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

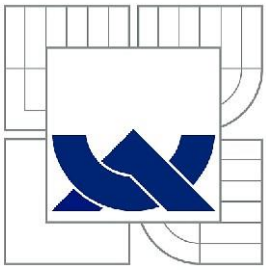
ZNAKOVĚ-ORIENTOVANÉ METODY DNA BARCODINGU

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

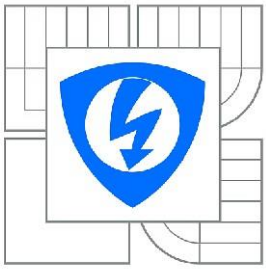
Bc. KATEŘINA LESÁKOVÁ

BRNO 2015



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

ZNAKOVĚ-ORIENTOVANÉ METODY DNA BARCODINGU

CHARACTER-BASED METHODS FOR DNA BARCODING

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

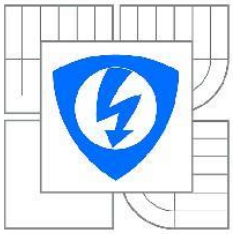
AUTOR PRÁCE — Bc. KATEŘINA LESÁKOVÁ

AUTHOR

VEDOUCÍ PRÁCE — Ing. DENISA MADĚRÁNKOVÁ

SUPERVISOR

BRNO 2015



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Diplomová práce

magisterský navazující studijní obor
Biomedicínské inženýrství a bioinformatika

Studentka: Bc. Kateřina Lesáková ***ID:*** 138944
Ročník: 2 ***Akademický rok:*** 2014/2015

NÁZEV TÉMATU:

Znakově-orientované metody DNA barcodingu

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši na téma znakově-orientované metody DNA barcodingu. 2) Na vhodně zvoleném souboru dat vyzkoušejte alespoň dvě z volně dostupných znakových metod pro identifikaci organismů, např. metodu CAOS a BLOG. 3) Výsledky identifikace mezi metodami porovnejte a diskutujte. 4) Navrhněte vylepšení jedné ze znakově-orientovaných metod či vlastní metodu a toto implementujte v libovolném programovém prostředí. 5) Pomocí navržené metody proveďte identifikační analýzu různých typů organismů a výsledky porovnejte s dříve testovanými metodami.

DOPORUČENÁ LITERATURA:

[1] SARKAR, I. N.; PLANET, P. J. a DESALLE, R. CAOS software for use in character-based DNA barcoding. Molecular Ecology Resources. 2008, 8, 1256-1259.
[2] WEITCHEK, E.; VAN VELZEN, R.; FELICI, G. a BERTOLAZZI, P. BLOG 2.0: a software system for character-based species classification with DNA Barcode sequences. What it does, how to use it. Molecular Ecology Resources. 2013, 13, 1043-1046.

Termín zadání: 9.2.2015 ***Termín odevzdání:*** 22.5.2015

Vedoucí práce: Ing. Denisa Maděránková
Konzultanti diplomové práce:

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

UPOZORNĚNÍ:

ABSTRAKT

Diplomová práce se zabývá studiem znakově orientovaných metod DNA barcodingu. Úvod obsahuje informace o DNA barcodingu a znakově orientovaných metodách DNA barcodingu. V teoretické části je popsána metoda CAOS a metoda BLOG. V praktické části jsou popsány oba programy pro znakově orientované metody k analýze genomických sekvencí. V praktické části je rovněž popsána teorie a realizace vlastní metody. V závěru jsou shrnuty výsledky analýzy.

KLÍČOVÁ SLOVA

DNA barcoding, znakově orientované metody DNA barcodingu, fasta, nexus, COAS, BLOG

ABSTRACT

This work deals with character-based DNA barcoding. DNA barcoding and character-based DNA barcoding methods are described in the introduction. Another part contains information of method CAOS (Characteristic Attributes Organization) and method BLOG (Barcoding with LOGic). Programs are described in the practical part. The end contains results.

KEYWORDS

DNA barcoding, character-based DNA barcoding methods, fasta, nexus, CAOS, BLOG

Bibliografická citace mé práce:

LESÁKOVÁ, K. *Znakově-orientované metody DNA barcodingu.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2015. 50 s. Vedoucí diplomové práce Ing. Denisa Maděránková.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Znakově-orientované metody DNA barcodingu“ jsem vypracovala samostatně pod vedením vedoucí diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

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

Brno . . . .

(podpis autora)

PODĚKOVÁNÍ

Děkuji své vedoucí diplomové práci Ing. Denise Maděránkové za příkladné vedení, cenné rady a trpělivost při vzniku této práce.

V Brně dne . . . .

(podpis autora)

Obsah

Seznam Tabulek

1 Úvod

Tato diplomová práce se zabývá studiem znakově orientovaných metod DNA barcodingu. Práce krátce popisuje DNA Barcoding a jeho znakově orientované metody. Blíže je zde popsán princip dvou v současnosti dostupných metod, které se danou problematikou zabývají. Jedná se o metodu CAOS (Characteristic Attributes Organization) a metodu BLOG (Barcoding with LOGic). U obou metod je rovněž v praktické části diplomové práce popsáno uživatelské rozhraní dostupných programů a návod na manipulace s nimi.

Podstatou této diplomové práce je především osvojit si obě metody po teoretické stránce. A jejich následné testování pomocí vhodně zvoleného souboru dat, kdy budou popsány výsledky analýzy. Cílem je poté navrhnout vlastní znakově orientovanou metodu pro analýzu DNA barcodových sekvencí.

Výsledky identifikace organismů pomocí obou zmíněných metod budou porovnány a případné klady a zápory diskutovány v průběhu práce. Na základě získaných informací o obou metodách bude jedna z uvedených metod vyhodnocena jako vhodnější pro klasifikaci pomocí znakově orientovaného DNA barcodingu.

Na základě této skutečnosti budou základní principy vybrané metody realizovány v praktické části diplomové práci, a to v libovolném programovém prostředí. Po vytvoření této metody bude následně provedena analýza na testovaném souboru dat a výsledky analýzy nudou porovnány s vybranou metodou v této diplomové práci.

1.1 DNA barcoding

Barcode v překladu z anglického jazyka znamená čárový kód, což nám může napovědět, jak by tato metoda mohla vypadat. O DNA barcodingu se mohla vědecká komunita zaměřená na tuto oblast poprvé dozvědět již v roce 2003, kdy výzkumná skupiny Paula Heberta z kanadské univerzity Guelph v Ontariu publikovala článek s názvem "Biologické identifikace prostřednictvím DNA barcodingu". Ten spočívá převážně v tom, že navrhuje nový systém určování živočišných a rostlinných druhů včetně objevování druhů nových organismů pomocí krátkého úseku DNA ze standardizované oblasti genomu. Tyto sekvence DNA mohou být použity k identifikaci různých druhů, podobně jako skener v supermarketech používá známé černé pruhy čárového kódu k identifikaci vašeho nákupu.

DNA barcoding může sloužit k dvojímu účelu, a to jako nový nástroj v taxonomickém toolboxu rozšiřující jeho znalosti a zároveň jako inovativní zařízení pro laiky, kteří potřebují rychlou identifikaci neznámých druhů.

Úsek genu, který je používán pro téměř všechny živočišné skupiny, je úsek dlouhý 648 párů bází pocházející z buněčných mitochondrií a nachází se právě v mitochondriálním cytochromu c oxidasy genu 1. V praxi je označován jako "CO1". Tento úsek se ukázal jako velmi účinný při určování zástupců ptáků, motýlů, ryb, much a mnoha dalších zvířecích skupin. Výhodou použití CO1 je, že je tato sekvence dost krátká na to, aby mohla být rychle a snadno sekvenována a zároveň dost dlouhá na to, aby mohla sloužit k identifikaci variací mezi druhy. [1]

1.2 Znakově orientované metody DNA barcodingu

Na následujících stránkách diplomové práce budou detailně popsány dvě nedávno zveřejněné metody DNA barcodingu, které jsou založené na hledání znaků v analyzovaných sekvencích.

Jedná se konkrétně o již zmíněné metody CAOS (Characteristic Attributes Organization) a BLOG (Barcoding with LOGic). Pomocí nalezených společných znaků v sekvencích u příbuzných zástupců, jsme ve většině případů schopni analyzovat neznámé sekvence a přiřadit je k již známé skupině živočichů, rostlin či hub.

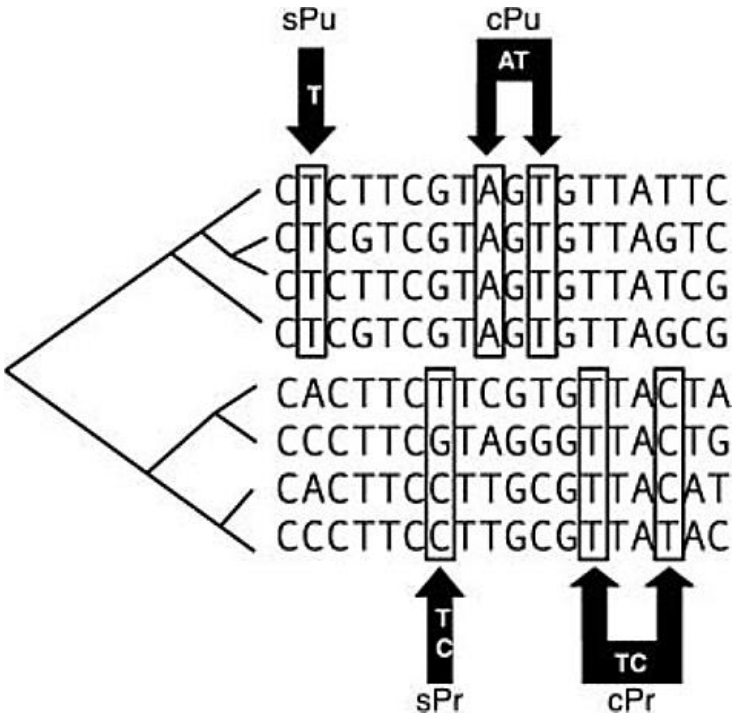
2 Teoretická část

2.1 Metoda CAOS

2.1.1 Princip

Metoda CAOS (Characteristic Attributes Organization), volně přeloženo jako systém organizace založený na vyhledávání charakteristických atributů v sekvencích, byl zveřejněn v roce 2008. Jedná se o automatizovanou a systematickou metodu, která nalézá neměnné charakteristické stavy ve fylogenetických stromech nebo ve skupinách kategorických informací.

CAOS definuje tzv. atribut testy v místě každého uzlu ve fylogenetickém stromu. Jedná se vlastně o znakové stavy v sekvencích, zde je nazýváme jako charakteristické atributy. Na rozdíl od rozhodovacích algoritmů při tvorbě fylogenetických stromů, systém CAOS nebere v potaz všechny atributy, ale pouze ty diagnosticky informativní, tedy významné charakteristické atributy. [2]



Obrázek 1: CAOS Systém

2.1.2 Popis atributů CAOS

Na obrázku č. 1 si můžeme názorně ukázat s jakými atributy metoda CAOS pracuje a co si pod nimi můžeme představit. Jako první zde tedy můžeme vidět atribut T, který se řadí do skupiny atributů Simple Pure CAs (sPu), neboť ji tvoří jediná báze T (thymin). Tato báze je přítomna na všech pozicích v první skupině znaků, ale již se nevyskytuje v druhé, alternativní skupině.

Další atribut definován jako AT je označen jako Compound Pure CAs (cPu), a to z důvodu, že je víceznakový, tedy že obsahuje více znaků než jeden. Tento atribut existuje napříč celou skupinou, ale tato kombinace není zastoupena odpovídajícími pozicemi ve druhé, alternativní skupině sekvencí.

V pořadí první atribut s označením TC je Simple Private CAs, neboť báze T a C jsou přítomny u některých členů druhé skupiny, ale v alternativní, první skupině je již nenalezneme. Poslední atribut TC je Compound Private CAs, je tvořen více znaky a kombinaci bází TC nenajdeme v alternativní, první skupině. [2]

2.1.3 Popis postupu analýzy CAOS

Na obrázku uvedeném pod tímto odstavcem si můžeme graficky představit v čem systém CAOS spočívá.

Jedná se o identifikaci jednotlivých typů atributů (CAs). Základními typy atributů jsou atributy Pure (čistý) a atributy Private (uzavřený). Pure atributy (PU) jsou tzv. znakové stavy, které existují ve všech složkách (napříč všemi členy) dané skupiny, ale nejsou členy alternativní skupiny. Private atributy (Pr) jsou přítomny pouze u některých členů skupiny, ale chybí v alternativní skupině.

CAOS nejprve identifikuje všechny jednotky CAs, které jsou složené z jednoho znakového stavu. Mluvíme zde o Simple (jednoduché) CAs (**sCAs**), jsou to atributy na jednotlivých pozicích ve fylogenetickém stromu. Opět zde rozlišujeme dva typy atributů: Simple Pure CAs (**sPus)** a Simple Private CAs (**sPrs).**

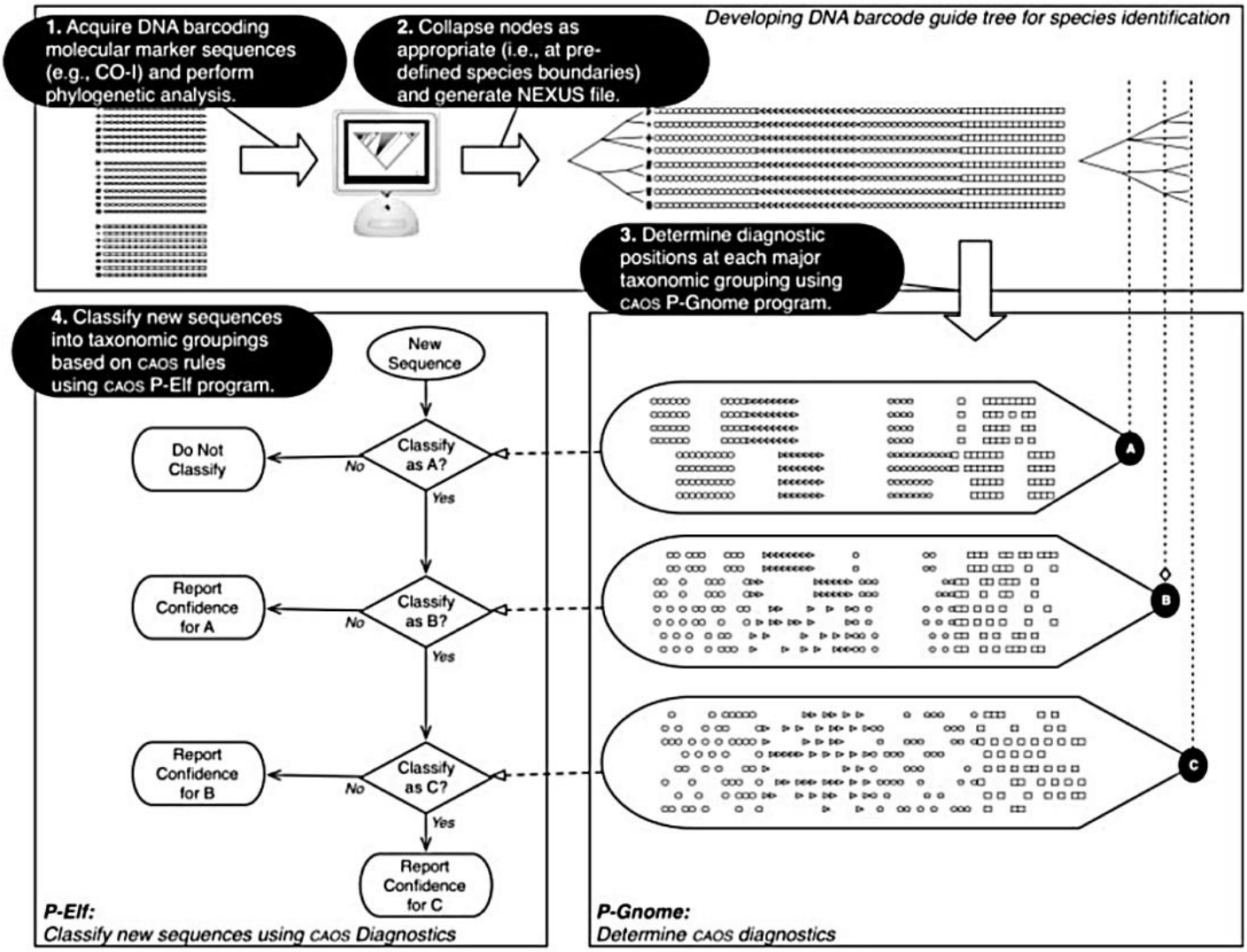
CAOS poté vyhledává víceznakové stavy s názvem Compound CAs (cCAs). Atribut cCAs není nikdy složen z žádných jednotek CAs. Opět existují Compound Pure atributy a Compound Private atributy, které jsou označeny písmeny **cPu** a **cPr.**

Takto nalezené atributy CAs jsou organizovány do indukční sady pravidel, která může být použita ke klasifikaci nové sekvence (nového taxonu) do již vytvořeného fylogenetického stromu, odkud byly jednotlivé atributy předem získány. [2]

2.1.4 Klasifikace

CAOS klasifikuje nový taxon pomocí sady pravidel zastoupených atributy (CAs), která jsou aplikována v předem definovaném pořadí kvůli správné identifikaci. Různé kombinace jednotek CAs mohou být použity pro klasifikaci různých taxonů na stejném uzlu ve stromu. Zároveň je zde možno provést klasifikaci v každém uzlu fylogenetického stromu.

Zatímco topologie původního fylogenetického stromu je použita jako vodítko pro rozhodovací strom, rozhodování v každém uzlu je stanoveno za použití dílčího shlukovacího algoritmu. Přehled celého CAOS procesu je znázorněn na obr. 2.



Obrázek 2: Přehled schématu metody CAOS

Na následujících řádcích jsou popsány kroky programu, které jsou znázorněny na obrázku výše. CAOS systém je realizován jako proces o čtyřech krocích.

Prvním je získání DNA sekvencí a provedení fylogenetické analýzy, probíhá tedy tvorba fylogenetického stromu (1). Následuje tzv. zhroucení uzlů ve stromě podle potřeby, tedy v předem určených druhových hranicích a vytvoření datového souboru ve formátu NEXUS (2). Poté probíhá stanovení jednotlivých diagnostických pozic, zmiňovaných atributů, v každé z hlavních taxonomických skupin pomocí programu CAOS P-Gnome (3). Posledním krokem je klasifikace nových sekvencí do taxonomických seskupení pomocí programu CAOS P-Elf (4). [2]

2.1.5 Popis programu

CAOS systém se skládá ze dvou samostatně realizovaných programů s názvy P-Gnome a P-Elf. P-Gnome slouží jako generátor diagnostických pravidel, který prohledává dané matice dat a stanoví zmíněné diagnostické sady pravidel pro každý z determinovaných subjektů v matici dat. Program P-elf pak může klasifikovat soubor dotazovacích sekvencí podle pravidel generovaných programem P-Gnome.

Program P-Gnome používá specifický formát souborů s názvem NEXUS, který se skládá z matice obsahující DNA data, převáděcího bloku, který převádí názvy taxonů na celočíselné hodnoty ve stromové reprezentaci a tzv. Newickova stromu, který je výsledkem zhroucení uzlů vzhledem k zájmovému taxonomickému seskupení (tj. determinovanými hranicemi druhů). Tento proces je zde značně usnadněn pomocí fylogenetických zkušebních balíčků jako Macclade (Maddison a Maddison 2005) nebo Mesquite (Maddison a Maddison 2007). [2]

Výstup programu P-Gnome je umístěn v několika souborech, z nichž soubory označené jako atributy a tzv. skupinové soubory jsou nejvíce relevantní pro stanovení diagnostiky. Soubor s názvem *skupiny* představuje vstupní strom a umístění všech koncových větví do vnořených skupin, které mohou být dále použity nebo vyřazeny jako hypotézy při procesu seskupování druhů. Soubor s názvem *atributy* uvádí diagnostiku a úroveň její spolehlivosti pro každou ze skupin stanovených v souboru *skupiny*.

Aby bylo možné lépe uspořádat výsledky analýzy CAOS, výstup z P-Gnome je navržen a formátován tak, že lze manipulovat v tabulkových aplikacích, jako je Microsoft Excel. Na základě souboru pravidel vytvořených pomocí P-Gnome, program P-Elf čte FASTA soubory a vrací soubor s nejlepší identifikací každého vstupu buď jednotlivě, nebo skupinově. Tento výstup je rovněž jednoduchý na manipulaci v tabulkách. [2]

2.2 Metoda BLOG

2.2.1 Princip metody BLOG

Metoda BLOG (Barcoding with LOGic) opět pracuje s Barcodovými DNA sekvencemi. I zde se jedná o nalézání pravidel pro klasifikaci sekvencí z hlediska umístění klíčových nukleotidů, které slouží k diagnostice analyzované sekvence. Zde se k získání DNA barcodové sekvence jako obvykle využívá gen cytochromu C oxidázy I (COI) z části mitochondriální DNA.

Princip klasifikace touto metodou můžeme popsat následujícím způsobem. Máme referenční knihovnu složenou s DNA barcodových sekvencí známých druhů a jednu neznámou DNA barcodovou sekvenci. Cílem metody je tuto sekvenci rozpoznat a přiřadit ke druhu, který se již nachází v referenční knihovně.

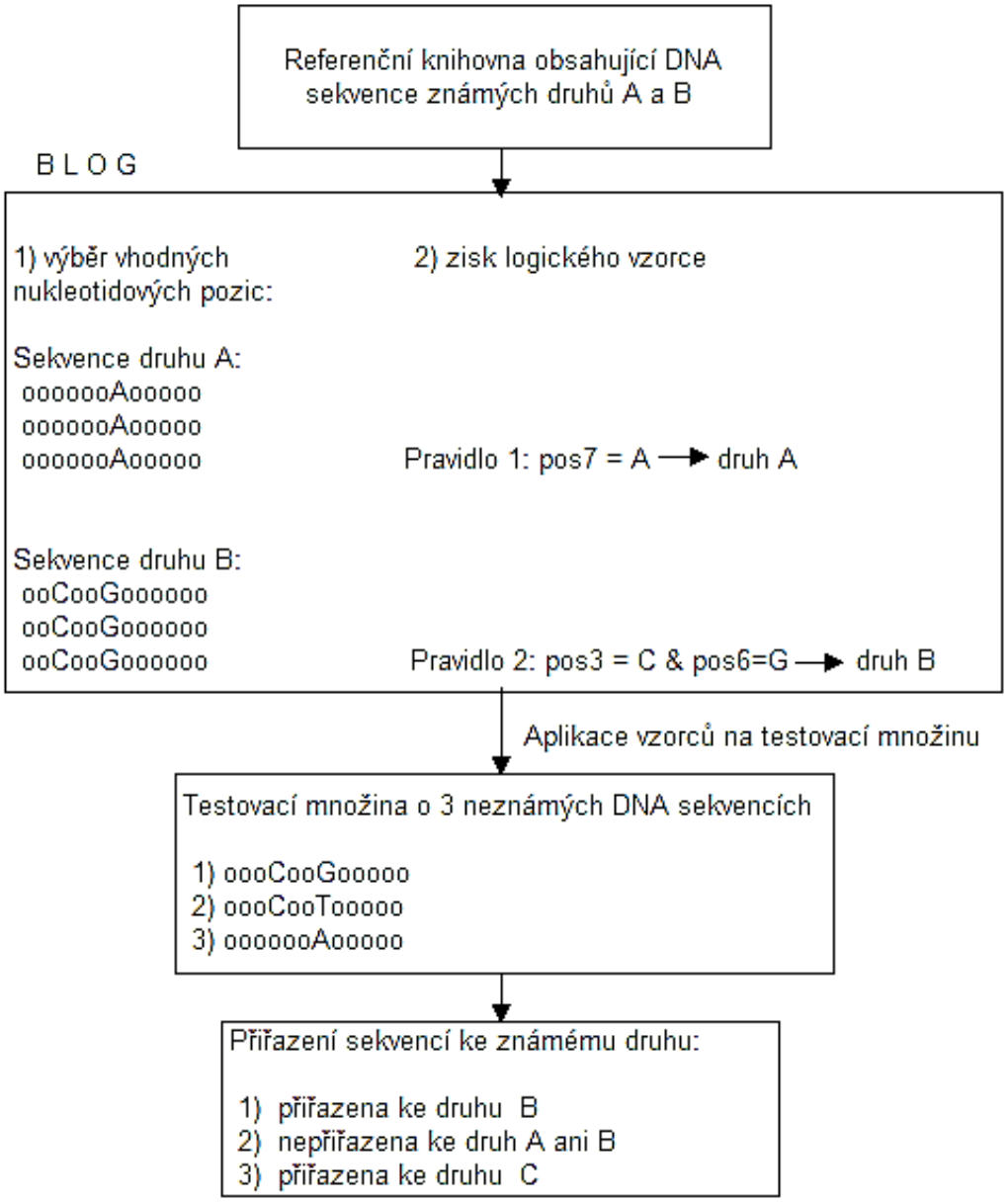
Metoda BLOG identifikuje pro každý druh v příruční referenční knihovně charakteristické pozice nukleotidů v DNA barcodové sekvenci a zároveň přiřazuje každému druhu logickou klasifikační formuli. Jedná se o taková malá pravidla ve tvaru „jestliže – pak“, která jsou schopna charakterizovat druhy kompaktním způsobem. Můžeme si zde uvést jednoduchý příklad takového pravidla BLOGU: " Pokud pos40 = T a pos265 = T, pak je vzorek klasifikován jako Ompok bimaculatus ". Tedy pokud se na pozici číslo 40 a 265 nachází báze thyminu, jedná se sekvenci příslušného druhu, zde Ompok bimaculatus.

Výrazná výhoda BLOGu ve srovnání s jinými dostupnými metodami je, že takové logické vzorce nabízejí další, přídavnou úroveň informace o jednotlivých druzích, která může být použita mimo rámec DNA barcodingu, například v druhovém popisu, v molekulární detekci nebo ve fylogenetické analýze. [3]

Metoda BLOG je založena na dvou hlavních výpočetních krocích. Zaprvé se provede tzv. volba znaku, tzn. že BLOG vybere malý soubor obsahující pozice DNA barcodových sekvencí, které jsou vhodné k rozlišování druhů v příruční referenční knihovně. Za druhé se provede tzv. získání vzorce, tzn. že BLOG vypočítá logické vzorce, které jsou poté schopny klasifikovat každý druh přítomný v příruční referenční knihovně.

Metoda BLOG používá přístup tzv. strojového učení s učitelem. Ten kdo pracuje s touto metodou, tedy uživatel, musí poskytnout tzv. trénovací množinu, která obsahuje vzorky DNA sekvencí s již známými členy jednotlivých druhů. Na základě této trénovací množiny software vybere vhodné nukleotidové pozice (tzv. znakový výběr) a vypočítá logické vzorce pro klasifikaci druhů (tzv. získání vzorce). Následně lze tyto logické vzorce aplikovat na testovací množinu DNA sekvencí, která obsahuje neznámé vzorky, které vyžadují klasifikaci. Testovací množina může obsahovat vzorky s neznámou příslušností ke druhu, nebo alternativně ty

vzorky, které mají a priori známou příslušnost k druhu, což nám umožňuje ověření správné klasifikace vzorků. [3]



Obrázek 3: Schéma metody BLOG

Metoda BLOG je převážně určena k identifikaci umístění klíčových diagnostických pozic nukleotidů u každého druhu zvlášť v plně definované trénovací množině. Pro získání spolehlivých výsledků musí tedy zkušební množina obsahovat pouze vzorky druhů, které jsou přítomny v trénovací množině. Stejně tak je nutná kompletní příruční knihovna, tzn. knihovna, ve které budou i polymorfismy přítomné u jednotlivých druhů v trénovací množině, aby se vyloučily falešně negativní výsledky.

2.2.2 Vstupy a výstupy programu BLOG

Vstupní soubory jsou DNA barcodové sekvence ve standardním FASTA formátu. Tedy soubor obsahující tzv. hlavičku s popisem sekvence, za kterou následuje samotná DNA sekvence. Trénovací sekvence musí pocházet z téhož regionu genu nebo musí být předběžně zarovnané do stejného regionu před zpracováním metodou BLOG.

Výstupem BLOGu jsou logické formule pro klasifikaci druhů, míra klasifikace a matice záměn. Logické formule, které přidělují vzorek ke druhu, jsou to zmíněná malá pravidla ve tvaru "jestliže - pak“. Míry klasifikace jsou stanoveny jako počet a procento správně, nesprávně klasifikovaných a neklasifikovaných vzorků. Matice záměn poskytují detailní informace o správnosti klasifikace a o případné křížové klasifikaci.

Hodnota na pozici i-j buňky této matice představuje počet vzorků jednotlivých druhů, kde se předpokládá zařazení do druhu příslušící sloupci j. Správně klasifikované prvky jsou na hlavní diagonále matice záměn. [3]

První výpočetní krok BLOGu je výběr druhově specifických pozic sekvencí DNA barcodingu z trénovací množiny. Přístup volby znaku BLOGu je založen na matematické optimalizační formulaci, kterou najdeme blíže popsanou v článku od Bertolazziho z roku 2010. BLOG přejímá účinný heuristický algoritmus založený na náhodném vyhledávání, který je schopen produkovat vysoce kvalitní řešení po omezenou dobu. (Proveditelné řešení je produkováno v lineárním čase v závislosti na velikosti problému).

Předchozí verze BLOGu aplikovala krok volby znaku na všechny druhy v referenční databázi současně. Nicméně znaky, které umožňují oddělení jednoho druhu, nemusí být užitečné pro separaci jiného. BLOG 2.0 může aplikovat krok volby znaku odděleně pro každý druh v příruční knihovně zvlášť. V každém kroku volby znaku, je uvažovaným druhům přiřazena třída A a všem ostatním druhům třída B. V důsledku toho je potřeba vyřešit problém volby znaku pro m případů v každém běhu analýzy, kde m je počet druhů v trénovací množině. Dlouhá doba výpočtu bude zapotřebí v případě použití přesných algoritmů. [3]

2.2.3 Tvorba logických vzorců

Cílem kroku, označovaném jako získání vzorce, je vytvořit logický vzorec nebo pravidlo, schopné oddělit každý druh. BLOG přijímá tzv. metodu Lsquare, kde se logické formule získají řešením posloupnosti dobře známého a zároveň těžkého logického optimalizačního problému v podobě Minimum Cost Satisfiability Problems (MinSat).

Každý literál vzorce, chápáno jako prvek nesoucí pevně danou hodnotu, představuje přiřazení nukleotidu (tj. Adeninu, Thyminu, Guaninu nebo Cytozinu) na konkrétní pozici v sekvenci DNA barcodu.

Předchozí verze BLOGu obvykle produkovaly vzorce s pozitivními i negativními literály, aby se minimalizovala velikost vzorců. Příklad negativního literálu je např. pos40 = NOT T, tedy že na pozici 40 v sekvenci se nevyskytuje báze Thymin. Nicméně, negativní literály rozpoznávají tři různé nukleotidy, což z nich dělá potenciálně méně přesné literály než literály pozitivní. Pro představu si můžeme uvést následující příklad. Vzorec Pos40 = G OR pos40 = C by byl přesnější vzorec než vzorec pos40 = není T. Proto verze BLOG 2.0 zavádí zvýšení nákladů na negativní literály v MinSat problémových vzorcích ku celkovému prospěchu převážně výstupních pozitivních literálů. [3]

Před vyhodnocením zkušební množiny provede BLOG 2.0 vyhodnocení trénovací množiny s cílem přiřadit relativní váhy všem logickými vzorcům podle algoritmu popsaného v článku od Weitschek z roku 2011 se využívá se tzv. Laplaceovo skóre, tedy falešně pozitivní a pravdivě pozitivní míry jsou vypočítány pro každý logický vzorec v referenční knihovně a tyto výsledky jsou pak zohledněny v testovací množině při provádění úkonů klasifikace.

Kromě slibných výsledků klasifikace je výraznou výhodou BLOGu výstup z modelu, který dává kompaktní a přesný popis druhů v příruční knihovně. BLOG navíc nabízí další informace o druzích v podobě logických klasifikačních vzorců, které mohou být také použity mimo rámec DNA barcodingu v popisu druhů nebo molekulární detekce. [3]

3 Praktická část

3.1 Popis pracovního souboru

V praktické části se budu zabývat samotným použitím popsaných metod CAOS a BLOG určeným k analýze DNA barcode sekvencí, a to pomocí již vytvořených programů, které jsou volně dostupné na webových stránkách, které se touto problematikou zabývají.

K testování jsem si vytvořila soubor ve formátu FASTA, který obsahuje více sekvencí, které byly postupně získávány z jednotlivých FASTA souborů a uloženy postupně do jednoho. FASTA formát obsahuje hlavičku, ve které jsou většinou informace typu ID sekvence, název organismu, umístění v genu atd. Tyto informace jsou odděleny svislými čarami. Dále je v souboru obsažena už jen samotná sekvence jako sled nukleotidových bází. Následuje ukázka FASTA souboru.

>LGSMC585-05|Adela caeruleella|COI-5P|GU088564
AACTTTATATTTTATTTTTGGAATTTGATCAGGATTACTAGGAACTTCATTAAGTTTATTAATTCGAA
CAGAATTAGGAATACCTGGATCTTTAATTGGAAATGATCAAATTTATAATACAATTGTCACAACTC
ATGCTTTTATTATAATTTTTTTTTATAGTAATACCAATTATAATTGGAGGATTTGGTAATTGATTAATT
CCTTTAATGCTTGGGGCTCCAGATATAGCTTTTCCTCGTCTTAATAATA

Pracovní soubor byl vytvořen z DNA sekvencí živočišného řádu *Lepidoptera*, česky motýli. Tyto sekvence jsem získala z databáze Barcode of Life Database (BOLD), která byla vytvořena a je nadále udržována pomocí University of Guelph v Ontariu. Tato databáze nabízí vědcům způsob, jak shromažďovat, spravovat a analyzovat data DNA barcodingu. BOLD slouží jako uložiště pro barcodové záznamy, kde je možno vyhledávat čárové kódy, ukládat vzorky dat a obrázky, stejně jako sekvence a stopové soubory. [1]

Celý soubor obsahuje 50 sekvencí, které přísluší dohromady čtyřem živočišným druhům, a to *Adela caeruleella, Aidos perfusa admiranda, Anthela ocellata a Atteva aurea*. Jedná se o krátké sekvence čítající okolo 700bp, pocházející z genu cytochromu C oxidázy I (COI) z části mitochondriální DNA. [4]

3.2 Metoda CAOS

3.2.1 Popis programu

Program COAS je umístěn na webové stránce http://bol.uvm.edu/caos-workbench/caos.php, kde jej můžeme používat on-line. Vstupem do programu je soubor ve formátu NEXUS, který byl popsán výše. K jeho vytvoření z FASTA souboru s více DNA sekvencemi, potřebujeme některé z následujících programů MacClade nebo Mesquite. Zde jsem pro naše potřeby použila program Mesquite (Maddison and Maddison, 2007). Jeho úvodní okno můžeme vidět v příloze č 1. [5]

Do programu vstupuje testovací soubor sekvencí ve formátu FASTA. Testované sekvence je nutné před použitím globálně zarovnat pomocí jakéhokoli programu pro práci s fylogenetickými daty. Zde byl pro zarovnání použit program ClustalX2.1. Při vkládání souboru do programu Mesquite pomocí *File → Open File* se soubor rovnou uloží do formátu NEXUS. Ovšem pro potřeby analýzy je nutno takto vytvoření soubor znovu otevřít v programu Mesquite a pomocí nástrojů pro Matice *(Matrix)* nastavit jako typ not interleave, kdy tento údaj bude odebrán ze souboru. [6]

Dalším krokem je provedení fylogenetické analýzy vstupních sekvencí, tedy vytvoření fylogenetického stromu. Fylogenetický strom můžeme vytvořit například pomocí programů PAUP nebo PHYLIP. Zde využívám nástrojů programu PAUP. Jedná se o zkratku programu Phylogenetic Analysis Using Parsimony, který využívá optimalizační metodu maximální věrohodnosti pro vícenásobné zarovnání sekvencí (Maximum Parsimony). Úvodní okno s již otevřeným NEXUS souborem můžeme vidět v příloze č. 2. Strom si vygenerujeme pod záložkou *Trees* pomocí příkazu *Generate trees* a uložíme opět jako NEXUS formát. [7]

Dalším krokem je upravení takto vytvořených dat, tedy souboru se sekvencemi a vytvořeného fylogenetického stromu, v programu Mesquite následovně. Otevřeme již vytvořený NEXUS soubor se sekvencemi, pomocí *Open File*, poté musíme přes záložky *Taxa&Trees → Import File with Trees → Include contents* otevřít soubor s vytvořeným stromem. Tento strom je zde potřeba otevřít příkazem *Show tree* a upravit všechny uzly tak, aby z nich vycházely pouze dvě větve a opět uložit pomocí *Tree → Store tree*. Strom můžeme upravit buď ručně nebo za pomocí záložky *Trees → Alter/Transform Tree → Resolve polytomies,* která úpravu provede za nás. Správnou podobu stromu můžeme vidět v příloze č.3. Ověřit si tuto informaci můžeme ve vlastnostech stromu, kdy nám program vypíše informaci „*no polynomies*“.

Nyní je vytvořen jeden konečný NEXUS soubor potřebný pro vstup do programu COAS. Vlastně zde došlo k úpravě souboru se sekvencemi pomocí souboru s fylogenetickým stromem, kam se přenesla informace o stavbě stromu. [6]

3.2.2 Postup analýzy

Na webových stránkách s programem CAOS je analýza rozdělena do čtyř kroků s názvy CAOS Analyzer, CAOS Barcoder, CAOS Classifier a CAOS Library. [5]

CAOS Analyzer

První z uvedených částí, které je nutno použít, je CAOS Analyzer. Vstupem do této části je konečný vytvořený soubor se DNA barcode sekvencemi s informací o stromové struktuře ve formátu NEXUS. Výstupem jsou čtyři soubory ve formátu TXT.

Jedná se o soubor s názvem Atributy (*CAOS_attributesFile.txt*), který obsahuje nalezené atributy metodou CAOS a obsahuje šest sloupců. První a druhý sloupec uvádějí pozici skupin, které byly vytvořeny v souboru skupiny. Třetí sloupec ukazuje pozici zarovnání, která je potenciálně diagnostická, a čtvrtý sloupec udává míru důvěry v diagnóze. Pouze diagnostiky s hodnotami 1,00 jsou považovány za čistě diagnostické.

Další soubor má název Skupiny (*CAOS_groupFile.txt*) a seskupuje názvy a číslování skupin a vede záznamy o uzlech během klasifikace, kdy byla vytvořena pravidla skupiny pro daný uzel. Dalšími dvěma soubory, už pro analýzy méně podstatnými, je soubor *CAOS_overviewFile.txt,* který obsahuje vstupní data ve formátu CAOS a soubor *help.txt,* který podrobně vysvětluje výstup. [5]

CAOS Barcoder

Vstupem do další části programu jsou vygenerované soubory Atributy a Skupiny společně se souborem s analyzovanými sekvencemi ve formátu FASTA. Zde dochází k vytvoření souboru CAOS barcode. Výsledek tohoto nástroje bude obsahovat referenční matici *(Ref_matrix.xls),* která obsahuje seznam všech jedinečných znaků pro každý uzel v jediném grafu. Pokud se soubory bude vstupovat i soubor se zarovnanými sekvencemi ve formátu FASTA, bude výsledek také obsahovat jeden nebo více přehledů o souborech, kde se vytřídily všechny jedinečné znaky jednotlivých taxonů pro všechny uzly v samostatných grafech.

V referenční matici první sloupec (s výjimkou prvního řádku) znázorňuje polohu znaku v rámci zarovnání a dále zobrazuje názvy taxonů přítomné ve větvi v daném místě ve stromu. Číslo za znakem ukazuje hodnotu spolehlivosti. Každé dva řádky ve schématu představují jeden uzel ve fylogenetickém stromu. [5]

CAOS Classifier

Další blok již slouží k samotné klasifikaci neznámé sekvence. CAOS může klasifikovat sekvence do existujícího hierarchického systému pomocí znakově orientovaného barcodingu. Stačí zadat buď vlastní čárový kód (barcode) vytvořený datovým souborem CAOS, pomocí bloků popsaných výše, nebo vybrat stávající barcode z CAOS Library. Ve druhém kroku se vybere FASTA soubor, který obsahuje jednu nebo více sekvencí, které chceme, aby byly klasifikovány.

Výsledkem tohoto nástroje bude taxonomické seskupení, které bude obsahovat analyzovanou sekvenci a odpovídající geneticky nejbližší sekvenci, a nejlepší shoda pro obě sekvence ve FASTA souboru. Je zde i možnost se podívat na odkaz se zarovnáním těchto dvou sekvencí. [5]

CAOS Library

Tato sekce obsahuje databázi jak souborů FASTA, tak souborů NEXUS, včetně již vytvořených CAOS barcode pro potřeby klasifikace, které mohou být také využívány. [5]

3.2.3 Výsledky analýzy testovaného souboru

Analýza byla provedena na již popsaném souboru s názvem *CAOSTRAIN.fasta*, čítající padesát DNA barcode sekvencí, které byly globálně zarovnány, a rovněž obsahují čtyři živočišné druhy. Z tohoto FASTA souboru byl pomocí programu Mesquite vytvořen požadovaný formát NEXUS s názvem *CAOSTRAIN.nex*. Následně byl pomocí takto vytvořeného souboru vytvořen fylogenetický strom v programu PAUP s názvem *CAOSTRAIN.tre*, který splňoval výše uvedené podmínky. Poté proběhla úprava souboru NEXUS za pomocí vytvořeného stromu opět v programu Mesquite, tak aby konečný NEXUS soubor se sekvencemi obsahoval i informaci o topologii stromu.

Takto vytvořený soubor vstupoval do první části programu CAOS, tedy CAOS Analyzer, kde se po úspěšném proběhnutí programu vytvořily čtyři textové soubory *CAOS_attributesFile.txt, CAOS_groupFile.txt , CAOS_overviewFile.txt* a soubor *help.txt.*

Tímto způsobem získané soubory byly poté použity jako vstupy do dalšího bloku programu CAOS Barcoder. Jako možný výstup z nabízených možností byl zvolen soubor, který bude obsahovat pouze atributy sPu + sPr + další. Po proběhnutí programu byl na výstupu získán soubor s názvem *Ref_matrix.cvs,* kde první sloupec (s výjimkou prvního řádku) zobrazuje názvy taxonů přítomné ve větvi v daném místě ve stromu. V dalších

sloupcích jsou uvedené charakteristické znaky, kde číslo za tímto znakem ukazuje hodnotu spolehlivosti. Každé dva řádky ve schématu představují jeden uzel ve fylogenetickém stromu.

Dalším krokem bylo použití vytvořeného souboru s atributy, k analýze neznáme vstupující sekvence. Tento poslední krok sice proběhl, ale bohužel neúspěšně. Došlo ke srovnání testované sekvence se sekvencemi v souboru pro klasifikaci, ovšem ve výsledné tabulce v poli pro shodu, byla vypsána chyba programu, která dle informací obsažených v nápovědě, může proběhnout z důvodu nedostatečně kvalitních získaných atributů, respektive jejich nedostatečného počtu pro klasifikaci.

Na současném výstupu programu CAOS v jeho závěrečné fázi se dle autorů webových stránek stále pracuje takovým způsobem, aby analýza byla spolehlivější. Z uvedeného důvodu nebylo možné analyzovat vybraná data a podat konečné výsledky analýzy testovaného souboru touto metodou.

Aby testování programu CAOS nebylo ukončeno neúspěšnou analýzou a mohla se ověřit jeho funkčnost, rozhodla jsem se pro vyzkoušení dalšího FASTA souboru s názvem *QuickTest.fas,* který byl k dispozici právě na webových stránkách programu COAS.

Daný soubor byl upraven stejným postupem jako předešlý. Do programu CAOS Analyzer vstupoval hotový soubor ve formátu NEXUS, obsahující zarovnané sekvence a informaci o podobě fylogenetického stromu. Byly získány čtyři textové soubory *CAOS_attributesFile.txt, CAOS_groupFile.txt , CAOS_overviewFile.txt* a soubor *help.txt.*

Tímto způsobem získané soubory byly opět poté použity jako vstupy do dalšího bloku programu CAOS Barcoder. Jako možný výstup z nabízených možností byl zvolen soubor, který bude obsahovat pouze atributy sPu + sPr + další. Po proběhnutí programu byl na výstupu získán soubor s názvem *Ref_matrix.cvs.* Dalším krokem bylo použití tohoto vytvořeného souboru s atributy pro klasifikaci sekvence druhu.

Konečná analýza v tomto případě proběhla i s uvedením míry shody, která zde byla vyčíslena na 100% u celkem osmi přiřazených druhů. Jediným problémem bylo, že ve výsledný soubor neobsahovat názvy jednotlivých sekvenci, ale pouze jejich ID. Tento výstupu nalezneme v příloze č. 6.

Při snaze analýzu zakončit i s obsahem jmen jednotlivých druhů, byl celý postup analýzy opakován od počátečního kroku. Jedinou změnou v postupu bylo využití programu pro počáteční zarovnání sekvencí, kdy byla využita webová aplikace MUSCLE. Výsledný soubor již obsahovat jména i ve sloupci s názvem Best hit, tedy pro sekvence s nejlepší shodou, ovšem se zde nepodařilo míru shody vyčíslit, ale byla opět vypsána chyba jako v případě testování předešlého souboru.

Toto pouze dokazuje, že práce s tímto programem je velice složitá, převážně při tvorbě správné podoby vstupujícího formátu NEXUS, kdy malé niance v podobě souboru mohou ovlivnit výslednou analýzu.

3.3 Metoda BLOG

3.3.1 Popis programu

Program ve verzi BLOG2.0 je dostupný na webové stránce http://dmb.iasi.cnr.it/blog.php. Jeho instalace je snadná a práce s programem je také lehce pochopitelná. Náhled uživatelského rozhraní je v příloze č. 4. Pro práci s programem jsou nejdůležitější tlačítka vlevo. A to **Open file**, kam zadáváme trénovací soubor s DNA sekvencemi a **Open Test file**, kam se ukládá cesta k souboru s DNA sekvencemi, které mají být analyzovány. Analýza se poté spustí pomocí tlačítka **Run Blog.**

Do programu vstupují sekvence ve FASTA formátu a je třeba se zaměřit na název vstupujícího souboru, neboť pokud se v názvu objeví nějaké neočekávané znaky typu mezera nebo podtržítko, program neproběhne úspěšně.

Pomocí tlačítka s názvem **Set Parameters** si můžeme nastavit parametry analýzy pro naše potřeby. Na začátku jsou hodnoty nastaveny defaultně, lze zde například nastavit zkracování sekvencí dle nejkratší nebo typy zarovnání sekvencí. [8]

Pokud navolíme cestu k souboru pomocí políčka **Open File**, objeví se nám okno v podobě, jaké je na obrázku v příloze č. 4. Vidíme, že došlo k naplnění obsahu pole s názvem **Data**, ve kterém můžeme najít sloupce s názvy *Name* obsahující ID sekvence, *Class* obsahující název sekvence (organismu) a *Sequence* s DNA sekvencí. Pod tímto oknem vidíme okno grafu s názvem **Plot**, který představuje množství nukleotidů v dané sekvenci. Je zobrazena aktuálně ta sekvence, kterou si označíme kurzorem myši. Napravo od tlačítka **Set Parameters**, se objevila rozbalovací lišta, která obsahuje název nalezených druhů organismů.

Jestliže si zde zvolíme druh a pozici nukleotidu v sekvenci pomocí čítače vedle, a stiskneme tlačítko **Get**, v okně pod rozbalovací lištou můžeme vyčíst průměrné zastoupení jednotlivých nukleotidů na této pozici u tohoto druhu v procentech.

Pokud nyní stiskneme tlačítko **Run Blog** aniž bychom zadali testovací soubor, který by měl obsahovat sekvence k analýze, program i tak proběhne úspěšně. Ze zadaného souboru sám automaticky si vytvoří jak množinu trénovací, kterou tvoří asi 80% vstupních sekvencí, tak množinu testovací, která obsahuje zbylé sekvence. Po proběhnutí programu se v okně objeví šest nových výstupů, které se zobrazí za záložkou s názvem **Data**, zároveň dochází k vytvoření šesti souborů ve formátu MS Excel s totožným obsahem, které jsou umístěny ve složce s analyzovaným FASTA souborem.

Jestliže bychom navolili zároveň cestu k testovací množině, byla by tu i záložka s názvem Test data, která má stejné parametry jako záložka data, ale obsahuje informace o testovací množině. [8]

3.3.2 Popis výstupních dat

Záložka **Blog Output** nám pouze sděluje, jak program proběhl, zda nedošlo k chybě a podobně. Soubory **TEST Statistics** a **TRAIN Statistics** nám představují tzv. klasifikační matice. Jedná se vlastně o míry klasifikace, které jsou stanoveny jako počet a procento správně nebo nesprávně klasifikovaných vzorků a neklasifikovaných vzorků. Podobu takto vytvořených matic v programu BLOG2.0 vidíme na obrázcích v příloze č. 5.

V souboru **TEST formulas** nalézáme právě zmiňované logické formule, které přidělují vzorek ke druhu, což jsou malá pravidla ve tvaru "jestliže - pak“, na kterých je metoda BLOG (Barcoding with LOGic) založena. Ty se vytvořily na základě zadané trénovací množiny a můžeme si je prohlédnout v příloze č. 5.

Soubory TEST Confmatrix a TRAIN Confmatrix nám představují, tzv. matice záměn poskytující detailní informace o správnosti klasifikace a o případné křížové klasifikaci. Tyto soubory můžeme vidět na v příloze č. 5. [8]

3.3.3 Výsledky analýzy testovaného souboru

Pro potřeby analýzy jsem si pracovní FASTA soubor o 50 sekvencích rozdělila na dva soubory o 25 sekvencích, obsahující stejné již zmíněné čtyři živočišné druhy, tak abych získala trénovací množinu s názvem BLOGTRAIN a testovací množinu s názvem BLOGTEST. Tyto soubory jsem vložila do programu a provedla analýzu výše popsaným způsobem. Níže budou uvedena a pospsána výsledná výstupní data.

Vzhledem k dobře zvoleným vstupním sekvencím proběhla celá analýza naprosto úspěšně. Informace v soubory obsahující klasifikační matice (TRAIN Statistics, TEST Statistics) jsou velice podobné. Sdělují nám tedy, že soubory shodně obsahují 25 sekvencí, správně klasifikovanách sekvencí bylo 25, špatně klasifikovaných sekvencí bylo 0, neklasifikovaných sekvencí bylo též 0. Totéž je vyjádřeno v procentech.

Jediným rozdílem těchto vstupních dat bylo, že zde vstupoval různý počet zástupců od každého druhu. V trénovací množině bylo 6 sekvencí druhu *Adela caeruleella*, 7 sekvencí druhu *Aidos perfusa admiranda*, 5 sekvencí druhu *Anthela ocellata* a 7 sekvencí druhu *Atteva aurea.* V testovací množině bylo 6 sekvencí druhu *Adela caeruleella*, 9 sekvencí druhu *Aidos perfusa admiranda*, 5 sekvencí druhu *Anthela ocellata* a 5 sekvencí druhu *Atteva aurea.* Všechny tyto informace můžeme vidět v následujících tabulkách, která program vytváří ve formátu CSV. [8]

Tabulka 1: Soubor TRAIN Statistics ve formátu CSV

STATISTICS						
	tot	unknown	adela_ caeruleella	aidos_perfusa_ admiranda	anthela_ ocellata	atteva_ aurea
No. El[1]	25	0	6	7	5	7
No. Correct Class. [2]	25	0	6	7	5	7
No. Wrong Class. [3]	0	0	0	0	0	0
No. Not Class.[4]	0	0	0	0	0	0
Pct. Correct Class. [5]	100	0	100	100	100	100
Pct. Wrong Class. [6]	0	0	0	0	0	0
Pct. Not Class.[7]	0	0	0	0	0	0

Tabulka 2: Soubor TEST Statistics ve formátu CSV

STATISTICS						
	tot	**unknown**	**adela_ caeruleella**	**aidos_perfusa_ admiranda**	**anthela_ ocellata**	**atteva_ aurea**
No. Of El	**25**	**0**	**6**	**9**	**5**	**5**
No. Correct. Class.	**25**	**0**	**6**	**9**	**5**	**5**
No. Wrong Class.	**0**	**0**	**0**	**0**	**0**	**0**
No. Not Class.	**0**	**0**	**0**	**0**	**0**	**0**
Pct. Correct Class.	**100**	**0**	**100**	**100**	**100**	**100**
Pct. Wrong Class.	**0**	**0**	**0**	**0**	**0**	**0**
Pct. Not Class.	**0**	**0**	**0**	**0**	**0**	**0**

Dalším výstupem funkce je soubor obsahující logické formule (TEST formulas) pro přidělování vzorků ke druhům, které se vytvořili ze zadané trénovací množiny. Tento výstup je umístěn v pořadí další tabulce.

Pro druh *Adela caeruleella* se vytvořil logický vzorec ve tvaru **pos38 = C**, což znamená, že pokud na pozici číslo 38 v testované sekvenci bude nukleotid s bází cythosinem, přiřadí se ke druhu *Adela caeruleella.* V programu můžeme zjistit, že u sekvencí tohoto druhu se na pozici 38 báze C vykytovala v celých 100 % případů z trénovací množiny. Vzhledem k charakteru všech vstupujících sekvencí, se toto pravidlo i násedující, jeví jako vhodné pro přiřazení testované sekvence ke známému druhu.

Pro další druh *Aidos perfusa admiranda* se vytvořil logický vzorec ve tvaru pos471 = C. U sekvencí tohoto druhu se na pozici 471 báze C vyskytovala rovněž ve 100% případů. Pro druh *Anthela ocellata* se vytvořil logický vzorec ve tvaru pos271 = C. U

[1] Počet elementů (prvků);
[2] Počet správně klasifikovaných elementů
[3] Počet špatně klasifikovaných elementů
[4] Počet neklasifikovaných elementů
[5] Procento správně klasifikovaných elementů
[6] Procento špatně klasifikovaných elementů
[7] Procento neklasifikovaných elementů

sekvencí tohoto druhu se na pozici 271 báze C vykytovala také ve 100% případů. Pro poslední druh *Atteva aurea* se vytvořil logický vzorec ve tvaru pos343 = A. U sekvencí tohoto druhu pro bázi A na pozici 341 platí totéž co u ostatních.

Tabulka 3: Soubor TEST formulas ve formátu CSV

CLASS 1:	adela_caeruleella		CLASS 3:	anthela_ocellata
pos38=C			**pos271=C**	
Coverage:	1		**Coverage:**	1
False Negative:	0		**False Negative:**	0
False Positive:	0		**False Positive:**	0
Score (Laplace):	0.7		**Score (Laplace):**	0.667
FP/TP:	0		**FP/TP:**	0
CLASS 2:	**aidos_perfusa_admiranda**		CLASS 4:	**atteva_aurea**
pos471=C			**pos343=A**	
Coverage:	1		**Coverage:**	1
False Negative:	0		**False Negative:**	0
False Positive:	0		**False Positive:**	0
Score (Laplace):	0.727		**Score (Laplace):**	0.727
FP/TP:	0		**FP/TP:**	0

Posledními výstupními soubory jsou soubory TRAIN Confmatrix a TEST Confmatrix, obsahující matice záměn. V tabulkách 4 a 5 můžeme pozorovat, že k žádné záměně ani křížové klasifikace nedošlo tedy, že nebyl žádný druh přiřazen chybně k neodpovídajícímu druhu.

Tabulka 4: Soubor TRAIN Confmatrix ve formátu CSV

	Unknown	**adela_ caeruleella**	**aidos_perfusa _admiranda**	**anthela _ocellata**	**atteva _aurea**
Unknown	0	0	0	0	0
adela_caeruleella	0	6	0	0	0
aidos_perfusa_admiranda	0	0	7	0	0
anthela_ocellata	0	0	0	5	0
atteva_aurea	0	0	0	0	7

Tabulka 5: Soubor TEST Confmatrix ve formátu CSV

	Unknown	**adela _caeruleella**	**aidos_perfusa _admiranda**	**anthela _ocellata**	**atteva _aurea**
Unknown	0	0	0	0	0
adela_caeruleella	0	6	0	0	0
aidos_perfusa_admiranda	0	0	9	0	0
anthela_ocellata	0	0	0	5	0
atteva_aurea	0	0	0	0	5

3.4 Vyhodnocení testovaných metod

Z výše popsaných výsledků plyne, že na testovaném souboru úspěšně proběhla pouze analýza pomocí programu BLOG2.0. Tato metoda má přívětivé uživatelské rozhraní a jeví se jako jednodušší i pro případnou implementaci v programovém prostředí. Dle výsledků z analýzy, za použití testovaného souboru, se tato metoda jeví jako spolehlivá a vhodná k analýze DNA barcodových sekvencí, neboť úspěšnost správné klasifikace zde byla 100%.

Analýza pomocí metody CAOS se na testovaném ani na alternativním souboru dat nezdařila, pravděpodobné kvůli nedokonalosti konečné analýzy souboru on-line programem. Tuto metodu, zejména její program považuji za velice náročnou pro začínající uživatele. Komplikace se objevují převážně v postupu správného vytvoření vstupních souborů, kdy je zapotřebí seznámit se a využívat další programy pro práci s genomickými daty. Samotná analýza na webových stránkách je rozsáhlá a v případě, že program oznámí chybu, není ve většině případů dostatek relevantních zdrojů s potřebnými informacemi, kde by byl popsán postup pro odstranění nahlášené chyby.

Z uvedených důvodů bych se v navazující praktické části diplomové práce zabývala vlastní realizací metody založené na principech metody BLOG, neboť se tato metoda na testovaném souboru osvědčila a výsledky byly více než uspokojivé.

3.5 Vlastní metoda

Jak již bylo zmíněno v odstavcích výše, tvorba vlastní metody bude založena na principech metody BLOG (Barcoding with LOGic) a to z důvodů snadnější realizace a kvalitě výsledků získaných analýzou. Vlastní metoda bude realizována v programovém prostředí Matlab s použitím dostupných nástrojů z bioinformatického toolboxu, které Matlab nabízí.

Bude se jednat o vytvoření referenčního souboru tzv. charakteristické druhové struktury, pomocí trénovací množiny obsahující zástupce odlišných druhů. Soubor bude obsahovat charakteristické informace o DNA sekvencích použitých zástupců živočišných druhů. Takto vytvořený soubor bude poté použit pro klasifikaci neznámé DNA sekvence, která má být analyzována.

Výsledkem a zároveň výstupem analýzy by měl být výčet tří zástupců druhů, se kterými si byla testovaná sekvence z hlediska hodnocených parametrů nejvíce podobná. Buse se samozřejmě jednat převážně o zástupce druhů, které byly použity pro tvorbu charakteristické druhové struktury.

3.5.1 Teoretický návrh metody

Program pro vytvoření charakteristické druhové struktury

Vstupem do programu bude soubor ve formátu FASTA, který by měl obsahovat více než jednoho zástupce živočišného druhu. Zároveň by v tomto souboru měl být daný druh zastoupený ideálně pěti a více sekvencemi, aby mohl být zahrnut do dalšího zpracování.

Ve FASTA souboru se sekvencemi bude poté provedena krátká analýza, která zjistí, kolik jednotlivých živočišných druhů se v tomto souboru vyskytuje a kolika sekvencemi jsou zde tyto jednotlivé druhy zastoupeny. Do dalšího zpracování postoupí pouze zástupci těch živočišných druhů, které jsou zastoupeny pěti a více sekvencemi.

Pro každý takto zvolený živočišný druh se v další části programu vytvoří tzv. charakteristická druhová struktura. Tato struktura bude obsahovat celkem pět polí, které budou reprezentovat charakteristické znaky vytvořené ze vstupních druhových sekvencí a bude sloužit k pozdější identifikaci sekvence neznámého druhu.

První pole této charakteristické druhové struktury bude obsahovat proměnnou s názvem konzervované pozice. Konzervovanými pozicemi máme na mysli ty pozice v testovaných sekvencích, ve kterých se vyskytoval stejný nukleotid u všech vstupních sekvencí, byl zde tedy z hlediska aktuálně vstupujících sekvencí tzv. zakonzervován. Mimo indexů těchto pozic, bude v tomto poli ještě vytvořeno tzv. sekvenční logo pro konzervované

pozice, které nám říká procentuální zastoupení jednotlivých nukleotidů (A, C, G, T) na dané pozici. Vzhledem ke konzervovaným pozicím se v sekvenčním logu bude vyskytovat pouze hodnota 100 pro přítomnost daného nukleotidu na dané pozici nebo hodnota 0 pro jeho nepřítomnost.

Druhé pole charakteristické druhové struktury bude obsahovat proměnnou s názvem nekonzervované pozice. Protože zde probíhá analýza i tzv. nekonzervovaných pozic, tedy pozic, kde není u vstupujících sekvencí na dané pozici vždy shodný nukleotid, ale objevují se zde dva až čtyři rozdílné nukleotidy. V tomto poli bude opět mimo indexů nekonzervovaných pozic uloženo sekvenční logo pro nekonzervované pozice, které bude obsahovat procentuální zastoupení jednotlivých nukleotidů na dané pozici v rozsahu 0 až sto a bude zaokrouhleno na celá čísla.

Hodnota délky konsensuální sekvence, vytvořené pomocí vstupujících druhových sekvencí, bude uložena do třetího pole charakteristické druhové struktury. Pod pojmem konsensuální sekvence si můžeme představit charakteristickou a relativně konzervovanou sekvence bází DNA společnou většímu počtu sekvencí. Samotná podoba konsensuální sekvence bude uložena do následujícího pole struktury.

Páté pole charakteristické druhové struktury můžeme označit pouze za doplňkové, jelikož nebude obsahovat žádná data, která by byla použita při analýze, ale bude obsahovat pouze název živočišného druhu pro pozdější identifikaci výsledků analýzy.

Výstupem tohoto programu bude charakteristická druhová struktura, popřípadě více charakteristických druhových struktury v jednom soubory v závislosti na počtu a charakteru vstupujícího FASTA souboru. Po novém spuštění programu s odlišným vstupním souborem sekvencí bude možno nově vytvořené struktury přidat do stávajícího souboru.

Program pro analýzu

Vstupem do programu pro analýzu bude soubor s charakteristickými druhovými strukturami, vytvořený pomocí prvního programu, a sekvence, která se má analyzovat.

První zásadní operací v tomto programu bude zarovnání testované sekvence jednotlivě a postupně se všemi konsensuálními sekvencemi z referenčních struktur pro potřeby další analýzy.

Poté se budou procházet pozice této sekvence odpovídající právě konzervovaným pozicím z hlediska referenční konsensuální sekvence. Bude se zde zkoumat přítomnost či nepřítomnost konzervovaného nukleotidu v testované vstupní sekvenci. Můžou zde nastat čtyři možné situace, které jsou ohodnoceny jistým skórovacím systémem, který každé z variant přiřadí nízkou numerickou hodnotu.

Situace si můžeme představit například z hlediska nukleotidu adeninu (A) a první pozice v sekvencích. Adenin se vyskytuje v testované sekvenci na první pozici a v referenční struktuře se buď vyskytuje, nebo nevyskytuje. Naopak se se Adenin nevyskytuje v testované

sekvenci na první pozici a v referenční struktuře se opět buď vyskytuje, nebo nevyskytuje. Největší hodnota ze skórovacího systému bude logicky přidělena situaci, kdy se Adenin vyskytuje na konzervované pozici v konsensuální sekvenci a zároveň na odpovídající pozici v testované sekvenci.

Totéž se provede v případě pozic odpovídajícím nekonzervovaným pozicím z hlediska referenční konsensuální sekvence. V tomto případě je skórovací systém modifikovaný vzhledem k sekvenčnímu logu nekonzervovaných pozic, kde na rozdíl od konzervovaných pozic nejsou jen hodnoty 0 a 100, ale hodnoty právě z tohoto intervalu. Skórovací systém zde také přiděluje rozdílné numerické hodnoty. Zde se hodnota skóre odvíjí od výskytu či nevýskytu daného nukleotidu v testované sekvenci v závislosti na míře výskytu daného nukleotidu na dané pozici v konsensuální sekvenci. Jsou zde stanoveny intervaly hodnot, do kterých mohou spadat hodnoty míry výskytu v sekvenčním logu nekonzervovaných pozic.

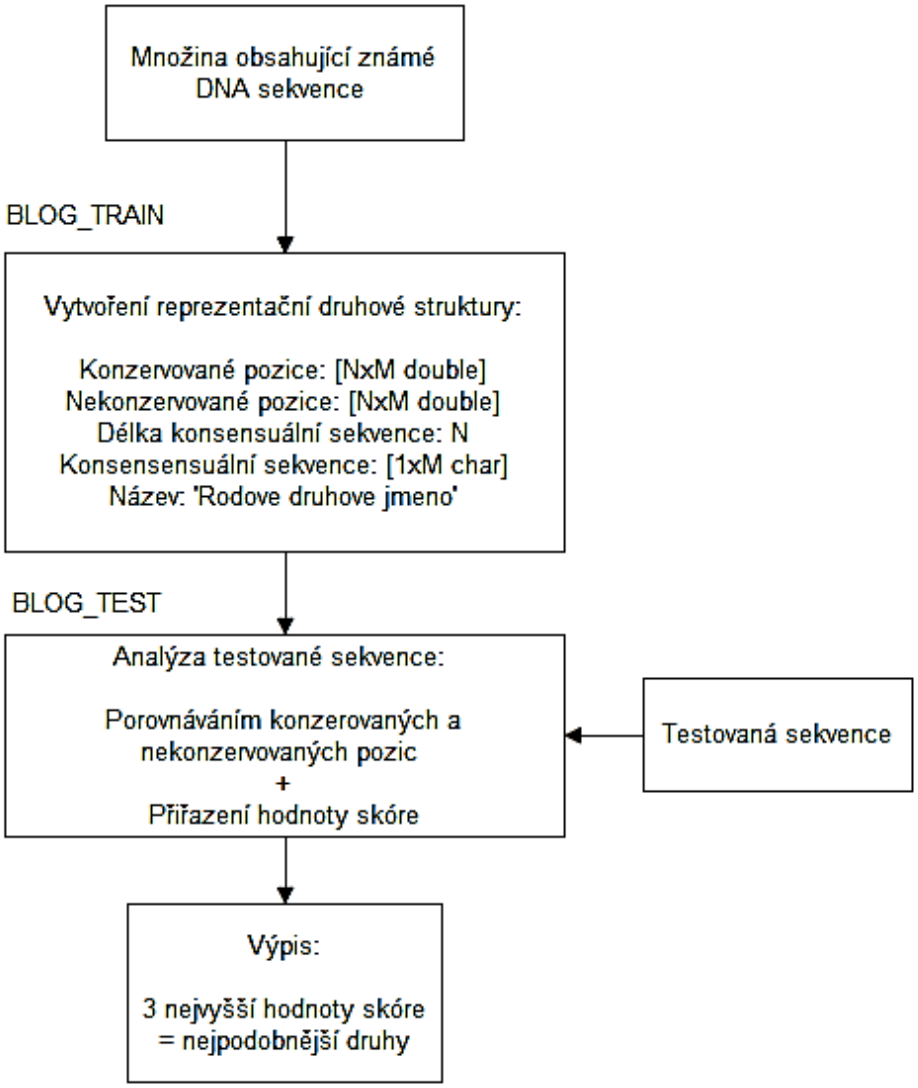
Je nutné zmínit, že celkově vyšší hodnoty přiřazujeme ve skórovacím systému v případě konzervovaných pozic, jelikož to je to hlavní, co daný druh charakterizuje. Je také potřeba brát v úvahu, že soubor sekvencí, z něhož je reprezentační struktura tvořena, nemusí být dostatečné obsáhlý a tím pádem dostatečně reprezentativní. Může se například stát, že při vstupu dalších nebo jiných sekvencí od jednoho živočišného druhu, se na místě konzervované pozice uloží jiný nukleotid než při předešlých sekvencích. Na tento fakt je potřeba dbát při volbě hodnot ve skórovacím systému, aby zde neměly příliš vysokou váhu právě konzervované pozice.

Výstupem tohoto programu budou nejvyšší tři hodnoty dosaženého celkového skóre, které vzniklo sumací dvou předešlých zmíněných skóre z konzervovaných a nekonzervovaných pozic. U těchto hodnot bude zároveň přiřazeno jméno příslušného živočišného druhu z reprezentační struktury.

3.5.2 Popis programů

Z již zmíněného výše vyplívá, že bylo nutné vytvořit dva oddělené programy pro potřeby popsané vlastní metody. Program pro vytvoření charakteristické – reprezentační struktury pro živočišný druh a druhý program pro samotnou analýzu testované sekvence. Programy byly vytvořeny jako volatelné funkce v programovém prostředí Matlab. Názorné blokové schéma, které nám pomůže se v programech zorientovat, si můžeme prohlédnout na obrázku níže.

Při vývoji funkcí byla použita data získaná z databáze BOLD. Konkrétně FASTA soubor, obsahující DNA sekvence živočišných druhů, které spadají do různých kmenů a tříd napříč živočišnou říší. Variabilita dat zde byla zajištěna zástupci z kmene strunatců (*Chordata*), v němž byla zastoupena třída savců (*Mammalia*), třída obojživelníků *(Amphibia*) a třída ptáků (*Aves*), dále z kmene členovců (*Arthorpoda*), zde byla třída Hmyz (*Insekta*) a nakonec z kmene měkkýšů (*Mollusca*).



Obrázek 4: Blokové schéma vlastní metody

Program pro vytvoření charakteristické druhové struktury

Program pro vytvoření charakteristické – reprezentační druhové struktury má pracovní název **BLOG_TRAIN.m.** Vstupem je soubor sekvencí ve formátu FASTA, který se musí zadávat pomocí apostrofů a koncovky.fas. Na následujících řádcích budou podrobně popsány jednotlivé části tohoto programu.

Soubor se sekvencemi pro tvorbu charakteristické reprezentační struktury se načítá klasickým příkazem *fastaread,* kdy se do dvou samostatných proměnných uloží jak názvy sekvencí, tak sekvence samotné.

První oddíl se věnuje selekci sekvencí dle jejich druhových názvů pro následující vstup do další části programu. Z proměnné s názvy jsou získány jména jednotlivých zástupců druhů pomocí vyhledávání svislých znaků v hlavičce sekvence za pomocí příkazu *strfind* a následnou indexací. Poté se v jednoduchém cyklu *for* vygenerují pouze rozdílná jména a jejich počet za pomocí příkazu pro porovnání řetězců *strcmp.* Na následujících řádcích programu jsou získány ve vnořených cyklech *for* indexy shodných druhů a do další části programu jsou poslány pouze ty, které jsou zastoupeny pěti a více sekvencemi.

Do dalšího obsáhlého cyklu postupně vstupují sekvence od shodného druhu za použití získaných indexů v předešlé části. Následuje blok, který upravuje vstupující sekvence tím způsobem, že odejímá znaky ‚N' a ‚-', reprezentující mezery a neidentifikované nukleotidy, ze začátků a konců sekvencí, tak aby zbytečně neovlivňovali další zpracování. Po této úpravě jsou vstupující sekvence zarovnány do bloku pomocí příkazu *multialign* a tento blok je následně upraven tak, aby na první pozici měly začátek aspoň dvě ze vstupujících sekvencí.

Následuje rychlé a snadné vytvoření konsensuální sekvence za pomocí příkazu *seqconsensus* z bioinformatického toolboxu Matlabu pro práci se sekvencemi. Konsensuální sekvence je vložena do čtvrtého pole vznikající reprezentační struktury. Třetí pole obsahuje v této souvislosti délku konsensuální sekvence.

Nalezení konzervovaných pozic je postupně prováděno za pomocí srovnávání vstupujících sekvencí s vytvořenou konsensuální sekvencí. Pokud se na dané pozici v konsensuální sekvenci a zároveň ve všech vstupujících sekvencích vyskytuje shodný nukleotid, je tato pozice vyhodnocena jako konzervovaná. Indexy těchto pozic jsou uloženy na první řádek matice, uložené v prvním poli vznikající reprezentační struktury. Do stejného pole je vloženo i sekvenční logo konzervovaných pozic, které má podobu matice. Jednotlivé řádky reprezentují nukleotidy v posloupnosti A, C, G, T a hodnota 100 značí přítomnost daného nukleotidu a 0 jeho nepřítomnost na dané konzervované pozici.

Nekonzervované pozice jsou získány jednoduchým systémem a to pomocí identifikací zbylých pozic konsensuální sekvence, které se vytvořili odejmutím již zjištěných pozic konzervovaných. Indexy nekonzervovaných pozic jsou spolu s vytvořeným sekvenčním logem pro nekonzervované pozice uloženy do druhého pole vznikající reprezentační struktury. Hodnoty sekvenčního loga na dané pozici a daném řádku, opět odpovídajícím danému nukleotidu, byly vypočítány jako počet daného nukleotidu na odpovídající nekonzervované pozici ve všech vstupujících druhových sekvencích, a tento počet byl následně podělen celkovým počtem sekvencí a převeden na procenta.

Závěrečnou úpravou postupně vytvořené reprezentační struktury je přidání jména zpracovávaného druhu do jejího posledního pole. Výpis struktury pro jeden druh v prostředí Matlab může vypadat následujícím způsobem.

```
          Konz_pozice: [5x532 double]
        Nekonz_pozice: [5x187 double]
Delka_kons_sekvence: 719
     Konsens_sekvence: [1x719 char]
                Nazev: 'Felis catus'
```

V souhrnu tedy první řádek charakteristické reprezentační druhové struktury obsahuje pole s názvem *Konz_pozice,* který obsahuje indexy konzervovaných pozic z hlediska konsensuální sekvence a sekvenční logo, značící zastoupení jednotlivých nukleotidů na dané pozici. Druhý řádek obsahuje proměnnou s názvem *Nekon_pozice*, ve které jsou též indexy a sekvenční logo pro pozice nekonzervované. Oba tyto řádky obsahují hodnoty tybu double.

Třetí pole v buňce nese název *Delka_kons_sekvence* a zaznamenává právě počet párů bází odpovídající konsensuální sekvenci. Další řádek s názvem *Konsens_sekvence* obsahuje konsensuální sekvenci jako sled nukleotidových bází A, C, G, T. Název sekvence je logicky uložen v poli *Nazev.*

Výstupem tohoto programu je jediný soubor, který v sobě čítá jednotlivé reprezentační struktury pro daný druh. Je zde nastavena možnost rozšíření vytvořené reprezentační struktury o nově vzniklé struktury. Toto je provedeno v závěrečné části, kdy program zjistí, zda daný soubor již existuje v cílové složce a pokud ano, rozšíří ji o nově vzniklou strukturu, která vznikla pomocí jiného souboru dat, na kterém byl program spuštěn.

Program pro analýzu

Vstupem do dalšího programu s pracovním názvem **BLOG_TEST.m** je soubor s vytvořenými reprezentačními strukturami a sekvence, která má být analyzována. Vstupní sekvence je upravena stejným způsobem jako v minulém programu, jsou tedy odejmuty nadbytečné znaky ‚N' a ‚-', na začátku a konci sekvence.

Sekvence je následně postupně zarovnána se všemi přítomnými konsensuálními sekvencemi pomocí příkazu *nwalign,* provádějící globální zarovnání nukleotidů pomocí Needleman - Wunschova algoritmu, a je zde zároveň vytvořeno i jakési vlastní sekvenční logo, reprezentující přítomnost nukleotidů na daných pozicích, pro potřeby následné analýzy. Přítomnost určitého nukleotidu je reprezentována jedničkou v odpovídajícím řádku z celkových čtyř řádků, reprezentující postupně Adenin, Cytosin, Guanin a Tymin.

Před samotným porovnáváním testované sekvence s konzervovanými a nekonzervovanými pozicemi je upraveno indexování testované sekvence, vzhledem k možnému posunu pozic, který mohl být způsoben předchozím zarovnáním. Způsob porovnávání pozic testované sekvence s konzervovanými pozicemi konsensuální sekvence, je založen na porovnání nukleotidu zastoupeného na dané konzervované pozici v konsensuální sekvenci s nukleotidem, který na dané pozici obsahuje testovaná sekvence. V tomto okamžiku se definují hodnoty pro první ze skórovacích systémů. Hodnoty byly experimentálně určeny následovně.

Pokud je určitý nukleotid zastoupen v testované sekvenci a rovněž je zastoupený na konzervované pozici v konsensuální sekvenci, přiřadí se hodnota 0.6. Pokud je určitý nukleotid zastoupen v testované sekvenci, ale není zastoupený na konzervované pozici v konsensuální sekvenci, přiřadí se hodnota 0.1. Pokud je určitý nukleotid zastoupený na konzervované pozici v konsensuální sekvenci, ale není zastoupený v testované sekvenci, přiřadí se také hodnota 0.1. A pokud se určitý nukleotid, například Adenin, nevyskytuje ani na konzervované pozici v konsensuální sekvenci ani v testované sekvenci, je přiřazena hodnota 0.2. Přehledně zvolené hodnoty skóre ukazuje následující tabulka.

Tabulka 6: Hodnoty skóre pro konzervované pozice

Přítomnost nukleotidu na dané pozici (hodnota v sekvenčním logu)	Testovaná sekvence	Konsensuální sekvence	Skóre
	Ano (1)	Ano (100)	0.6
	Ano (1)	Ne (0)	0.1
	Ne (0)	Ano (100)	0.1
	Ne (0)	Ne (0)	0.2

Způsob porovnávání testované sekvence s nekonzervovanými pozicemi konsensuální sekvence, je založen na porovnání nukleotidu zastoupeného na dané konzervované pozici testované sekvence, jedničkou reprezentována jeho přítomnost, s procentuálním zastoupením nukleotidu na dané pozici v konsensuální sekvenci. Přehledně hodnoty skóre stanovené experimentálně pro nekonzervované pozice ukazuje opět následující tabulka.

Tabulka 7: Hodnoty skóre pro nekonzervované pozice

	Testovaná sekvence	Interval konsensuální sekvence	Skóre
Přítomnost nukleotidu na dané pozici / hodnota v sekvenčním logu	Ano (1)	< 0 -25 >	0
		< 26 -50 >	0.1
		< 51 -75 >	0.2
		< 76 -100 >	0.3
	Ne (0)	< 0 -25 >	0.2
		< 26 -50 >	0.1
		< 51 -75 >	0.05
		< 76 -100 >	0.05

Hodnota výsledného skóre pro danou reprezentační strukturu je vypočítána jako suma ze skóre pro konzervované pozice a ze skóre nekonzervované pozice. Konečným výstupem programu je výpis tří nejvyšších hodnot skóre s názvem odpovídajícího druhu, reprezentovaného vybranou strukturou. Tato data jsou rovněž exportována do souboru MS Excel.

3.5.3 Výsledky analýzy

Program byl vyvíjen a testován pomocí sekvencí z databáze BOLD, kde byly zastoupeny různé živočišné druhy, které jsou popsány v úvodu teoretické části této práce. Pro ověření správnosti fungování programu se nabízí otestovat, zda testovaná sekvence, která byla rovněž použita společně s ostatními sekvencemi pro tvorbu referenční struktury, bude opět správně přiřazena ke svému druhu.

Zástupci druhů, pro které se poprvé vytvářely referenční struktury a které obsahovali vždy pět až osm sekvencí, byly následující: *Adela caeruleella* (motýl), *Felis catus* (kočka), *Otus lempiji* (sova), *Theba subdentata* (hlemýžď) a *Mantidactylus femoralis* (žába). Všechny tyto sekvence byly sdruženy do FASTA souboru s názvem *Soubory.fas* a tento soubor byl dán na vstup programu BLOG_TRAIN.m, jehož výstupem byl soubor s názvem *struktura*, který

obsahoval pět rozdílných referenčních struktur pro každý zastoupený druh zvlášť. Vybraná data z tohoto soubory vidíme v tabulce pod textem.

Tabulka 8:Data z druhové struktury pro soubor Soubory.fas

Pořadí	Druh	Délka konsensuální sekvence	Počet konzervovaných pozic	Počet nekonzervovaných pozic
1	*Adela caeruleella*	658	653	5
2	*Felis catus*	719	532	187
3	*Otus lempiji*	686	683	3
4	*Theba subdentata*	682	488	194
5	*Mantidactylus femoralis*	626	385	241

V tabulce vidíme délky konsensuálních sekvencí, které jsou pro všechny vstupující druhy srovnatelné. Z počtu konzervovaných a nekonzervovaných pozic můžeme říci, že velice podobné sekvence mají zástupci druhu *Adela caeruleella a Otus lempiji.* Naopak zbývající druhy jsou charakterizovány velkým počtem rozdílných pozic mezi sekvenci téhož druhu, což nám může značně zkomplikovat další analýzu.

Získaný soubor *struktura,* dále vstupoval společně s testovanou sekvencí do programu s názvem BLOG_TEST.m. Testovaná sekvence byla postupně zastoupena všemi pěti zástupci druhů zmíněnými výše. Jaké byly výstupy programu při tomto postupu a experimentálně stanovených hodnotách ve skórovacích systémech, si můžeme prohlédnout v tabulce níže.

Tabulka 9: Výsledky analýzy sekvencí ze souboru s názvem Soubory.fas

Pořadí	Vstupující sekvence	1. Nejvyšší skóre	Odpovídá druhu	2. Nejvyšší skóre	Odpovídá druhu	3. Nejvyšší skóre	Odpovídá druhu
1	*Adela caeruleella*	787.65	*Adela caeruleella*	663.6	*Otus lempiji*	613.45	*Felis catus*
2	*Mantidactylus femoralis*	669.9	*Otus lempiji*	641.3	*Felis catus*	633.1	*Mantidactylus femoralis*
3	*Felis catus*	765.65	*Felis catus*	709.85	*Otus lempiji*	656.1	*Adela caeruleella*
4	*Otus lempiji*	820.05	*Otus lempiji*	643.65	*Adela caeruleella*	634.8	*Felis catus*
5	*Theba subdentata*	711.05	*Theba subdentata*	657.55	*Adela caeruleella*	654.45	*Otus lempiji*

V tabulce vidíme, že ve většině byl k testované sekvenci přiřazen správný druh. Ovšem v případě, kdy do programu vstupovala sekvence druhu *Mantidactylus femoralis*, jako druh nejvíce podobný a s největším skóre, byl vyhodnocen *Otus lempiji*, poté *Felis catus* a až třetí nepodobnější v pořadí byl *Mantidactylus femoralis*. Tato situace mohla být mimo jiné způsobena, že testovaná sekvence byla o více jak 100bp kratší než ostatní sekvence. Podobnou zásluhu na špatné identifikaci mohou mít i nevhodně zvolené hodnoty ve skórovacích systémech pro konzervované a nekonzervované pozice konsensuálních sekvencí.

Zároveň si zde můžeme všimnout, že hodnoty skóre se od sebe příliš neliší, jedná se o rozdíly v řádu desítek mezi jednotlivými druhy. Podobné výsledky mohou být opět způsobeny hodnotami ve skórovacích systémech.

Z tabulky dále vyplývá, že program v této podobě, tedy se zvolenými hodnotami ve skórovacím systému, nemá stoprocentní výsledky, ale výsledky se dají považovat za přinejmenším uspokojivé. Při daném rozsahu testování se pohybujeme na úspěšnosti 80%.
Pro dosažení kvalitnějších výsledků je potřeba provést další analýzy a na základě jejich výsledků hodnoty skórovacího systému upravovat.

Dalším souborem, na kterém byla vytvořená metoda testována, byl soubor s pracovním názvem *Soubory2.fas*. Tento soubor obsahoval sedm zástupců druhu *Notomyotida* (hvězdice), osm zástupců druhu *Anthopleura elegantissima* (sasanka), šest zástupců druhu *Haplothrips aculeatus* (truběnka - hmyz), pět zástupců druhu *Osteolaemus tetraspi* (krokodýl), osm zástupců druhu *Crocidura buettikoferi* (bělozubka) a čtyři zástupce třídy *Agelenopsis utahana* (pavouk).

Do programu vstupovalo celkem šest zástupců živočišných druhů. Referenčních druhových struktur bylo vytvořeno pouze pět. Struktura se nevytvořila pro druh *Agelenopsis utahana* z důvodu, že byl v souboru zastoupen pouze čtyřmi sekvencemi. Hlavní údaje z referenčních druhových struktur najdeme v tabulce pod textem.

Tabulka 10: Data z druhové struktury pro soubor Soubory2.fas

Pořadí	Druh	Délka konsensuální sekvence	Počet konzervovaných pozic	Počet nekonzervovaných pozic
1	*Notomyotida*	658	*630*	28
2	*Anthopleura elegantissima*	658	*608*	50
3	*Haplothrips aculeatus*	659	*341*	318
4	*Osteolaemus tetraspis*	513	*304*	209
5	*Crocidura buettikoferi*	635	*614*	21

V tabulce jsou vyčísleny počty konzervovaných pozic a nekonzervovaných pozic pro jednotlivé druhy společně s délkou konsensuální sekvence. Můžeme si zde všimnout, že konsensuální sekvence mají srovnatelnou délku (počet bp) až na druh *Osteolaemus tetraspis,* který obsahoval o cca 100 bp méně.

Dále vidíme, že zatímco druhy *Notomyotida, Anthopleura elegantissima a Crocidura buettikoferi* obsahují vysoký počet konzervovaných pozic vzhledem k délce konsensuální sekvence, jsou si tedy vzájemně velice podobné, tak zbývající druhy *Haplothrips aculeatus a Osteolaemus tetraspis* obsahují daleko menší počet konzervovaných pozic vzhledem ke konsensuální sekvenci a tím pádem, zde bylo nalezeno i mnoho pozic nekonzervovaných. Velký počet nekonzervovaných pozic nám značí, že vstupující sekvence tohoto druhu se vzájemně velice odlišovali v umístění nukleotidů na jednotlivých pozicích. Takto vytvořená reprezentační druhová struktura nemusí být plně funkční pro analýzu testovaných sekvencí.

Takto vytvořený soubor s referenčními druhovými strukturami byl dán na vstup programu pro analýzu společně se sekvencemi jednotlivých druhů. Výstupy analýzy ze souboru Soubory2.fas jsou shrnuty v následující tabulce.

Tabulka 11: Výsledky analýzy sekvencí ze souboru s názvem Soubory2.fas

Pořadí	Vstupující sekvence	1. Nejvyšší skóre	Odpovídá druhu	2. Nejvyšší skóre	Odpovídá druhu	3. Nejvyšší skóre	Odpovídá druhu
1	*Notomyotida*	765.8	*Notomyotida*	626.15	*Anthopleura elegantissima*	614.95	*Crocidura buettikoferi*
2	*Anthopleura elegantissima*	758.7	*Anthopleura elegantissima*	641.45	*Notomyotida*	614.95	*Crocidura buettikoferi*
3	*Haplothrips aculeatus*	617.55	*Haplothrips aculeatus*	593.45	*Notomyotida*	587.55	*Anthopleura elegantissima*
4	*Osteolaemus tetraspis*	594.95	*Notomyotida*	584.25	*Crocidura buettikoferi*	582.55	*Anthopleura elegantissima*
5	*Crocidura buettikoferi*	631.8	*Notomyotida*	619.7	*Crocidura buettikoferi*	606.1	*Anthopleura elegantissima*

V případě analýzy na tomto souboru úspěšnost identifikace sekvencí poklesla. Tři z druhů byly správně přiřazeny ke svému druhu jako nejpodobnější. U druhu *Crocidura buettikoferi*, program vyhodnotil, že je tato vstupující sekvence více podobná s druhem *Notomyotida* a poté až s korektním druhem *Crocidura buettikoferi*, ovšem hodnoty skóre v tabulce jsou si velice blízké pro obě varianty.

Negativně dopadla analýza druhu *Osteolaemus tetraspis*, tato sekvence byla postupně vyhodnocena jako nejpodobnější s druhy, které nebyly korektní. Toto mohlo být způsobeno nedostatečné reprezentativní charakteristickou strukturou pro tento druh, jelikož zde bylo velké množství nekonzervovaných pozic. Nesprávnou identifikaci mohly rovněž ovlivnit nevhodně zvolené hodnoty ve skórovacím systému, ovšem vzhledem k podobě referenční struktury lze tento vliv zanedbat.

Z výsledků této konkrétní analýzy můžeme učinit dílčí závěr, že pokud požadujeme, aby byla charakteristická druhová struktura dostatečně reprezentativní, je potřeba, aby do programu pro tvorbu této charakteristické struktury vstupoval dostatečný počet zástupců daného druhu a tyto druhy byly sami o sobě reprezentativní, což se týče převážně jejich dostateční informativní hodnoty. Informativní hodnotou mám na mysli její dostatečnou sekvenaci, tedy že se v sekvenci nebude vyskytovat příliš mnoho mezer a neurčených nukleotidů.

Posledním souborem, na kterém byla metoda testována, byl soubor s názvem Soubory3.fas. Tento soubor obsahoval sekvence popsané v úvodu práce. V souboru byli zástupci třídy motýlů a jednalo se o druhy *Adela caeruleella, Aidos perfusa admiranda, Anthela ocellata a Atteva aurea.*

Tento soubor se od předchozích odlišuje tím, že v sobě sdružuje velice příbuzné organismy, což nám může ovlivnit výsledky analýzy, vzhledem k teoretické podobnosti sekvencí zástupců těchto druhů. Vzhledem k dostatečnému počtu vstupujících sekvencí u jednotlivých druhů byly vytvořeny čtyři charakteristické druhové struktury, jejichž parametry si můžeme prohlídnout v tabulce č. 12.

Tabulka 12:Data z druhové struktury pro soubor Soubory3.fas

Pořadí	Druh	Délka konsensuální sekvence	Počet konzervovaných pozic	Počet nekonzervovaných pozic
1	*Adela caeruleella*	658	*653*	5
2	*Aidos perfusa admiranda*	658	*580*	78
3	*Anthela ocellata*	658	*302*	356
4	*Atteva aurea*	633	*386*	247

V tabulce je opět vidět, že velké množství nekonzervovaných pozic obsahují sekvence druhů *Anthela ocellata a Atteva aurea*.

Takto vytvořené druhové struktury byly použity opět k analýze vstupujících sekvencí. Výsledky analýzy shrnuje následující tabulka.

Tabulka 13:Výsledky analýzy sekvencí ze souboru s názvem Soubory3.fas

Pořadí	Vstupující sekvence	1. Nejvyšší skóre	Odpovídá druhu	2. Nejvyšší skóre	Odpovídá druhu	3. Nejvyšší skóre	Odpovídá druhu
1	*Adela caeruleella*	749.15	*Adela caeruleella*	668.25	*Aidos perfusa admiranda*	556.55	*Anthela ocellata*
2	*Aidos perfusa admiranda*	760.75	*Aidos perfusa admiranda*	723.3	*Adela caeruleella*	553.45	*Anthela ocellata*
3	*Anthela ocellata*	730	*Adela caeruleella*	692.7	*Aidos perfusa admiranda*	670.3	*Anthela ocellata*
4	*Atteva aurea*	724.1	*Adela caeruleella*	694.45	*Aidos perfusa admiranda*	673.6	*Atteva aurea*

Výsledky analýzy nejsou příliš dobré, neboť druhy *Anthela ocellata a Atteva aurea,* byly diagnostikovány až jako v pořadí třetí sekvence, které jsou si nejvíce podobné s konsensuální sekvencí vytvořenou z těchto druhů a s jejími konzervovanými, popřípadě nekonzervovanými pozicemi. Vliv na diagnostiku mimo zvolené hodnoty skórovacího systému mělo i vysoké zastoupení nekonzervovaných pozic u těchto druhů.

Ve snaze zajistit lepší výsledky analýzy v tomto souboru dat byly experimentálně změněny hodnoty ve skórovacím systému pro konzervované pozice na hodnoty, které ukazuje následující tabulka.

Tabulka 14: Upravené hodnoty skóre pro konzervované pozice

Přítomnost nukleotidu na dané pozici (hodnota v sekvenčním logu)	Testovaná sekvence	Konsensuální sekvence	Skóre
	Ano (1)	Ano (100)	0.6
	Ano (1)	Ne (0)	0.2
	Ne (0)	Ano (100)	0.1
	Ne (0)	Ne (0)	0.1

Výsledky analýzy po této úpravě hodnot skóre byly daleko uspokojivější než v předchozím případě. Testované sekvence a druhy, se kterými byly vyhodnoceny jako nejpodobnější, můžeme vidět opět v tabulce na následující stránce.

Tabulka 15:Výsledky analýzy sekvencí ze souboru s názvem Soubory3.fas po úpravě hodnot ve skórovacím systému pro konzervované pozice

Pořadí	Vstupující sekvence	1. Nejvyšší skóre	Odpovídá druhu	2. Nejvyšší skóre	Odpovídá druhu	3. Nejvyšší skóre	Odpovídá druhu
1	*Adela caeruleella*	553.25	*Adela caeruleella*	511.25	*Aidos perfusa admiranda*	473.15	*Anthela ocellata*
2	*Aidos perfusa admiranda*	586.75	*Aidos perfusa admiranda*	545.2	*Adela caeruleella*	473	*Atteva aurea*
3	*Anthela ocellata*	579.9	*Anthela ocellata*	552.7	*Adela caeruleella*	537.1	*Aidos perfusa admiranda*
4	*Atteva aurea*	558	*Atteva aurea*	548.6	*Adela caeruleella*	546.75	*Anthela ocellata*

V tabulce jasně vidíme, jak nepatrná změna v nastavení hodnot ve skórovacím systému může ovlivnit výsledky analýzy. Zároveň můžeme pozorovat, že se hodnoty skóre celkově snížili vzhledem k předchozímu případu. V tomto případě je úspěšnost analýzy stoprocentní, neboť všechny sekvence byly přiřazeny ke svému druhu jako k prvnímu v pořadí. V případě tohoto souboru, tedy můžeme vyhodnotit nastavení skórovacího systému jako účinnější.

3.6 Srovnání vlastní metody s metodou BLOG

Po realizaci vlastní metody, která byla inspirována principy analýzy DNA barcodových sekvencí metodou BLOG se nabízí jejich vzájemné srovnání.

Metoda BLOG je určena k identifikaci umístění klíčových diagnostických pozic nukleotidů u každého druhu zvlášť v plně definované trénovací množině. Ve vytvořené vlastní metodě můžeme pod diagnostickými pozicemi chápat právě zjištěné konzervované pozice vzhledem ke konsensuální sekvenci.

Metoda BLOG poté přiřazuje v závislosti na klíčových diagnostických pozicích, každému druhu logickou klasifikační formuli, tedy malá pravidla ve tvaru „jestliže – pak“, která jsou schopna charakterizovat druhy kompaktním způsobem. Je schopna tuto sekvenci rozpoznat a přiřadit ke druhu, který se již nachází v referenční knihovně. Ve vytvořené vlastní metodě se taková pravidla nevytvářejí, ale jsou nahrazena skórovacím systémem. Za pravidlo můžeme považovat pouze to, že se testovaná sekvenci přiřadí ke druhu, který dosáhl nejvyššího skóre.

Metoda BLOG dále nabízí další získané informace o analyzovaných sekvencích. Vlastní metoda v současné podobě tato data, která mohou obsahovat novou přidanou informaci o sekvencích, vytvářet neumí.

Abychom porovnali i praktickou část obou metod, tedy účinnost analýzy metodou BLOG a vlastní metodou, provedeme testování souborů s názvy *Soubory.fas, Soubory2.fas a Soubory3.fas* ještě metodou BLOG.

Výsledky analýzy prvního souboru shrnují následující dvě tabulky. První z nich označuje trénovací množinu, na základě které se tvořila pravidla. Na prvním řádku vidíme, kolik sekvencí od jednotlivých druhů bylo použito na trénovací množinu. Vidíme, že zde byla 100% úspěšnost analýzy.

Tabulka 16: Soubor TRAIN Statistics z Soubory.fas

	tot	unknown	Adela caeruleella	Felis catus	Otus lempiji	Theba subdentata	Mantidactylus femoralis
No. Of El	23	0	5	3	5	5	5
No. Correct. Class.	23	0	5	3	5	5	5
No. Wrong Class.	0	0	0	0	0	0	0
No. Not Class.	0	0	0	0	0	0	0
Pct. Correct Class.	100	0	100	100	100	100	100
Pct. Wrong Class.	0	0	0	0	0	0	0
Pct. Not Class.	0	0	0	0	0	0	0

Druhá z tabulek již prezentuje samotnou analýzu sekvencí, které byly při vstupu přiřazeny do testovací množiny. Vidíme, že analýza neproběhla úspěšně u všech vstupujících sekvencí, ale pouze u druhu *Adela caeruleella*. Jediného vstupujícího zástupce druhu *Felis Catus* se nepodařilo správně přiřadit. U zbývajících druhů byla úspěšnost 50%.

Nekvalitní výsledky mohou být způsobeny nedostatečně reprezentativní trénovací množinou, jak uvádí autoři této metody.

Tabulka 17:Soubor TEST Statistics z Soubory.fas

	tot	unknown	Adela caeruleella	Felis catus	Otus lempiji	Theba subdentata	Mantidactylus femoralis
No. Of El	9	0	2	1	2	2	2
No. Correct. Class.	5	0	2	0	1	1	1
No. Wrong Class.	2	0	0	0	1	0	1
No. Not Class.	2	0	0	1	0	1	0
Pct. Correct Class.	55.56	0	100	0	50	50	50
Pct. Wrong Class.	22.22	0	0	0	50	0	50
Pct. Not Class.	22.22	0	0	100	0	50	0

Pro přehlednost v tomto případě uvádím i tabulku reprezentující matici záměn, kde můžeme pozorovat, jak se druhy správně či nesprávně přiřazovaly k referenčním druhům na základě vygenerovaných pravidel. Hodnoty v diagonále značí počet správně diagnostikovaných zástupců druhů.

Tabulka 18: Matice záměn pro Soubory.fas

	Unknown	Adela caeruleella	Felis catus	Otus lempiji	Theba subdentata	Mantidactylus femoralis
Unknown	0	0	1	0	1	0
Adela caeruleella	0	2	0	0	0	0
Felis catus	0	0	0	0	0	0
Otus lempiji	0	0	0	1	0	0
Theba subdentata	0	0	0	0	1	1
Mantidactylus femoralis	0	0	0	1	0	1

V případě testování v pořadí druhého souboru Soubory2.fas, byla matice TRAIN Statistics opět stoprocentní, proto zde bude uvedena pouze matice TEST Statistics, kde vidíme, jak byly sekvence analyzovány dle sady pravidel vygenerovaných z trénovací množiny.

Tabulka 19: Soubor TEST Statistics z Soubory2.fas

	tot	unknown	notomyotida	Anthopleura elegantissima	Agelenopsis utahana	Haplothrips aculeatus	Osteolaemus tetraspis	Crocidura buettikoferi
No. Of El	10	0	2	2	1	2	1	2
No. Correct. Class.	5	0	2	1	1	0	0	1
No. Wrong Class.	0	0	0	0	0	0	0	0
No. Not Class.	5	0	0	1	0	2	1	1
Pct. Correct Class.	50	0	100	50	100	0	0	50
Pct. Wrong Class.	0	0	0	0	0	0	0	0
Pct. Not Class.	50	0	0	50	0	100	100	50

V tabulce můžeme pozorovat analýzu sekvencí ze souboru Soubory2.fas. Analýza zde neproběhla též úplně korektně, což dávám opět za vinu nedostatečné trénovací množině, která byla poslána na vstup program BLOG.

Co se týče analýzy třetího souboru metodou BLOG, tak ta proběhla již v první části této práce v rámci seznámení se s metodou a jejím programem. Výsledky programu naleznete v tabulkách 1-5.

Pokud bychom tedy měli vyhodnotit schopnosti analýzy obou programů na použitých datech, můžeme výsledky označit za přinejmenším srovnatelné. Obě metody měly vysoké procento správně přiřazených druhů, ovšem se snižující se kvalitou vstupních souborů pro tvorbu reprezentační struktury u vlastní metody, či ke tvorbě pravidel u metody BLOG, se míra úspěšnosti klasifikace lineárně snižovala.

4 Závěr

Tato diplomová práce seznamuje s problematikou znakově orientovaných metod DNA barcodingu. Jako dvě hlavní metody byly vybrány CAOS a BLOG. V úvodu práce je popsán jak princip klasifikace, na kterém jsou obě metody postaveny, tak fungování a používání jejich programů, včetně popisu vstupních a výstupních dat.

Cílem práce bylo se seznámit s principy výše uvedených metod a na základě výsledků analýzy zvolit jednu z nich jako vhodnější k tvorbě vlastní znakově orientované metody DNA barcodingu, sloužící ke klasifikaci sekvencí do jednotlivých druhů.

Z důvodu snadnější realizace a celkově kvalitnějších výsledků testování byly vybrány k realizaci principy metody BLOG. Vlastní metoda byla navržena a teoreticky popsána. V praktické části byly poté vytvořeny dva samostatné programy v programovém prostředí Matlab. První program s názvem BLOG_TRAIN.m slouží k vytvoření charakteristické druhové struktury ze vstupujícího FASTA souboru obsahujícího zástupce jednotlivých živočišných druhů. Tato struktura poté slouží k samotné identifikaci druhů. Druhým programem je BLOG_TEST.m, který provádí klasifikaci vstupující testované sekvence za pomocí definovaného skórovacího systému.

Konečným výstupem metody je přiřazení testované sekvence ke druhu, u kterého bylo dosaženo nejvyššího skóre, tedy největší podobnosti v analyzovaných parametrech obsažených v charakteristické druhové struktuře. Jako parametry zde můžeme označit konzervované a nekonzervované pozice v konsensuální sekvenci reprezentující daný druh.

Oba zrealizované programy jsou funkční a úspěšnost v průběhu testování byla odhadnuta na 70% až 80% v závislosti na objemu analyzovaných dat.

V závěru práce byla srovnána vlastní navržená metoda s metodou BLOG. Obě metody v závislosti na vstupních datech, vykazovaly srovnatelné výsledky. Vytvořená vlastní metoda má potenciál být dále rozšiřována a zdokonalována pro potřeby klasifikace DNA barcodových sekvencí.

5 Seznam literatury

[1] Barcode of life: What is DNA Barcoding. Barcode of Life. [online]. 19.5.2015 [cit. 2015-05-19]. Dostupné z: www.barcodeoflife.org/content/about/what-dna-barcoding

[2] BLOG – Barcoding with LOGic formulas. . [online]. [cit. 2015-05-19]. Dostupné z: http://dmb.iasi.cnr.it/blog.php

[3] Tjard Bergmann. CAOS- Workbench (Character Attribute Organization system). . [online]. [cit. 2015-05-19]. Dostupné z: http://bol.uvm.edu/caos-workbench/caos.php

[4] The Mesquite Project Team. Mesquite: A modular system for evolutionary analysis. . [online]. [cit. 2015-05-19]. Dostupné z: http://mesquiteproject.org

[5] David Swofford. PAUP: Phylogenetic Analysis Using Parsimony. PAUP*. [online]. [cit. 2015-05-19]. Dostupné z: http://paup.csit.fsu.edu/index.html

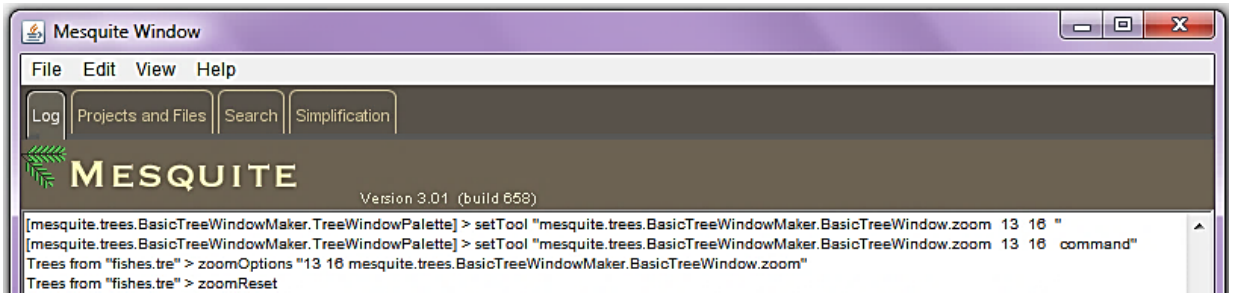
[6] Ratnasingham, S. & Hebert, P. D. N. BOLD: The Barcode of Life Data System. . [online]. [cit. 2015-05-19]. Dostupné z: http://www.boldsystems.org/

[7] SARKAR, I.N.; PLANET, P.J. a DESALLE,R.. CAOS sofware for use character- based DNA barcoding. New York: Molecular Ecology Resources, 2008. ISBN 1256-1259.

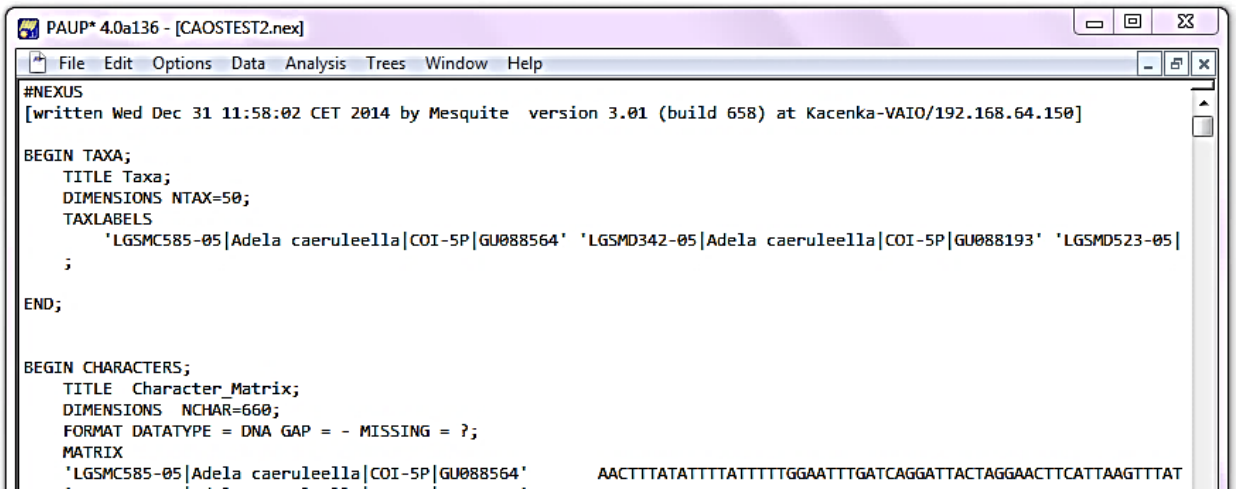
[8] WEITCHEK, E.; VAN VELZEN, R.; FELICI, G. a BERTOLAZZI P. BLOG 2.0: a software system for character-based species classification DNA Barcode sequences. What it does, how to use it.. Italy: Molecular Ecology Resources, 2013. ISBN 1043-1046.

6 Seznam příloh

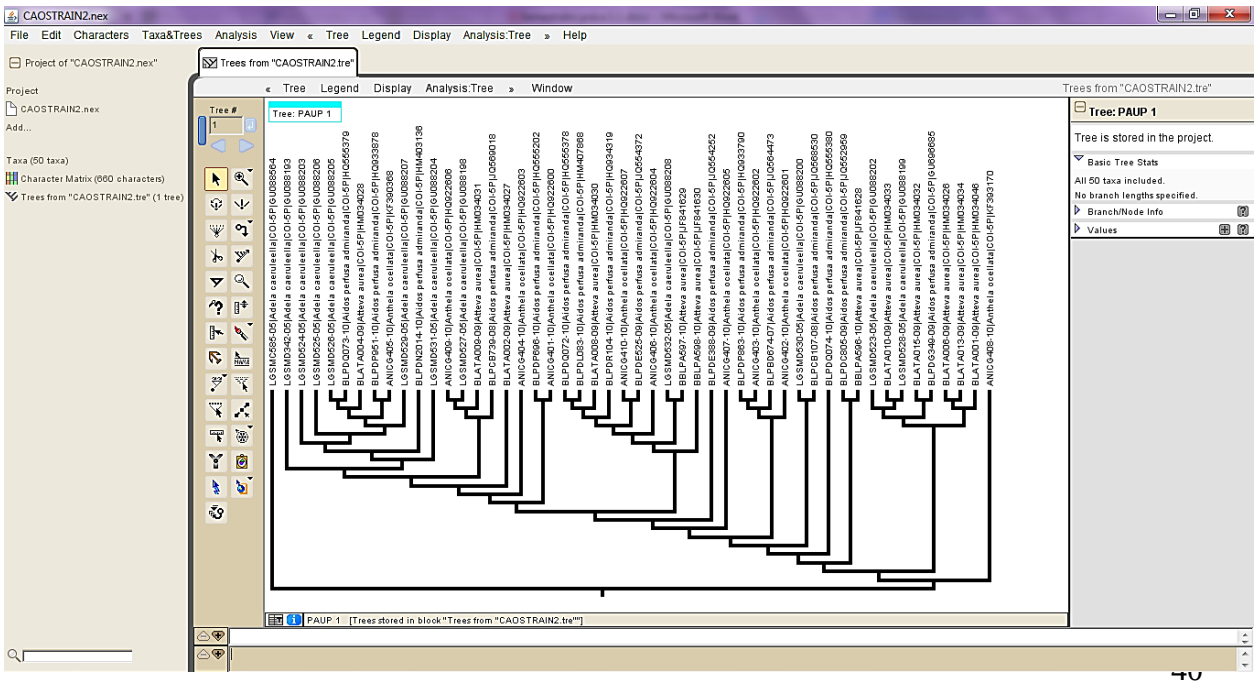
Příloha č. 1.



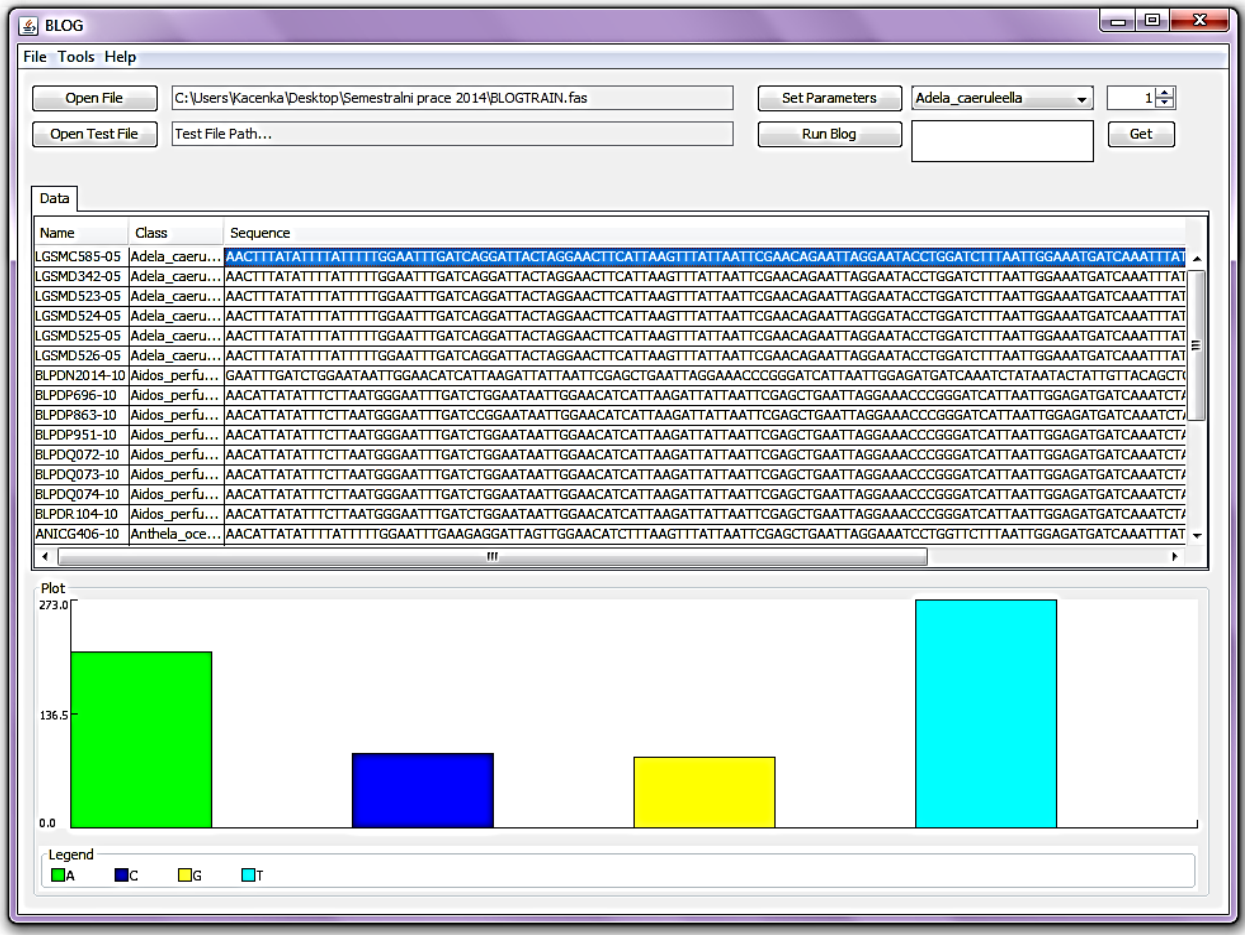
Příloha č. 2



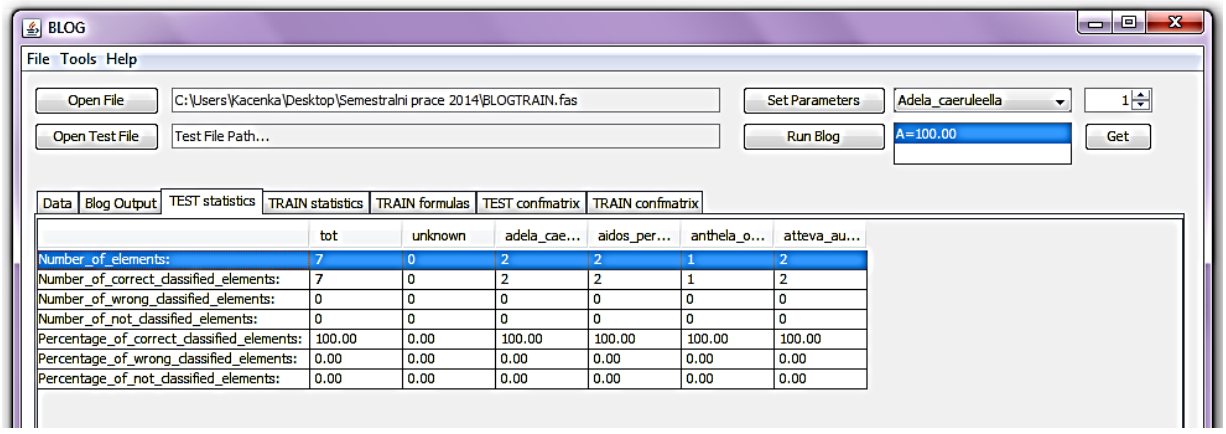
Příloha č. 3



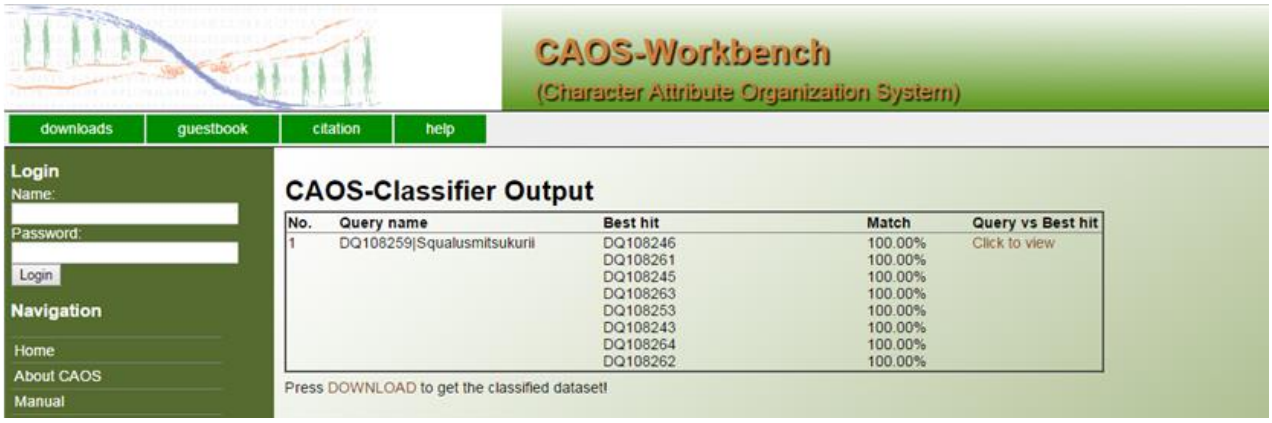
Příloha č. 4



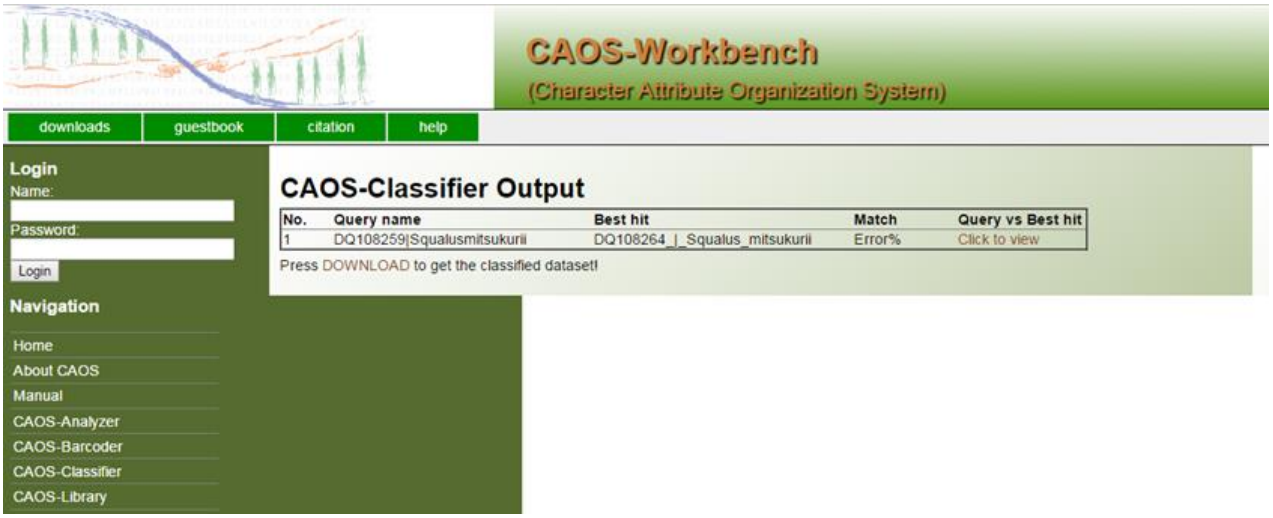
Příloha č. 5



Příloha č. 6



Příloha č. 7



Příloha č. 8

CD s programy BLOG_TRAIN a BLOG_TEST