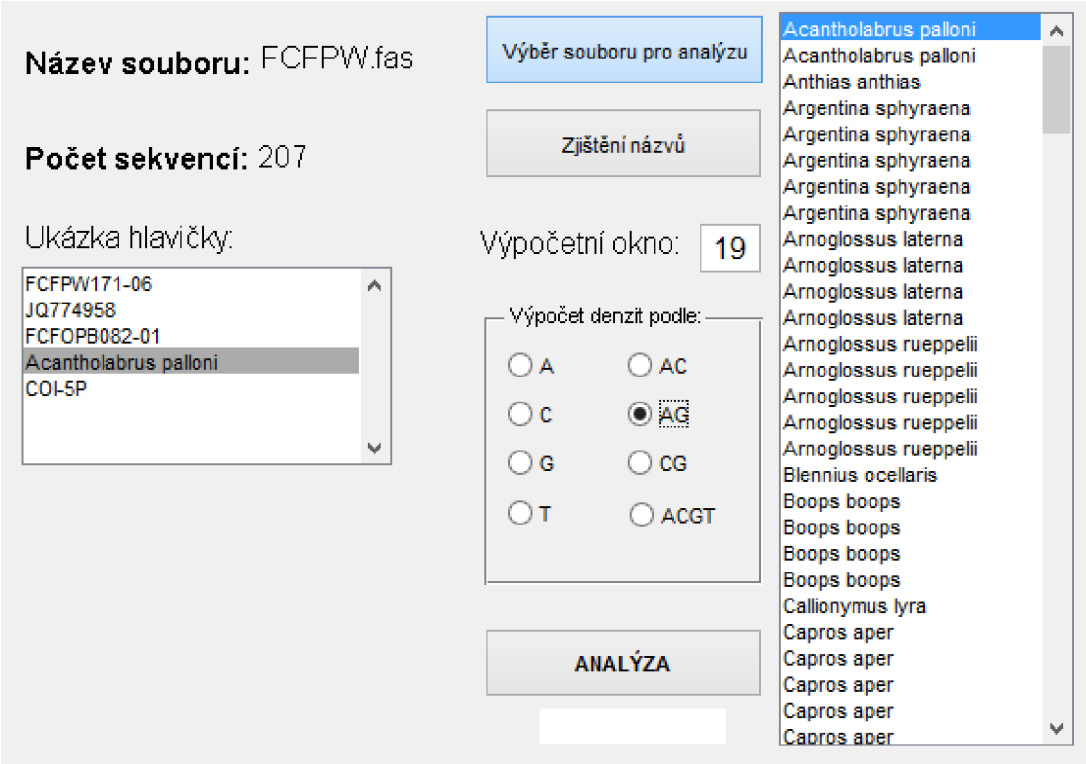


Obr. 21: Ukázka hlavičky pro první druh referenčního souboru FCFP.

Následně je potřeba napsat do editačního okna *Výpočetní okno* hodnotu posuvného okna, se kterou mají být vypočteny denzitní vektory a Reference se tvoří po kliknutí na tlačítko *Výpočet referencí*, které si na výpočet volá již dříve popsané funkce z kapitoly 7.2.1, a to funkce ZAROVNANI*_BUNEK_GUI.m* (je to upravená funkce *ZAROVNANI_BUNEK.m* pro GUI), *DENZITA.m* a *PRUMEROVANI.m*. Jakmile je soubor referencí vytvořen, objeví se opět okno, které se uživatele zeptá, kam chce soubor uložit. Název souboru je generován automaticky. Je to vždy slovo „reference“, k němuž je za podtržítkem připsána délka výpočetního okna. Pro lepší představu je na Obr. 22 znázorněna levá část čelního panelu ukazující tvorbu referencí ze souboru FCFP s délkou výpočetního okna 19 bp.



Obr. 22: Ukázka levé části čelního panelu při tvorbě referencí ze souboru FCFP.



Obr. 23: Ukázka pravé části čelního panelu při identifikační analýze souboru FCFPW.

Pravá část čelního panelu, znázorněná na začátku této kapitoly, slouží již k identifikaci druhů a obsahuje velmi podobná tlačítka jako jeho levá část. Oproti ní se ale zde nachází panel pro výběr nukleotidů, podle nichž mají být počítány euklidovské vzdálenosti, na jejichž základě dochází k přiřazování referencí k analyzovaným druhům, a tlačítko *ANALÝZA*, které spouští celý výpočet. Ten zprostředkovává funkce *ANALYZUA_GUI.m* (je to upravená funkce *ANALYZA.m* pro GUI). Pravou část čelního panelu s načteným souborem FCFPW a výpočetním oknem 19 bp je možné vidět na Obr. 23. Vzhledem k tomu, že výpočet je časově náročný, je aplikace doplněna o bílé pole (hned pod tlačítkem ANALÝZA), do kterého se červeným písmem napíše „HOTOVO“, jakmile je výpočet u konce. Výsledná tabulka je uložena do xls souboru, který má automaticky generovaný název.

7.3 Výsledky identifikační analýzy

Testování metody identifikace porovnáváním referenčních denzitních vektorů pro druh s denzitními vektory analyzované sekvence proběhlo na referenčních souborech popsaných v kapitole 7.2.1 a analyzovaných souborech z kapitoly 7.2.2. Pro každý z pěti souborů určených k analýze byla vytvořena tabulka, jež ukazuje úspěšnost přiřazení analyzovaných sekvencí ke správnému referenčnímu druhu a také přiřazení analyzovaných sekvencí nereferenčních druhů alespoň k referenčnímu druhu stejného rodu. Analýza byla provedena pro všechny délky výpočetních oken (3 až 19) a všechny varianty výběru nukleotidů pro výpočet euklidovské vzdálenosti (viz kapitola 7.2.2).

Hodnoty do tabulek úspěšností identifikační analýzy pro druh byly získány pomocí funkce *vyhodnoceni_pro_druh.m*. Tato funkce zjistila, které z analyzovaných druhů mají referenci a kolik druhů s referencí bylo správně přiřazeno. Po vydělení těchto hodnot a následném vynásobení stem byl získán výsledek v procentech. Vyhodnocení úspěšnosti identifikace rodů proběhlo analogicky díky funkci *vyhodnoceni_pro_rod.m*. Výsledky obou funkcí se automaticky zapsaly na druhý a třetí list patřičného xls souboru, který byl popsán v kapitole 7.2.2.

Při hodnocení úspěšnosti nás nejvíce zajímá, jaký výběr nukleotidů a jaká délka výpočetního okna byly pro analýzu nejvhodnější. Tedy kdy došlo k nejúspěšnější identifikaci organismů. Předem je důležité připomenout, že každý soubor obsahoval jiné typy organismů, a proto u každého vyšly jiné výsledky. Výsledky jsou kromě vlastností metody identifikace také ovlivněny samotnými vlastnostmi sekvencí, tj. vnitrodruhovou a mezidruhovou variabilitou sekvencí. K tomu, aby identifikace proběhla správně, je totiž potřeba co nejmenší vnitrodruhová a co největší mezidruhová variabilita. Pro

některé druhy to neplatí, což snižuje procento úspěšné identifikace. Referenční sekvence pro takovéto druhy není zcela adekvátní – na základě minimální euklidovské vzdálenosti se analyzovaný druh přiřadí k jinému, bližšímu, referenčnímu druhu (kvůli malé mezidruhové variabilitě). Navíc testované soubory obsahují relativně malý počet druhů a ty nedostatek jedinců. Tudíž jsou reference tvořeny z malého množství sekvencí a vnitrodruhová variabilita není zcela podchycena.

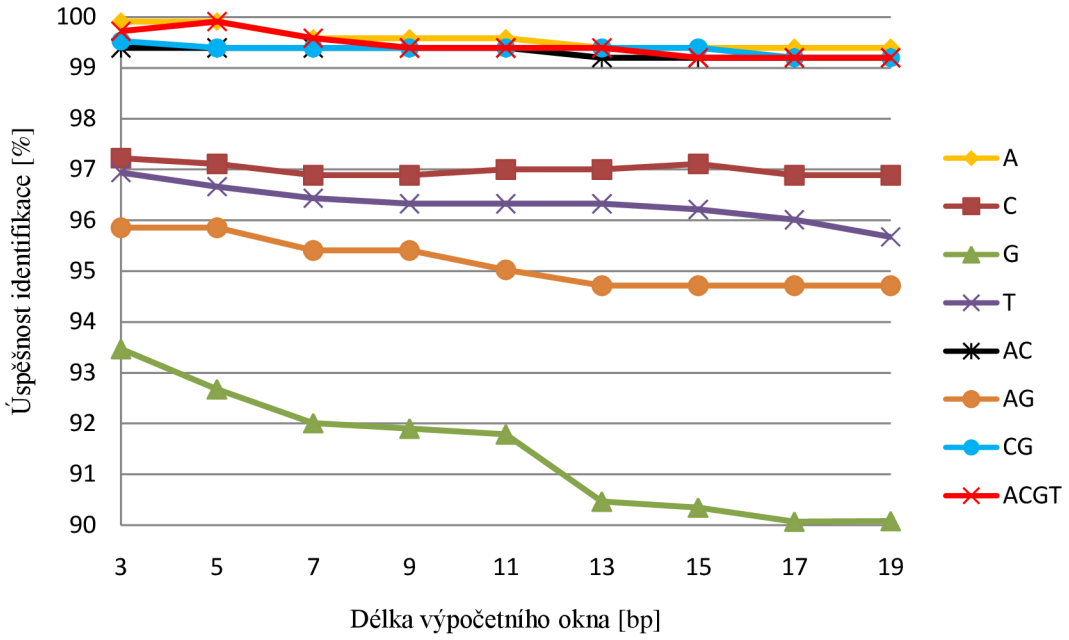
7.3.1 Úspěšnost správné identifikace druhu

Pro každý z pěti analyzovaných souborů vyšly vysoké hodnoty úspěšností přiřazení k druhům i rodům. Pouze pro soubor BRAS a počítání euklidovských vzdáleností jen mezi denzitními vektory pro nukleotid guanin došlo k nejnižšímu procentu úspěšného přiřazení. Zatímco u ostatních souborů úspěšnost v závislosti na velikosti výpočetního okna a volbě výpočtu euklidovských vzdáleností neklesla pod 90,16%, u tohoto souboru při výpočtu euklidovských vzdáleností podle G došlo v nejhorším případě ke správnému přiřazení pouze u 61,18% druhů, viz Tab. č. 14.

Dále se soubory shodují ve vlivu okna – buď nemá vliv vůbec, což platí pouze pro soubor FCFPW (viz Tab. č. 11) nebo mírně snižuje úspěšnost přiřazení. Je to tím, že se s rostoucím oknem zmenšuje rozdíl mezi sekvencemi. Zde jsou ale dvě výjimky, a to opět u souboru BRAS, kde toto tvrzení neplatí v několika málo případech - jak je možné vidět v Tab. č. 14 a u souboru BCDR, kde to neplatí pro nukleotid G a délku výpočetního okna 19 bp, což ukazuje Tab. č. 12.

Jako nejvhodnější soubor pro tuto analýzu vyšel soubor FCFPW. U něj totiž došlo ke správnému přiřazení všech analyzovaných druhů. U dalších čtyř souborů už byl patrný vliv velikosti výpočetního okna. Nejmenší u souboru DSTRI, kde plovoucí okno ovlivňovalo úspěšnost pouze při volbě nukleotidů G, T, což ukazuje Tab. č. 13 a u souboru BCDR, kde k ovlivnění úspěšnosti posuvným oknem došlo pouze pro nukleotidy G a sumy AG, jak je možné pozorovat v Tab. č. 12. Nejrozmanitější výsledky analýzy vyšly u souborů FCFPS a BRAS, jež jsou v Tab. č. 10 a Tab. č. 14.

Celkově se dá říci, že jako nejvýhodnější pro analýzu vyšlo použití samostatného A a sum AC, CG a ACGT při velikosti okna 3 a 5 bp. Nejméně se osvědčilo využití samotného G. To dokazuje Obr. 24, ve kterém jsou zobrazeny vážené průměry úspěšností identifikace druhů v závislosti na délce výpočetního okna.



Obr. 24: Vážené průměry úspěšností identifikace druhů pro všechny analyzované soubory.

7.3.2 Úspěšnost správné identifikace rodu

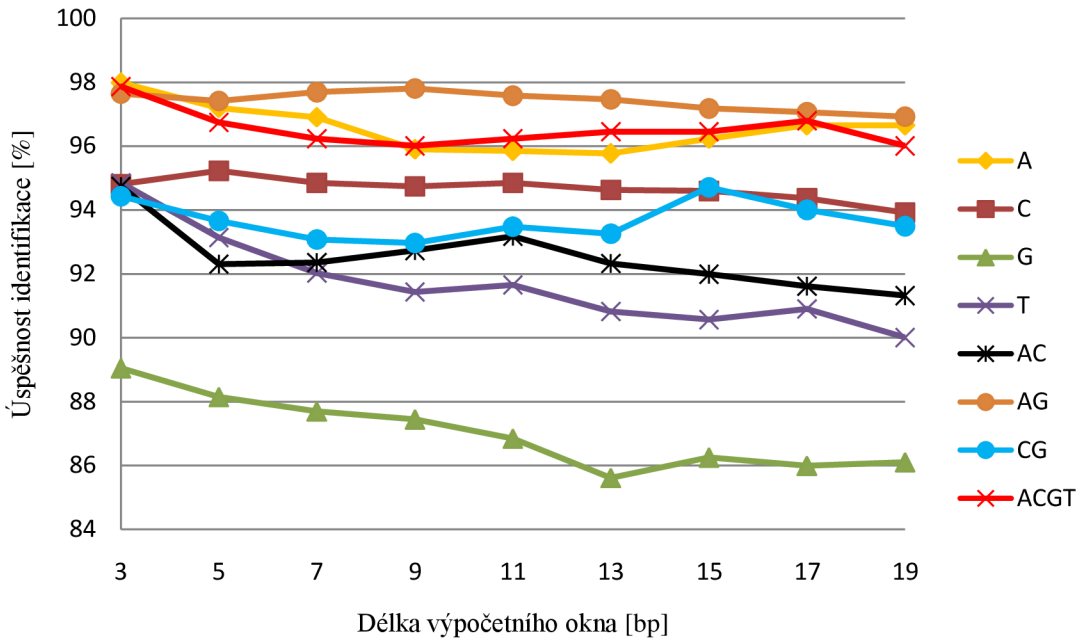
Výsledky úspěšnosti identifikační analýzy pro rod jsou pro všech pět souborů nižší než úspěšnosti druhové identifikace. Úspěšnosti neklesly pod 76,56%, což je o 13,6% méně než při identifikaci druhů. U souboru BRAS a výběru G došlo opět k nejnižší úspěšnosti identifikace a to pouze k 53,79% správně identifikovaných rodů. Celková nižší úspěšnost identifikace rodu je způsobena tím, že rod je vyšší taxonomická skupina než druh, tudíž je pro ni gen *cox1* variabilnější a identifikace pak není tak úspěšná.

Velikost okna při identifikaci rodů měla vliv u každého souboru. Neprojevila se pouze u souboru FCFPW v případě počítání euklidovských vzdáleností podle A, AG a ACGT, a u souboru BCDR, kde se neprojevila při výpočtech podle CG, viz Tab. č. 16 a Tab. č. 17.

U ostatních souborů nevyšly výsledky na základě žádného pravidla. Někdy došlo k nejvyšší úspěšnosti při délce okna 3 a jindy třeba při délce 19 bp, někdy pro výběr samotného A (viz Tab. č. 19), jindy pro sumy AG (viz Tab. č. 16, Tab. č. 17 a Tab. č. 18

Tab. č. 18) nebo při výběru všech čtyř nukleotidů dohromady (viz Tab. č. 15 a Tab. č. 16).

Kdyby měly být k identifikaci rodu vybrány pouze jedna délka výpočetního okna a jeden typ nukleotidů pro všechny soubory, byly by to 9 bp a suma AG. To ukazuje Obr. 25, na kterém je možné pozorovat závislost vážených průměrů úspěšností identifikace na délce výpočetního okna. Jak je vidět, nejméně se opět osvědčil samotný G.



Obr. 25: Vážené průměry úspěšností identifikace rodů pro všech 5 analyzovaných souborů.

7.3.3 Zhodnocení výsledků identifikační analýzy

Jak je možné pozorovat z výše vložených grafů a tabulek v příloze, identifikační analýza měla větší úspěch při identifikaci druhu. Je to tím, že druh je nižší taxonomická skupina, což bylo komentováno již v předchozí podkapitole.

Tab. č. 8: Průměrné variability a jejich směrodatné odchylky jednotlivých nukleotidů pro druhy v souborech FCFPS a FCFPW.

Analyzovaný soubor	**FCFPS**		**FCFPW**	
Nukleotid	**Průměr variabilit**	**Směrodatná odchylka**	**Průměr variabilit**	**Směrodatná odchylka**
A	0,0082	0,0116	0,0443	0,1531
C	0,0101	0,0127	0,0049	0,0067
G	0,0083	0,0105	0,0052	0,0074
T	0,0081	0,0104	0,0040	0,0051

Výsledné hodnoty byly ověřeny vypočítáním variabilit jednotlivých nukleotidů

mezi sekvencemi totožných druhů. Myšlenka je taková, že čím nižší variabilita pro analyzovaný nukleotid vyjde, tím více se v tomto nukleotidu sekvence shodují. Tím pádem by to mělo znamenat vyšší procento správně identifikovaných druhů na základě právě tohoto nukleotidu. Jak ale dokazují Tab. č. 8 a Tab. č. 9, hodnoty variabilit úplně nekorespondují s očekáváním. To však omlouvají vypočtené směrodatné odchylky – v souborech jsou totiž druhy, pro které je vše v pořádku (identifikace proběhne zcela správně), ale také druhy s vyššími variabilitami než je variabilita průměrná. Právě tyto zavádí do výsledků chybu. Samozřejmě se musí brát v úvahu i to, že testované soubory obsahují relativně malé množství sekvencí jednotlivých druhů, takže zde uvedené hodnoty variabilit plně neodpovídají skutečné variabilitě celé populace druhu.

Tab. č. 9: Průměrné variability a jejich směrodatné odchylky jednotlivých nukleotidů pro druhy v souborech BCDR, DSTRI a BRAS.

Analyz. soubor	**BCDR**		**DSTRI**		**BRAS**	
Nukleotid	**Průměr variabilit**	**Směr. odchylka**	**Průměr variabilit**	**Směr. odchylka**	**Průměr variabilit**	**Směr. odchylka**
A	0,0222	0,1095	0,0085	0,0109	0,0985	0,0469
C	0,0302	0,1129	0,0127	0,0098	0,1027	0,0420
G	0,0276	0,1112	0,0180	0,0185	0,1067	0,0574
T	0,0218	0,1061	0,0089	0,0100	0,0985	0,0492

Analýza úspěšnosti identifikace se měla zaměřit na úspěšnost při výpočtech denzitních vektorů pro samostatné nukleotidy a pro sumy purinových/pyrimidinových nukleotidů. Z toho hlediska se tedy nejlépe při identifikaci druhu osvědčilo použití samotného adeninu, při identifikaci rodu využití sum purinových/pyrimidinových nukleotidů - ty při identifikaci druhu vyšly jako druhé nejméně výhodné. Analýza se shoduje v tom, že nejnižší úspěšnost identifikace nastala, jak pro rod i druh, při použití samotného guaninu.

ZÁVĚR

Cílem této bakalářské práce byla programová realizace výpočtu nukleotidových denzitních vektorů a metody identifikační analýzy pomocí porovnávání nukleotidových denzit referenčního souboru sekvencí s nezávislým souborem sekvencí stejných druhů a následné vyhodnocení úspěšnosti identifikace při použití samostatných denzitních vektorů pro jednotlivé nukleotidy a sumy denzit nukleotidů stejných biochemických vlastností.

Programová realizace práce byla provedena v prostředí Matlab, kde byly zrealizovány jednotlivé funkce pro výpočet nukleotidových denzitních vektorů s proměnnou délkou výpočetního okna a s výběrem ošetření okrajových hodnot vektorů. Následně byla vytvořena uživatelská GUI aplikace, pomocí které proběhla identifikační analýza organismů na základě porovnávání nukleotidových denzit.

Z databáze BOLD byly vybrány čtyři referenční a pět analyzovaných FASTA souborů. Pro reference se jednalo o ryby z portugalského pobřeží (188 sekvencí, 48 druhů), chrostíky z Churchillu (716 sekvencí, 48 druhů), neotropické ptáky (758 sekvencí, 43 druhů) a netopýry z Guyany (840 sekvencí, 94 druhů). Analyzované soubory byly nezávislé soubory stejných druhů – ryby z portugalského jižního a západního pobřeží (200 a 207 sekvencí, 55 a 49 druhů), chrostíci z Churchillu (zkrácené z 540 na 71 sekvencí a z 24 na 20 druhů), neotropičtí ptáci (zkrácené z 637 na 145 sekvencí a z 432 na 145 druhů) a netopýři z Guyany (zkrácené z 4132 na 267 a z 69 na 61 druhů).

Druhové denzitní reference se tvořily tak, že se sekvence z FASTA souborů, určených pro tvorbu referencí, převedly na nukleotidové denzitní vektory a následně zprůměrovaly pro každý druh. To pro délky výpočetního okna 3 – 19 bp. Přiřazování referenčních druhů k analyzovaným bylo uskutečněno na základě nejmenší euklidovské vzdálenosti mezi referenčními a analyzovanými sekvencemi. Výpočet těchto euklidovských vzdáleností proběhl pro samostatné nukleotidy a i jejich sumy dle biochemických vlastností, opět pro délku výpočetního okna opět 3 – 19 bp. Tato analýza byla uskutečněna pro rody i druhy. Z výsledných hodnot byly vytvořeny grafy vážených průměrů úspěšností identifikace pro druh i rod.

Výsledky identifikační analýzy mohou být ovlivněny tím, že na testování byly použity pouze určité skupiny organismů, které jsou zastoupeny malým počtem druhů. V souborech je jen málo sekvencí k jednotlivým druhům a navíc zastoupené druhy pochází z malého množství rodů a čeledí. Nelze z toho tedy vyvozovat obecné závěry, že identifikace bude stejně dobře fungovat i pro jiné skupiny organismů.

Z verifikační identifikační analýzy vyplynulo, že nejúspěšnější pro identifikaci druhu je použití samostatného adeninu, a že identifikace rodu měla největší úspěch při volbě sum purinových/pyrimidinových nukleotidů. Naopak využití pouze guaninu vyšlo v obou případech jako nejméně vhodná volba. Velikost výpočetního okna neměla buď žádný, nebo malý vliv – s rostoucím výpočetním oknem se procento úspěšné identifikace mírně snižovalo. Nejvyšší hodnoty úspěšností vážených průměrů v rámci všech souborů jsou pro druh 99,91% a pro rod 97,8%.

LITERATURA

[1] DOLEŽALOVÁ, Ivana. *Druh jako základní článek evoluce rostlin, speciace* [online prezentace]. [cit. 2013-10-30]. Dostupné z: http://www.old.botany.upol.cz/prezentace/dolezal/4.pdf

[2] FLEGR, Jaroslav. *Evoluční biologie*. 2., opr. a rozš. vyd. Praha: Academia, 2009, 569 s. ISBN 978-80-200-1767-3.

[3] NOVÁK, Vladimír J. *Historický vývoj organismů: fylogenese mikroorganismů, rostlin a živočichů*. 1. vyd. Praha: Academia, 1969, 835 s.

[4] JELÍNEK, Jan a Vladimír ZICHÁČEK. *Biologie pro gymnázia: (teoretická a praktická část)*. 9. vyd. Olomouc: Nakladatelství Olomouc, 2007, 575 s., [92] s. barev. obr. příl. ISBN 978-80-7182-213-4.

[5] FLEGR, Jaroslav, Šárka KAŇKOVÁ, Jitka LINDOVÁ a Petr SYNEK. *Základy evoluční biologie pro gymnázia: projekt JPD3 - Přírodovědná gramotnost*. Vyd. 1. Praha: Univerzita Karlova v Praze, Přírodovědecká fakulta, 2008, 94 s. ISBN 978-80-86561-68-4.

[6] CAMPBELL, Neil A a Jane B REECE. *Biologie*. Vyd. 1. Brno: Computer Press, 2006, xxxiv, 1332 s. ISBN 80-251-1178-4.

[7] FLEGR, Jaroslav. *Úvod do evoluční biologie*. Vyd. 1. Praha: Academia, 2007, 544 s. Galileo, sv. 13. ISBN 978-802-0015-396.

[8] ROSYPAL, Stanislav. *Úvod do molekulární biologie*. 3. inovované vyd. Brno: Stanislav Rosypal, 2000, 604-900 s. ISBN 80-902562-2-8.

[9] Mutace. *Genetika - Biologie* [online]. ©2010-2013 [cit. 2013-11-29]. Dostupné z: http://www.genetika-biologie.cz/mutace

[10] ROSYPAL, Stanislav. *Nový přehled biologie*. 1. vyd. Praha: Scientia, 2003, xxii, 797 s. ISBN 80-7183-268-5.

[11] DROZD, Pavel. *Principy systematiky a taxonomie*. Vyd. 1. Ostrava: Ostravská univerzita, 2004, 89 s. ISBN 80-7042-995-X.

[12] HAMPL, Vladimír. Molekulární taxonomie: Úvod, taxonomie, molekulární znaky, sekvenace DNA. *Evoluntionary protistology group.* [Online]. [cit. 2013-10-30]. Dostupné z: http://www.protistologie.cz/files/MolTax/Molekularni%20taxonomie1-text.pdf

[13] ALBERTS, Bruce. *Základy buněčné biologie: úvod do molekulární biologie buňky*. 2. vyd. Ústí nad Labem: Espero Publishing, 2004, xxvi, 630 s. ISBN 80-

902906-0-4.

[14] SCHEFFLER, Immo E. Mitochondria. 2nd ed., Wiley-Blackwell, 2007, 472 s. ISBN 978-0-470-04073-7.

[15] 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.

[16] HEBERT, P. D. N., A. CYWINSKA, S. L. BALL a J. R. DEWAARD. Biological identifications through DNA barcodes: taxonomy, species delimitation and DNA barcoding. *Proceedings of the Royal Society B: Biological Sciences*. 2003-02-07, vol. 270, issue 1512, s. 313-321. DOI: 10.1098/rspb.2002.2218. Dostupné z: http://rspb.royal societypublishing.org/cgi /doi/10.1098/rspb.2002.2218

[17] TAYLOR, H. R. a W. E. HARRIS. An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources*. 2012, vol. 12, issue 3, s. 377-388. DOI: 10.1111/j.1755-0998.2012.03119.x. Dostupné z: http://doi.wiley.com/10.1111/j.1755-0998.2012.03119.x

[18] DESALLE, R., Mary G. EGAN a Mark SIDDALL. The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2005-10-29, vol. 360, issue 1462. DOI: 10.1098/rstb.2005.1722.

[19] ZÁVESKÁ DRÁBKOVÁ. Čárový kód života: Utopie, či realita? *Vesmír*. únor 2012, č. 91, s. 4.

[20] What is DNA barcoding?: How DNA Barcoding Works and What it Will Do. *Barcode of life* [online]. © 2013 [cit. 2013-11-29]. Dostupné z: http://ibol.org/ about-us/chat-is-dna-barcoding/

[21] BERGMANN, T., H. HADRYS, G. BREVES a B. SCHIERWATER. Character-based DNA barcoding: a superior tool for species classification. *Berliner Und Munchener Tierarztliche Wochenschrift*. 2009, č. 122, s. 446-450. DOI: 10.2376/0005-9366-122-446.

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

[23] What is DNA Barcoding?. *Barcode of Life: Identifying Species with DNA Barcoding* [online]. © 2010–2013 [cit. 2013-11-29]. Dostupné z: http://www.barcodeoflife.org /content/about/what-dna-barcoding

SEZNAM ZKRATEK

A	adenin
ATP	adenosintrifosfát
BCBNC	FASTA soubor netopýrů z Guyany
BOLD	Barcode od Life Database
BCDR	FASTA soubor netopýrů z Guyany
bp	base pair
BRAS	FASTA soubor neotopických ptáků
C	cytozin
CAOS	Characteristics Attributes Organization Systém
cox1	cytochrom c oxydáza podjednotka 1
CUCAD	FASTA soubor chrostíků z Churchillu
DSTRI	FASTA soubor chrostíků z Churchillu
FCFP	FASTA soubor ryb z portugalského pobřeží
FCFPS	FASTA soubor ryb z portugalského jižního pobřeží
FCFPW	FASTA soubor ryb z portugalského západního pobřeží
G	guanin
MNCM	FASTA soubor neotropických ptáků
MOTU	Molecular Operational Taxonomis Unit
mRNA	mediánová RNA
mtDNA	mitochondriální DNA
RNA	ribonukleová kyselina
rRNA	ribozomální RNA
T	thymin
tRNA	transferová RNA
UPGMA	Unweighted Pair Group Method with Arithmetic Mean

SEZNAM PŘÍLOH

PŘÍLOHA 1: Tabulky úspěšností identifikace druhu

Tab. č. 10: Tabulka úspěšnosti identifikační analýzy druhu pro soubor FCFPS.

Nukleotid pro výpočet	A	C	G	T	AC	AG	CG	ACGT
Délka okna	Úspěšnost správného přiřazení druhů [%]							
3	100	98,54	100	99,27	98,54	100	100	100
5	100	98,54	100	98,54	98,54	100	98,54	100
7	98,54	98,54	100	98,54	98,54	98,54	98,54	98,54
9	98,54	98,54	100	98,54	98,54	98,54	98,54	98,54
11	98,54	98,54	100	98,54	98,54	98,54	98,54	98,54
13	98,54	98,54	99,27	98,54	98,54	98,54	98,54	98,54
15	98,54	98,54	99,27	98,54	98,54	98,54	98,54	98,54
17	98,54	98,54	98,54	98,54	98,54	98,54	98,54	98,54
19	98,54	98,54	98,54	98,54	98,54	98,54	98,54	98,54

Tab. č. 11: Tabulka úspěšnosti identifikační analýzy druhu pro soubor FCFPW.

Nukleotid pro výpočet	A	C	G	T	AC	AG	CG	ACGT
Délka okna	Úspěšnost správného přiřazení druhů [%]							
3	100	100	100	100	100	100	100	100
5	100	100	100	100	100	100	100	100
7	100	100	100	100	100	100	100	100
9	100	100	100	100	100	100	100	100
11	100	100	100	100	100	100	100	100
13	100	100	100	100	100	100	100	100
15	100	100	100	100	100	100	100	100
17	100	100	100	100	100	100	100	100
19	100	100	100	100	100	100	100	100

Tab. č. 12: Tabulka úspěšnosti identifikační analýzy druhu pro soubor BCDR.

Nukleotidy pro výpočet	**A**	**C**	**G**	**T**	**AC**	**AG**	**CG**	**ACGT**
Délka okna	**Úspěšnost správného přiřazení druhů [%]**							
3	100	100	97,24	100	100	90,94	100	100
5	100	100	96,85	100	100	90,94	100	100
7	100	100	96,06	100	100	90,55	100	100
9	100	100	96,06	100	100	90,55	100	100
11	100	100	96,06	100	100	90,55	100	100
13	100	100	93,70	100	100	90,16	100	100
15	100	100	93,31	100	100	90,16	100	100
17	100	100	93,31	100	100	90,16	100	100
19	100	100	94,09	100	100	90,16	100	100

Tab. č. 13: Tabulka úspěšnosti identifikační analýzy druhu pro soubor DSTRI.

Nukleotidy pro výpočet	**A**	**C**	**G**	**T**	**AC**	**AG**	**CG**	**ACGT**
Délka okna	**Úspěšnost správného přiřazení druhů [%]**							
3	97,5	97,5	97,5	97,5	97,5	97,5	97,5	97,5
5	97,5	97,5	97,5	97,5	97,5	97,5	97,5	97,5
7	97,5	97,5	95	97,5	97,5	97,5	97,5	97,5
9	97,5	97,5	95	97,5	97,5	97,5	97,5	97,5
11	97,5	97,5	95	97,5	97,5	97,5	97,5	97,5
13	97,5	97,5	95	97,5	97,5	97,5	97,5	97,5
15	97,5	97,5	95	97,5	97,5	97,5	97,5	97,5
17	97,5	97,5	95	95	97,5	97,5	97,5	97,5
19	97,5	97,5	95	95	97,5	97,5	97,5	97,5

Tab. č. 14: Tabulka úspěšnosti identifikační analýzy druhu pro soubor BRAS.

Nukleotidy pro výpočet	**A**	**C**	**G**	**T**	**AC**	**AG**	**CG**	**ACGT**
Délka okna	**Úspěšnost správného přiřazení druhů [%]**							
3	100	97,65	80	96,47	98,82	91,76	97,65	98,82
5	100	97,65	76,47	96,47	98,82	91,76	98,82	100
7	100	96,47	71,76	97,65	98,82	91,76	98,82	100
9	100	96,47	69,41	97,65	98,82	91,76	98,82	98,82
11	100	96,47	68,24	97,65	98,82	89,41	98,82	98,82
13	98,82	96,47	64,71	96,47	97,65	88,24	98,82	98,82
15	98,82	96,47	65,88	96,47	97,65	88,24	98,82	97,65
17	98,82	96,47	64,71	96,47	97,65	88,24	97,65	97,65
19	98,82	97,65	61,18	96,47	97,65	88,24	97,65	97,65

PŘÍLOHA 2: Tabulky úspěšností identifikace rodu

Tab. č. 15: Tabulka úspěšnosti identifikační analýzy rodu pro soubor FCFPS.

Nukleotidy pro výpočet	A	C	G	T	AC	AG	CG	ACGT
Délka okna	Úspěšnost správného přiřazení rodů [%]							
3	100	98,72	94,23	99,36	98,72	100	100	100
5	100	98,72	94,23	98,72	94,87	100	98,72	100
7	98,72	98,72	94,23	97,44	93,59	98,72	96,15	98,72
9	94,23	98,72	93,59	96,79	92,95	98,72	96,15	98,72
11	95,51	98,72	92,95	96,79	92,95	98,72	97,44	98,72
13	96,15	98,72	92,31	93,59	91,67	98,72	97,44	98,72
15	98,72	98,08	92,31	92,95	91,67	97,44	97,44	98,72
17	98,08	98,08	91,67	92,95	91,67	97,44	96,79	98,72
19	98,08	98,08	91,67	92,95	91,67	96,79	95,51	98,72

Tab. č. 16: Tabulka úspěšnosti identifikační analýzy rodu pro soubor FCFPW.

Nukleotidy pro výpočet	A	C	G	T	AC	AG	CG	ACGT
Délka okna	Úspěšnost správného přiřazení rodů [%]							
3	100	94,87	96,15	91,03	95,51	100	94,87	100
5	100	94,23	96,15	87,18	91,67	100	92,31	100
7	100	93,59	96,15	86,54	91,67	100	92,31	100
9	100	93,59	96,15	86,54	94,87	100	92,31	100
11	100	93,59	96,15	86,54	96,79	100	92,31	100
13	100	93,59	96,15	86,54	96,79	100	92,31	100
15	100	93,59	100	86,54	96,79	100	96,15	100
17	100	93,59	100	86,54	96,15	100	96,15	100
19	100	93,59	100	86,54	94,87	100	96,15	100

Tab. č. 17: Tabulka úspěšnosti identifikační analýzy rodu pro soubor BCDR.

Nukleotidy pro výpočet	A	C	G	T	AC	AG	CG	ACGT
Délka okna	Úspěšnost správného přiřazení rodů [%]							
3	99,25	97,75	94,38	99,63	97,75	99,63	97,38	99,63
5	97,38	97,75	94,01	97,75	97,75	99,25	97,38	97,75
7	97,38	97,75	93,26	97,75	97,38	99,25	97,38	97,75
9	97,38	97,38	93,26	97,38	97,38	99,25	97,38	97,75
11	97,38	97,38	93,63	97,38	97,38	99,25	97,38	97,75
13	97,38	97,38	91,39	97,38	97,75	99,25	97,38	97,75
15	97,38	97,38	91,01	97,38	97,75	99,25	97,38	97,75
17	97,38	97,38	91,01	97,38	97,75	99,25	97,38	97,38
19	97,38	97,38	91,76	97,38	97,75	99,25	97,38	97,38

Tab. č. 18: Tabulka úspěšnosti identifikační analýzy rodu pro soubor DSTRI.

Nukleotidy pro výpočet	A	C	G	T	AC	AG	CG	ACGT
Délka okna	Úspěšnost správného přiřazení rodů [%]							
3	91,55	91,55	81,69	100	98,59	91,55	91,55	98,59
5	91,55	100	80,28	100	91,55	90,14	91,55	91,55
7	91,55	100	80,28	94,37	98,59	97,18	91,54	91,55
9	91,55	100	80,28	91,55	97,18	97,18	91,55	91,55
11	91,55	100	74,65	94,37	97,18	97,18	91,55	91,55
13	91,55	97,18	73,24	92,96	91,55	97,18	90,14	94,37
15	91,55	97,18	73,24	92,96	88,73	98,59	97,18	94,37
17	98,59	97,18	73,24	97,18	84,51	98,59	90,14	98,59
19	98,59	91,55	74,65	90,14	85,92	98,59	90,14	91,55

Tab. č. 19: Tabulka úspěšnosti identifikační analýzy rodu pro soubor BRAS.

Nukleotidy pro výpočet	**A**	**C**	**G**	**T**	**AC**	**AG**	**CG**	**ACGT**
Délka okna	**Úspěšnost správného přiřazení rodů [%]**							
3	93,10	85,52	65,52	82,76	80,69	90,34	82,07	88,28
5	91,72	84,83	61,38	82,07	80	90,34	82,76	88,28
7	91,72	83,45	60	80,69	79,31	90,34	82,76	86,89
9	91,72	83,45	59,31	80	78,62	91,03	82,07	85,52
11	89,66	84,14	58,62	80	78,62	89,66	83,45	86,9
13	88,28	84,14	55,86	80	77,24	88,97	82,76	86,9
15	87,59	84,83	55,86	79,31	76,55	88,28	82,76	86,9
17	87,59	83,45	55,17	79,31	77,24	87,59	82,76	87,59
19	87,59	83,45	53,79	77,24	76,55	87,59	81,38	86,21